Noble Deceit: Optimizing Social Welfare for Myopic Multi-Armed Bandits

Ashwin Maran[1], Jeremy McMahan[1], and Nathaniel Sauerberg[1]

University of Wisconsin-Madison
{amaran,jmcmahan,nsauerberg}@wisc.edu

Abstract. In the information economy, consumer-generated information greatly informs the decisions of future consumers. However, myopic consumers seek to maximize their own reward with no regard for the information they generate. By controlling the consumers' access to information, a central planner can incentivize consumers to produce more valuable information for future consumers. The myopic multi-armed bandit problem is a simple model encapsulating these issues. In this paper, we construct a simple incentive-compatible mechanism achieving constant regret for the problem. We use the novel idea of *independent phases* to selectively reveal information towards the goal of maximizing social welfare. Moreover, we characterize the distributions for which an incentive-compatible mechanism can achieve the first-best outcome and show that our general mechanism achieves first-best in such settings.

Keywords: Algorithmic Mechanism Design · Information Design · Multi-Armed Bandits · Learning Theory · Online Algorithms

1 Introduction

With society's deeper integration with technology, information generated by consumers has become an incredibly valuable commodity. Entire industries solely exist to compile and distribute this information effectively. Some of this information, such as product ratings or traffic reports, is critical for informing a consumer's decisions. For example, an agent may wish to buy the best phone case from an online retailer. Without accurate product ratings, an agent might purchase a low-quality item. However, if every agent were to simply purchase the highest rated product, alternate options would never get explored and the truly best product might never be discovered. Therefore, the seller needs some way to incentivize agents to explore new options even though they may desire to myopically choose the best option seen so far.

Waze is a popular GPS service similar to Google Maps that addresses these issues. It collects traffic information from drivers in real time and uses the observations to recommend fast routes to participants. In addition to routes it knows to be fast, Waze sometimes recommends lesser known routes to drivers in order to gather information about them. Drivers are uncertain whether a particular recommended route is known to be fast or not, but are willing to accept the recommendations since they expect the routes to be fast in general. This information asymmetry is critical to ensuring drivers do not deviate from the recommended routes. At the same time, this relationship between drivers and Waze encourages Waze to explore sparingly and only on reasonable routes so that drivers are willing to use the service.

All recommendation systems must overcome these problems. The designers of these systems want to attract and retain users, which they can plausibly do by maximizing the social welfare of their users. The critical issue for these systems is that agents both utilize information to decide on an action to take and then produce information since the system observes the outcome of their decisions. Although agents benefit from information generated in the past, they may not care about

generating new information for future agents. Rather, they just desire to exploit, achieving the best outcome they can for themselves. This fundamental trade-off between exploring and exploiting in the absence of incentive constraints is well known and traditionally captured by the *multi-arm bandit* (MAB) problem. In traditional MAB problems, a single player plays every round. With strategic agents, this would be equivalent to the system being able to force agents to take arbitrary actions. In our setting, however the best we can do is attempt to influence agents into taking an action that will maximize social welfare in the long run. We call this problem with incentives the *myopic multi-arm bandit problem* (MMAB).

The easiest way to incentivize agents to take an action they do not prefer is to pay them to do so. Many recommendation systems do this indirectly by offering discounts on particular products. When the designer is allowed to make monetary transfers, the problem is well understood [6]. An interesting and more applicable variation of the problem tries to incentivize the agents without the use of payments.

The main resource a designer has excluding payments is information asymmetry. Since these systems may not allow agents to interact with each other directly, the only information an agent has about past outcomes comes from the system itself. Using Amazon as an example, most people look at the reviews available on the website itself rather than contacting other Amazon users directly to ask about a product. Hence, by hiding certain reviews from users, Amazon could influence users to buy different products. A similar phenomenon happens with the examples of Waze and TripAdvisor.

The field of information design studies how to exploit such information asymmetries. In particular, strategically revealing information to agents could influence their decisions. To get a feel for this, we consider two simple information revealing mechanisms. The first, full transparency, reveals all past outcomes to the agents. Since agents are myopic, they will almost always choose the best option seen so far and will rarely explore. Consequently, this approach usually causes the agents to "herd", meaning the agents converge to one decision and so get stuck in a local optima. On the other extreme, the mechanism could opt to never reveal information about the past outcomes. In this no information mechanism, agents will always choose the action with the best a priori expected outcome. Consequently, it is intuitive that neither extreme is optimal and that a sophisticated mechanism which selectively reveals information is necessary to achieve optimal social welfare.

1.1 Past Work

In their seminal work, Kremer et al. [11] construct an optimal incentive-compatible (IC) mechanism for MMAB when restricted to two actions and deterministic rewards. In a follow up work, Cohen et al. [5] construct an optimal IC mechanism for any finite number of actions, for settings with only 2-point distributions and settings where the first distribution is over $[-1, 1]$ and the rest are over $\{-1, 1\}$.

In the heterogeneous agents setting, which generalizes our homogeneous setting, agents of different types may receive different rewards from the same arm. In this setting, Immorlica et al. [7] construct a constant regret mechanism for any finite number of actions when the support of the reward distributions is finite. The mechanism presented involves solving an LP. Similarly, Chen et al. [4] presents a constant regret mechanism under additional assumptions on the heterogeneity and using payments.

For stochastic rewards, several results are known. Most notably, Mansour et al. [12] gives a general reduction from MAB problems with incentives to MAB without incentive. Further, the regret achieved is nearly the best possible. Although slightly improving on this result may be

possible, many researchers consider the stochastic rewards case resolved [5, 7]. Consequently, most recent works extend this framework to more general information design problems or attempt to make the mechanisms more practical by having them reveal more information to the agents [12, 8, 14]. Notably, Mansour et al. [13] consider more complicated settings where multiple agents arrive at each time step and their interactions determine rewards. Sellke and Slivkins [15] consider how different mechanisms would work in this setting without incentives to quantify the penalty mechanisms incur from satisfying incentive-compatibility requirements.

Generally, MMAB can be seen to be at the intersection of information design and Bayesian Persuasion. Many interesting results for information design have been considered such as in [2, 9]. Similarly, Bayesian Persuasion has been an active field of research ever since the publication of [10] leading to interesting developments such as [3]. Closer to our setting, the general technique of exploiting information asymmetry is well-studied and much of what is known can be found in [16]. Furthermore, [1] delves into the tradeoffs that exist as additional competition is introduced.

Lastly, the study of the problem using monetary payments rather than information asymmetry to incentivize agents was initiated by Frazier et al. [6]. Several other works in this direction have been studied. The most similar to our setting is [4] which utilizes information asymmetry but also uses some small payments. Then, Wang and Huang [17] investigates achieving logarithmic regret at the cost of logarithmic payments.

1.2 Our Contributions

We present simple a mechanism for a fixed number of arms that always finds the best explorable arm. An arm is explorable if it is ever possible for an IC mechanism to recommend it. Therefore, whenever all arms are explorable, our mechanism achieves constant regret with respect to the strongest possible benchmark, the optimal offline mechanism, which simply picks the best arm at every time step. It also achieves constant regret with respect to the best IC mechanism unconditionally. Moreover, we show that our mechanism achieves the first-best outcome whenever it is possible for an IC mechanism to do so.

If there are m arms with discrete distributions of support size at most S, then our mechanism takes $O(S^m)$ time to make some pre-computations. After that, each agent can be recommended their arm online in constant time.

Our studied setting lies between several of the models above. On the one hand, our mechanism works on a much broader class of distributions than [5, 7, 4] (which consider discrete distributions). At the same time, our setting focuses on deterministic rather than stochastic rewards so is more restricted than that of [12, 8]. However, unlike in the stochastic setting, with deterministic rewards we can guarantee constant regret. This proves a strict separation in difficulty between the deterministic and stochastic settings that was only known before for specific classes of distributions.

2 Preliminaries

Suppose we have a set of m actions/arms $A = \{a_1, \ldots, a_m\}$. We denote by $R_1, \ldots, R_m$ random variables that represent the random rewards associated with each arm and $D_1, \ldots, D_m$ be the distributions of the rewards of each arm. The R_i are unknown a priori, but are persistent when realized, i.e. the rewards are determined by a single draw from the distribution. We assume the D_i's are each independent distributions and define $D = D_1 \times \ldots \times D_m$ to be the joint distribution of all the arms. Lastly, we define $\mu_i = E[R_i]$ and assume the arms are given in decreasing order

of expectation $\infty > \mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$. We use r_i to denote a particular realization of the random reward R_i. We overload the notation D_i to refer to the support of the distribution D_i. This will be clear from context as we will be treating it as a set in this case. We also define $D_{\max(k)} = \max\{D_1, \ldots, D_k\}$ to be the distribution of the maximum of independent draws from each of the first k arms.

In the *myopic multi-arm bandit problem* (MMAB), a (possibly infinite) sequence of agents arrive over a time horizon. When agent t arrives, he chooses an arm to pull and receives some reward. The goal of the mechanism designer or principal is to strategically reveal information about the past outcomes of actions in order to maximize the social welfare, which is the total reward received by all the agents.

We consider several performance benchmarks in this work.

1. *Offline-best*, the strongest benchmark, is the optimal offline mechanism. It is a mechanism which somehow pulls the arm with the highest realization at every time step.
2. *First-best*, the next strongest benchmark, is a mechanism that is optimal in the absence of all IC constraints. In other words, the first-best mechanism has the ability to pull the arms directly as in a MAB instance.
3. *Second-best*, the weakest benchmark, is the optimal mechanism that respects the IC constraints.

We define the *regret* of a mechanism with respect to one of these benchmarks as the difference between the expected reward of the benchmark and the expected reward of the mechanism.

We say arm i *dominates* arm j if $\inf D_i > \mu_j$. If arm j is not dominated by any other arm, we say that arm j is *explorable*. This is similar to the notion of explorability in [7]. If an arm is not explorable, no IC mechanism will ever be able to recommend it.

In fact, there exist instances where unexplorable arms cause any IC mechanism arbitrarily large regret. Consider an instance with just two arms and the first arm dominating the second, meaning the mean of the second distribution is smaller than any reward in the first distribution. However, the second distribution has a small, non-zero probability of achieving a massive reward. No mechanism can ever recommend arm 2, even though the realization of arm 2 may be huge. Thus, the regret of any mechanism will be at least the probability of this large reward times that large reward value times the number of agents. As the number of agents grows, this will be arbitrarily large.

2.1 Policies

Generally, we refer to any mechanism for MMAB as a *policy*. We will exclusively focus on a special type of policy called a *recommendation policy* originally defined in [11].

Definition 1. *A recommendation policy is a mechanism for MMAB in which at time t, the principal recommends an action $a^t \in A$ that is incentive compatible. That is, $\mathbb{E}[R_j - R_i | a^t = a_j] \geq 0$ for all $i, j \in [m]$.*

The goal of the policy is to determine which arm yields the best reward. To this end, the policy will generally need to get agents to explore arms to discover their realizations. It may occasionally be possible to guarantee that an arm is suboptimal without exploring it, using only the realizations of other arms. Whenever a policy explores an arm or otherwise determines that the arm is suboptimal, we say it has *learned* that arm. In general, a policy may attempt to learn an arm, give up and try learning a different arm, and then later come back to learning that arm. However, we will focus on

a special kind of recommendation policy called a *linear policy* that keeps trying to learn an arm once it starts trying.

Definition 2. *A linear policy is a recommendation policy which attempts to learn the arms in a fixed order. This means that once the policy begins the process of learning an arm, it must continue until the arm is learned. Then, it moves on to the next arm in the order.*

Notice that the definition above naturally breaks the procedure of the policy into distinct pieces. This leads to our definition of *phases*.

Definition 3. *The k-phase of a linear policy is the segment of the linear policy in which arm k is attempted to be learned.*

A linear policy is completely determined by the arm ordering and what happens in each phase. Since we will learn the arms in decreasing order of their means, this structure allows us to focus on the process of learning a single arm.

3 Mechanism

3.1 Intuition

Let us first consider the problem with just 2 arms, which has been solved optimally by Kremer et.al. [11]. The first agent will always explore arm 1. The key issue is that an agent who is recommended arm 2 may be hesitant to choose it because its expected payoff is less than that of arm 1. However, agents can be incentivized to pull arm 2 if they consider it likely that the arm has already been explored and found to have a higher realization.

This is achieved by partitioning the support of the distribution D_1 into continuous intervals $\{I_t\}_{t=2}^{T}$. Then, an agent j is recommended arm 2 for the first time if the realization r_1 of arm 1 lies in the interval I_j. After learning arm 2, the mechanism always recommends the better arm. By clever choice of the intervals, we can ensure that the agents know that the expected benefit from the case that arm 2 was already known to be the best arm outweighs the expected loss from the case that the agent is being asked to explore arm 2. This can be viewed as a linear recommendation policy with two phases.

We can now extend this idea to create a linear recommendation policy for k arms. In the k-phase, we imagine combining arms $1, \ldots, k-1$ into a virtual arm whose realization is the that of the best explored arm. We can view the values of this virtual arm as being drawn from the distribution $D_{\max(k-1)}$. If we ensure that it is IC for the agent to choose arm k over this virtual arm, it will also be IC for them to choose arm k over any actual arm among $1, \ldots, k-1$.

The key to making this strategy work is to ensure that each phase is independent in the sense that agents cannot learn about the arms' realizations using information from previous phases. Otherwise, agents might learn specific details about realizations of individual arms. This might allow agents to adopt a better strategy than to treat all arms $1, \ldots, k-1$ as being represented by the virtual arm. We do this by having the lengths of the phases depend only on the distributions and not the realizations. In particular, the k-phase does not necessarily end as soon as arm k is learned. Rather, the k-phase ends only when it is guaranteed that arm k would have been learned regardless of the actual realizations of the arms. This is analogous to the case of mixing times

in Markov chains, as we obfuscate the information learnt during the early part of each phase by sufficiently extending the length of the phase.

A minor issue is that the agent may learn enough information about the current best realization to prefer exploring a different arm, say $k+1$, over exploiting the current best realization. This will happen if the last agent in the k-phase is recommended arm i, telling them that arm k was already explored and found to be worse than i. Conditioned on this new information, it is possible for the expected value of arm i to be less than that of arm $k+1$. To avoid this issue, we simply recommend arm $k+1$ to agents who would prefer exploring it to exploiting the previous arms. However, in order to ensure that future agents do not gain any information from this exploration, we "forget" the realization of the reward r_{k+1}, and proceed as if we did not learn anything from the exploration.

3.2 Mechanism

Our mechanism, the *independent phases mechanism* (IPM), proceeds by having a k-phase for each arm k such that $\mu_1 \geq \cdots \geq \mu_m$.

0. Remove all unexplorable arms, and relabel the arms $1, 2, \ldots, m$ such that $\mu_1 \geq \cdots \geq \mu_m$.
1. Recommend arm 1 to agent 1 and learn the realization r_1.
2. For each $k \in \{2, \ldots, m\}$, begin the k-phase for the agents $j \in \{t_k \ldots, T_k\}$ where $t_2 = 2$, and $t_k = T_{k-1} + 1$ for all other k.
   (a) Find the intervals $\{I_t\}_{t=t_k}^{T_k}$ guaranteed by Theorem 2, and find $i = \arg\max_{i'<k} r_{i'}$.
   (b) Determine if there exists some $t \in \{t_k, \ldots, T_k\}$ such that $r_i \in I_t$. If no such t exists, then for each agent j who arrives in phase k, recommend arm i.
   (c) Otherwise, there is a unique agent t such that $r_i \in I_t$. Then, for each agent j who arrives in phase k:
      (i) If $j < t$, recommend arm i to the agent.
      (ii) If $j = t$, recommend arm k to the agent, and learns r_k.
      (iii) If $j > t$, r_k is known. If $r_k > r_i$, recommend arm k. Otherwise, recommend $\arg\max\{r_i, \mu_{k+1}\}$. (In the case of a tie, give priority to the arm with the smallest index.)
3. After all the phases end, recommend arm $i = \arg\max_{i' \leq m} r_{i'}$

4 Analysis

In this section, we will prove that given m explorable arms, it is possible to compute the intervals $\{I_t\}_{t=t_k}^{T_k}$ referenced in the mechanism above, for which the IPM achieves constant regret. It is important to note that these intervals can be pre-computed if the distributions are known beforehand, which makes our mechanism easy to implement. It will also be clear from the proof of Theorem 2, that if the distributions are discrete, then computing the intervals takes polynomial time.

We will assume for the rest of this paper that the distributions are all continuous to make our proofs easier. It can be verified that with minor changes, they hold true even for non-continuous distributions. In this section, we will prove the following theorem.

Theorem 1 (Main Result). *IPM is Bayesian incentive-compatible (BIC) and achieves constant regret with respect to the second best. Additionally, if all arms are explorable, then the mechanism achieves constant regret with respect to the offline best.*

We will first prove that IPM is BIC, and then later analyze its regret properties.

Theorem 2. *For each k-phase, there exists a collection $\{I_t\}_{t=t_k}^{T_k}$, such that the above mechanism is BIC.*

Proof. As defined in the mechanism, let i be the index of the best known arm:

$$i = \arg\max_{i'<k} r_{i'}$$

We will show that there is a partition $(-\infty, s^{t_k}], (s^{t_k}, s^{t_k+1}], \ldots, (s^{T_k-1}, s^{T_k}]$ that satisfies the requirements. The values are defined as follows

$$s^{t_k} = \sup\left\{ s \le \sup D_k : \mathbb{E}\left[\max_{i'<k} R_{i'} \middle| \max_{i'<k} R_{i'} \le s\right] \le \mu_k \right\}$$

For our proofs, it will be convenient to use the following equivalent definition for s^{t_k}

$$s^{t_k} = \sup\left\{ s \le \sup D_k : \int_{\max_{i'<k} R_{i'} \le s} (\max_{i'<k} R_{i'} - \mu_k) dD \le 0 \right\}$$

In Observation 1, we will show that our assumption that all arms are explorable would be violated if this set was empty. Therefore, $s^{t_k} > -\infty$.

Now, with s^t defined for $t \ge t_k$, we define s^{t+1} as follows

$$s^{t+1} = \sup\left\{ s \le \sup D_k : \int_{s^t < \max_{i'<k} R_{i'} \le s} (\max_{i'<k} R_{i'} - \mu_k) dD \le \int_{\substack{\max_{i'<k} R_{i'} \le s^t \\ R_k > \max_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD \right\}$$

Notice that since $s^t > -\infty$, $s^{t+1} > -\infty$ as well. If $s^t = \infty$, then we let $s^{t+1} = \infty$ as well, and we define

$$T_k = \inf\{t \ge 1 : s^t = s^{t+1}\}$$

Also, let e be the first agent who explores arm k in the k-phase, if any. So $r_i \in (s^{e-1}, s^e]$ where $s^{e-1} = -\infty$ for $e = t_k$. We will now prove that these intervals satisfy the IC constraints. In the k-phase, each recommendation is to pull arm i, arm k, or arm $k+1$. These recommendations are shown to be IC by Lemma 4.

The main idea behind our definition for s^{t_k} is to compute exactly what information to reveal to agent t_k so the agent is indifferent between exploiting the previous best arm and exploring arm k. In particular, if agent t_k were to choose exploiting the max arm given that its value is at most s, they would attain $\mathbb{E}\left[\max_{i'<k} R_{i'} \middle| \max_{i'<k} R_{i'} \le s\right]$ reward in expectation. On the other hand, since the recommendation only constrains the values of the max arm and gives no information about arm k by independence, the agent's expected reward for choosing arm k is exactly μ_k, despite

having additional information about the max arm. Thus, by choosing $s = s^{t_k}$, agent t_k will be indifferent and so exploring arm k will be incentive compatible. Then, the idea for defining s^{t+1} follows the same reasoning. We give the agents just enough information so that exploiting seems equally desirable to exploring arm k. Furthermore, exploring future arms will be undesirable since their means are smaller than that of μ_k, up to a special case we take care of later.

Observation 1. *For a phase k, if $s^{t_k} = -\infty$, then arm k is unexplorable.*

Proof. By definition of s^{t_k}, the event $s^{t_k} = -\infty$ occurs if and only if

$$\inf \left\{ \max_{i'<k} R_{i'} \right\} > \mu_k.$$

Note that

$$\inf \left\{ \max_{i'<k} R_{i'} \right\} = \max_{i'<k} \left\{ \inf R_{i'} \right\}.$$

Therefore, if $s^{t_k} = -\infty$, there must be some $\ell < k$ such that

$$\inf R_\ell > \mu_k$$

But then, we have by definition that arm ℓ dominates arm k. Hence, arm k is unexplorable.

Note that in the mechanism, since all unexplorable arms will be removed, the case in Observation 1 won't occur and all intervals will exist.

Observation 2. $s^{t_k} \geq \mu_k$

Proof. Just consider plugging in $s = \mu_k$ in the definition of s^{t_k}. This always gives a small enough conditional expectation and so $s = \mu_k$ is a possible candidate, implying the supremum will be at least this value.

Now, in phase k, let us consider an agent t. We want to show that all possible recommendations to agent t will be IC. We will also let e be the agent who first explores arm k in phase k.

Lemma 1. *In the k-phase, if an agent t is recommended arm i, the recommendation is IC.*

Proof. There are two cases in which the agent t would be recommended arm i. We find the expectation in each case and then apply the law of total expectation to get the result. Note that being recommended arm i in the k-phase implicitly implies that $t \neq e$. Thus, conditioned on this event we have $\Pr[t = e] = 0$. The common idea in both cases is that knowing $t \neq e$ and knowing i was recommended implies that i had the highest realization so far. Hence, conditioning on the recommendation is equivalent to conditioning on the realization of arm i being the largest of the previous arms' realizations. Then, depending on the case, we get further bounds on arm i's realization and this allows us to bound the utility for the agents.

Case 1: $(t < e)$ The principal has not recommended arm k to anyone yet, and arm i is the best arm explored so far. By definition of our intervals, this then means that $r_i > s^t$. Otherwise, t or a previous agent would have explored arm k. Hence, the agent can deduce that arm i has the largest realization so far and deduce a lower bound on r_i. Taking this information into the expectation yields the following inequalities. Overall,

$$\begin{aligned}\mathbb{E}[R_i|a^t = a_i, t < e] &= \mathbb{E}[R_i|\max_{i'<k} R_{i'}, R_i > s^t] \\ &\geq \mathbb{E}[R_\ell|a^t = a_i, t < e] \ \forall \ \ell < k\end{aligned}$$

Moreover, we can also conclude from continuity that

$$\begin{aligned}\mathbb{E}[R_i|a^t = a_i, t < e] &= \mathbb{E}[R_i|R_i = \max_{i'<k} R_{i'}, R_i > s^t] \\ &\geq \mathbb{E}[R_i|\max_{i'<k} R_{i'}, R_i > s^{t_k}] \\ &\geq \mathbb{E}[R_i|\max_{i'<k} R_{i'}, R_i \geq \mu_k] \\ &\geq \mu_k \\ &\geq \mu_\ell, \ \forall \ \ell \geq k\end{aligned}$$

$$\therefore \mathbb{E}[R_i - R_\ell|a^t = a_i, t < e] \geq 0 \ \forall \ \ell \in [m]$$

Here, we used the fact that the intervals are increasing and that conditional expectations are monotonic here, together with Observation 2.

Case 2: $(t > e)$ The principal has recommended arm k to a previous agent and found that arm i has the highest realization among all arms $\ell \leq k$. Furthermore, we know that $r_i \geq \mu_{k+1}$. Thus, we have,

$$\begin{aligned}\mathbb{E}[R_i|a^t = a_i, t > e] &= \mathbb{E}[R_i|R_i = \max_{i'\leq k} R_{i'}, R_i \leq s^{t-1}, R_i \geq \mu_{k+1}] \\ &\geq \mathbb{E}[R_\ell|R_i = \max_{i'\leq k} R_{i'}, R_i \leq s^{t-1}, R_i \geq \mu_{k+1}] \\ &= \mathbb{E}[R_\ell|a^t = a_i, t > e] \ \forall \ \ell \leq k\end{aligned}$$

$$\therefore \mathbb{E}[R_i - R_\ell|a^t = a_i, t > e] \geq 0 \ \forall \ \ell \leq k$$

Since in phase k, the planner's recommendations never depend on the realizations of any arm $\ell > k$, their expected values from the agent's perspective do not deviate from μ_ℓ. Hence, we have

$$\begin{aligned}\mathbb{E}[R_i|a^t = a_i, t > e] &= \mathbb{E}[R_i|R_i = \max_{i'\leq k} R_{i'}, R_i \leq s^{t-1}, R_i > \mu_{k+1}] \\ &\geq \mu_{k+1} \\ &= \mathbb{E}[R_{k+1}|a^t = a_i, t > e] \\ &\geq \mathbb{E}[R_\ell|a^t = a_i, t > e] \ \forall \ \ell > k\end{aligned}$$

$$\therefore \mathbb{E}[R_i - R_\ell|a^t = a_i, t > e] \geq 0 \ \forall \ \ell > k$$

Hence, i has a higher expected value than any other arm:

$$\therefore \mathbb{E}[R_i - R_\ell|a^t = a_i] \geq 0 \ \forall \ \ell \in [m]$$

In either case, we see that the recommendation is incentive compatible.

Lemma 2. *In phase k, if an agent t is recommended arm k, the recommendation is IC.*

Proof. Let us first consider the arms $\ell > k$. There are two cases in which an agent t would be recommended arm k.

Case 1: $(t = e)$ The principal is recommending arm k to agent t for the first time. Then,

$$\mathbb{E}[R_k - R_\ell | a^t = a_k, t = e] = \mu_k - \mu_\ell \geq 0$$

In other words, knowing the recommendation and knowing the place in line does not tell the agents anything about the expected rewards of arm k and arm ℓ. Specifically, the planner has only observed realizations of other arms since these two haven't been explored yet by assumption. Hence, independence of the arms implies those realizations observed by the planner don't affect the rewards of arm k and arm ℓ.

Case 2: $(t > e)$ The principal recommended arm k to a prior agent, and found that it is the best arm. Then,

$$\mathbb{E}[R_k - R_\ell | a^t = a_k, t > e] = \mathbb{E}[R_k | R_k = \max_{\ell \leq k} R_\ell, R_k \geq \mu_{k+1}] - \mu_\ell \geq \mu_{k+1} - \mu_\ell \geq 0$$

Therefore, it is clear that

$$\mathbb{E}[R_k - R_\ell | a^t = a_k] \geq 0 \ \forall \ \ell > k$$

Now, we consider past arms $\ell < k$. We begin with the case where $e = 1$. In this case, if the agent is recommended arm k, they know that they are exploring. But, they also know that $\max_{\ell < i} r_\ell \leq s^{t_k}$. By our choice of s^{t_k}, we know that

$$\int_{\max_{i' < k} R_{i'} \leq s^{t_k}} (\max_{i' < k} R_{i'} - \mu_k) dD = 0$$

It is also true that for all $\ell < k$,

$$\int_{\max_{i' < k} R_{i'} \leq s^{t_k}} (R_\ell - \max_{i' < k} R_{i'}) dD \leq 0$$

$$\therefore \int_{\max_{i' < k} R_{i'} \leq s^{t_k}} R_\ell dD \leq \int_{\max_{i' < k} R_{i'} \leq s^{t_k}} R_k dD$$

$$\therefore \mathbb{E}[R_\ell | \max_{i' < k} R_{i'} \leq s^{t_k}] \leq \mathbb{E}[R_k | \max_{i' < k} R_{i'} \leq s^{t_k}]$$

$$\therefore \mathbb{E}[R_\ell | a^t = k, e = 1] \leq \mathbb{E}[R_k | a^t = k, e = 1]$$

The LHS here denotes the utility that the agent t derives from pulling arm ℓ, and the RHS denotes the utility that the agent t derives from pulling the recommended arm k. Therefore, the agent would not prefer to pull an arm $\ell < k$ over arm k.

If $e > 1$, the agent no longer knows with certainty whether they are exploring or exploiting. By our choice of s^e and continuity, we know that

$$\int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} \left(\max_{i'<k} R_{i'} - \mu_k \right) dD \leq \int_{\max_{i'<k} R_{i'} \leq s^{e-1}, R_k > \max_{i'<k} R_{i'}} (R_k - \max_{i'<k} R_{i'}) dD$$

Now, by rearranging, we get

$$\int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} \max_{i'<k} R_{i'} dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} \max_{i'<k} R_{i'} dD \leq \int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} R_k dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} \mu_k dD$$

$$\therefore \mathbb{E}[\max_{i'<k} R_{i'} | a^t = k, e > 1] \leq \mathbb{E}[R_k | a^t = k, e > 1]$$

The LHS denotes the utility that the agent derives from pulling a virtual arm with the distribution $\max_{i'<k} D_{i'}$ and the RHS here denotes the utility that the agent t derives from following the recommendation to pull arm k. Note that for any $\ell < k$, it is true that

$$\int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} R_\ell dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} R_\ell dD \leq \int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} \max_{i'<k} R_{i'} dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} \max_{i'<k} R_{i'} dD$$

where the LHS here is the utility that the agent derives from ignoring the recommendation and pulling arm $\ell < k$.

$$\therefore \int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} R_\ell dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} R_\ell dD \leq \int_{\substack{\max_{i'<k} R_{i'} \leq s^{e-1} \\ R_k > \max_{i'<k} R_{i'}}} R_k dD + \int_{s^e \leq \max_{i'<k} R_{i'} < s^{e+1}} R_k dD$$

$$\therefore \mathbb{E}[R_\ell | a^t = k, e > 1] \leq \mathbb{E}[R_k | a^t = k, e > 1]$$

This proves that the agent will always follow the recommendation to pull the arm k.

Lemma 3. *In phase k, if an agent t is recommended arm $k+1$, the recommendation is IC.*

Proof. The mechanism recommends arm $k+1$ to an agent t during phase k if and only if there exists some e with $r_i \in I_e$, such that $t > e$, and $k+1 = \arg\max\{r_i, r_k, \mu_{k+1}\}$. Consider some $\ell \leq k$. Then,

$$\begin{aligned}
\mathbb{E}[R_\ell | a^t = k+1] &= \mathbb{E}[R_\ell | \max_{i' \leq k} R_{i'} \leq \mu_{k+1}, t > e] \\
&\leq \mathbb{E}[\max_{i' \leq k} R_{i'} | \max_{i' \leq k} R_{i'} \leq \mu_{k+1}, t > e] \\
&\leq \mu_{k+1} \\
\therefore \mathbb{E}[R_{k+1} - R_\ell | a^t = k+1] &\geq 0 \ \forall \ \ell \leq k
\end{aligned}$$

Hence, agent t knows that $r_i = \max_{i' < k} r_{i'}$

Since in phase k, the planner's recommendations never depend on the realizations of any arm $\ell > k$, their expected values from the agent's perspective do not deviate from μ_ℓ. Therefore,

$$\begin{aligned}\mathbb{E}[R_{k+1} | a^t = k+1] &= \mu_{k+1} \\ &\geq \mu_\ell, \quad \forall\ \ell > k+1\end{aligned}$$

$$\therefore \mathbb{E}[R_{k+1} - R_\ell | a^t = k+1] \geq 0,\ \forall\ \ell < k$$

Hence, the recommendation is always IC.

Lemma 4. *IPM is BIC.*

Proof. First, note that the first agent will always be willing to pull an arm with the highest expected reward, so step one of IPM is BIC. Now, note that the total number of intervals is fixed in advanced. This is because we use the distributions $D_{\max(k-1)}$ and D_k to compute the intervals for the k-phase and these distributions do not depend on the realizations seen so far. Since the agents in a phase and the length of the phase are fixed ex-ante, the agents cannot learn anything about the actions or realizations in previous phases from their given recommendation and a description of the mechanism. Suppose agent t is in the k-phase, there are several cases to consider.

Case 1: Suppose the recommendation is arm $i < k$. Then, the agent knows arm i has the best realization so far and they are being asked to exploit. The only bounds they can derive on the realization is $r_i \notin I_t$ and possibly $r_i \in I_{t'}$ for some $t' < t$ if arm k has already been explored. Given this information the agent may have, Lemma 1 shows that in either case, this recommendation is BIC.

Case 2: Suppose the recommendation is arm k. Then, the agent knows they are being asked to explore arm k or arm k was the best seen so far and they are asked to exploit it. The only information the agent can then derive is that $r_i \in I_t$, or $r_i \in I_{t'}$ for some $t' < t$ and $r_k > r_i$ for some $i < k$ but they do not know which of these events occurred. Given this information, Lemma 2 implies the recommendation is BIC.

Case 3: Suppose the recommendation is arm $k+1$. Then, they only know that arm k has already been explored and that $\max\{r_i, r_k\} < \mu_{k+1}$. Thus, Lemma 3 shows this recommendation is then BIC.

To complete the proof of Theorem 1, we now need to show that the IPM mechanism attains constant regret with respect to the number of agents. Note that the regret is not constant in the number of arms; indeed this is impossible in general because even in the case where the rewards are i.i.d., all arms must be explored.

Theorem 3. *IPM achieves constant regret with respect to the second-best mechanism. IPM additionally achieves constant regret with respect to the offline-best mechanism if all arms are explorable.*

Proof. We already know from Observation 1 that the intervals $\{I_t\}_{t=t_k}^{T_k}$ are well-defined since all unexplorable arms are removed at the start of the mechanism. Moreover, note that if $r_i \leq \sup D_k$, then $r_i \in (s^{t-1}, s^t]$ for some $t \in \{t_k, T_k\}$. Therefore, the realization of arm k will be determined by

the mechanism if it has non-zero probability of being higher than r_i at the start of the k-phase. In any case, we see that all explorable arms are learned by IPM.

Therefore, IPM will learn all arms and after T_m agents, all future agents will be recommended the best explored arm. Therefore, after T_m agents, all agents are recommended the best arm. So, if OPT is the best realization among all (explorable) arms, then

$$\mathbb{E}[\text{OPT-BEST}(n)] - \mathbb{E}[\text{IPM}(n)] \leq n \cdot \text{OPT} - (n - T_m) \cdot \text{OPT} = T_m \cdot \text{OPT}$$

where OPT-BEST(n) is the optimal-best mechanism on all explorable arms. So, if we can show that T_m is a constant, we will have shown that the regret is constant with respect to the second-best mechanism. Moreover, when all arms are explorable, it will follow that IPM achieves constant regret with respect to the optimal-best.

Note that

$$T_m = \sum_{k=1}^{m} (T_k - t_k + 1)$$

We will prove that for each $k \in [m]$, $T_k - t_k + 1$ is a finite constant that does not depend on the number of agents. For a fixed k, note that from our definitions,

$$\int\limits_{s^t \leq \max\limits_{i'<k} R_{i'} < s^{t+1}} (\max_{i'<k} R_{i'} - \mu_k) dD = \int\limits_{\substack{\max\limits_{i'<k} R_{i'} \leq s^t \\ R_k > \max\limits_{i'<k} R_{i'}}} (R_i - \max_{i'<k} R_{i'}) dD$$

for all $t_k \leq t \leq T_k$.
Note that

$$\sum_{t=t_k}^{T_k} \left(\int\limits_{s^t \leq \max\limits_{i'<k} R_{i'} < s^{t+1}} (\max_{i'<k} R_{i'} - \mu_k) dD \right) \leq \mu_{\max}(k) - \mu_k$$

where $\mu_{\max}(k) = \mathbb{E}[\max\limits_{i'<k} R_{i'}]$ and also note that

$$\int\limits_{\substack{\max\limits_{i'<k} R_{i'} \leq s^t \\ R_k > \max\limits_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD \leq \int\limits_{\substack{\max\limits_{i'<k} R_{i'} \leq s^{t+1} \\ R_k > \max\limits_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD$$

for all $t_k \leq t < T_k$.

$$\therefore (T_k - t_k + 1) \left(\int\limits_{\substack{\max\limits_{i'<k} R_{i'} \leq s^{t_k} \\ R_k > \max\limits_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD \right) \leq \sum_{t=t_k}^{T_k} \left(\int\limits_{\substack{\max\limits_{i'<k} R_{i'} \leq s^t \\ R_k > \max\limits_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD \right) \leq \mu_{\max}(k) - \mu_k$$

$$\therefore (T_k - t_k + 1) \leq \frac{\mu_{\max}(k) - \mu_k}{\int\limits_{\substack{\max_{i'<k} R_{i'} \leq s^{t_k} \\ R_k > \max_{i'<k} R_{i'}}} (R_k - \max_{i'<k} R_{i'}) dD}$$

Clearly, this is a constant ($< \infty$, since all arms are explorable) which does not depend on the number of agents. This proves that the mechanism achieves constant regret.

With Theorem 2 and Theorem 3, we have now completed the proof of Theorem 1.

5 Achieving First-Best

We give necessary and sufficient conditions for the first-best and second-best outcomes to coincide. In some sense, these are conditions under which the IC constraints do not increase the difficulty of the problem. We will ultimately show that our mechanism achieves first-best in such cases.

First, we derive the first-best mechanism. In this section, we again assume for simplicity that all distributions are continuous and do not contain point masses.

Lemma 5. *Suppose we have a MMAB instance with distributions D_i with means μ_i and suprema b_i. The following mechanism achieves the first-best outcome:*

1. *Order the arms in decreasing order of b_i, with μ_i as a tiebreaker*[1]
2. *Set $r_{max} := -\infty$ and $a_{max} := Null$*
3. *For $t \in [m]$:*
   (a) *If $r_{max} \geq b_t$, pull a_{max}*
   (b) *Otherwise, pull a_t and then if $r_j > r_{max}$, update $r_{max} = r_j$ and $a_{max} = a_j$.*
4. *Pull a_{max} at all remaining times.*

Proof. Any asymptotically optimal mechanism must determine the optimal arm. Define the exploration stage of a mechanism to be the set of time periods before it knows the optimal arm, and the exploitation stage to be the set of time periods after the mechanism knows the optimal arm. Any optimal mechanism must always pull the optimal arm during the exploitation stage (incurring 0 regret), so to minimize regret it is sufficient to minimize regret during the exploration stage.

Note that for a mechanism to determine the optimal arm, for all arms a_i, it has to learn all arms. In particular, pulling each arm once is sufficient to learn the optimal arm. Note also that there is nothing to be gained by pulling an arm for a second time during the exploration stage: If the arm is in fact optimal, there is no gain relative to ending the exploration stage 1 time step earlier, and if the arm is suboptimal there is some regret incurred. Therefore, to derive a first-best mechanism, it is sufficient to find an ordering of the arms in the exploration stage to minimize the regret incurred from exploring arms unnecessarily.

We argue that our mechanism is first-best with a greedy-stays-ahead style argument. We show inductively that our mechanism never explores an arm that a first-best mechanism does not also explore during the exploration stage. Then, since exploring a particular arm incurs the same regret regardless of the timing of that exploration and all arm values are independent, this shows that our mechanism is first-best almost surely.

[1] Any tie-breaking method leads to an optimal mechanism, but this particular tie-breaking causes the mechanism to be IC in the special case considered in Theorem 4.

First, we claim that with probability 1, a first-best mechanism must explore arm a_1. This is because since D_1 has the highest supremum, no other arm has a non-zero probability of attaining a realization at least b_1.

Now, we argue that, with probability 1, the k-th arm our mechanism explores in the exploration stage is also explored in any first-best mechanism, given that all arms previously explored are also explored by any first-best mechanism. If we explore a_k, then every explored arm must have a realization less than b_k. Additionally, since the distributions are continuous and all unexplored arms have supremum at most b_k, no unexplored arm has a non-zero probability of achieving at least b_k. Hence, with probability 1, any first-best mechanism must explore a_k.

Next, we give a necessary condition for a mechanism to achieve first-best. We essentially show that any first-best mechanism must order the exploration as in the mechanism from Lemma 5.

Lemma 6. *Consider an arm j, and let i and k be the smallest and largest indices (respectively) such that $b_i = b_j = b_k$. Then in any optimal mechanism, arm j must never be recommended during times $\{1, 2, \ldots, i-1\}$. Further, if arm j is ever recommended, then with probability 1 it must be recommended exactly once in the time steps $\{i, i+1, \ldots, k\}$.*

Proof. First, suppose for contradiction that a first-best mechanism M recommends arm j at time time $t < i$. Then by pigeonhole principle, there must exist some arm $\ell < i$ that has not yet been recommended by M, and further we know $b_\ell > b_j$. Since M is first-best, we can assume that none of the arms that have been sampled before time t attained a value greater than b_j. Since we know $b_\ell > b_j$, there is a non-zero probability that $r_\ell > r_j$, and hence since M is optimal and must determine the optimal arm, it must recommend arm ℓ at some time t' in the future. Consider a mechanism M' identical to M except that M' recommends arm ℓ rather than arm j at time t and then at time t', M' recommends arm ℓ if $r_\ell > b_j$ and arm j otherwise.

In the case that $r_\ell \le b_j$, the expected rewards of M and M' are identical since the distributions are all independent and hence the order they are sampled in doesn't matter. In the case that $r_\ell > b_j$, which occurs with strictly positive probability as noted before, M and M' differ only in their recommendation at time t'. At this time, M' attains reward of $r_\ell > b_j \ge r_j$ while M attains a reward of r_j. Hence, M' is strictly better than M, contradiction.

Next, suppose for contradiction that a first-best mechanism M never recommends arm j in times $\{1, \ldots, k\}$. Since there are only $k-1$ other arms ℓ with $b_\ell \ge b_j$, M must recommend some arm ℓ with $b_\ell < b_j = b_k$ at a time in $\{1, \ldots k\}$, which is not possible by previous case.

Finally, suppose for contradiction that a first-best mechanism M recommends arm j more than once in times $\{1, \ldots, k\}$. Let t be the second time j is recommended. Since M is first-best, we can again assume that none of the arms that have been sampled before time t attained a value greater than b_j. By pigeonhole, there must be some arm $\ell \in [k]$ that is never recommended in $\{1, \ldots, k\}$. Also, since the distributions are continuous, $r_j < b_j$ with probability 1. Therefore, since $b_\ell \ge b_j$, there is some strictly positive probability that $R_\ell > r_j$ for $R_\ell \sim D_\ell$, in which case arm ℓ would be optimal. Since M is first-best and must determine the optimal arm, it must recommend arm ℓ at some time t' in the future.

This is now exactly identical to the first case: Consider a mechanism M' identical to M except that M' recommends arm ℓ rather than arm j at time t and then at time t', M' recommends arm ℓ if $r_\ell > b_j$ and arm j otherwise. In the case that $r_\ell \le b_j$, the expected rewards of M and M' are identical since the distributions are all independent and hence the order they are sampled in doesn't matter. In the case that $r_\ell > b_j$, which occurs with strictly positive probability as noted

before, M and M' differ only in their recommendation at time t'. At this time, M' attains reward of $r_\ell > b_j \geq r_j$ while M attains a reward of r_j. Hence, M' is strictly better than M, contradiction.

Now, we apply our lemmas to prove the main result of this section. We show that the mechanism from Lemma 5 is incentive-compatible under our assumptions and use Lemma 6 to show that these assumptions are necessary for any IC mechanism to achieve first-best.

Theorem 4. *Suppose we have arms m arms labeled $\{1, 2, \dots m\}$ with values drawn from distributions D_i where the suprema of the distributions b_i are descending (with their means μ_i as a tiebreaker). We claim there exists an IC mechanism achieving first-best if and only if both of the following hold:*

- *The means μ_i of the distributions are descending: $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$.*
- *The distributions satisfy the property that, for all $k \in \{2, \dots, m\}$, for all $i \in [k-1]$, $\mathbb{E}_{r_i \sim D_i}[r_i | r_i < b_k] \leq \mu_k$.*

Proof. ($\Rightarrow$) We show that the first-best mechanism from Lemma 5 is incentive compatible in this setting. The incentive compatibility of the mechanism is straightforward. Consider the recommendation to agent k. We have two cases for the recommendation: it is a_k or it is a_j for some $j < k$. If the recommendation is a_j for $j < k$, the principal is recommending to exploit arm that it knows to be optimal among all arms, so it is clearly IC to follow the recommendation. If the recommendation is a_k, then the agent knows all a_i for $i < k$ have been explored and have realization less than b_k. Further, by condition 2, the expected values of these arms (conditional on the known information) are no more than μ_k. Finally, the agent knows that no arm with index greater than k has been explored, and the means of those distributions are also no more than μ_k. Therefore, the expected value of a_k is at least as high as that of any other arm, and so it is IC for the agent to follow the recommendation.

($\Leftarrow$) Suppose the first condition does not hold. Then we must have some arms a_i and a_j where $b_i > b_j$ (and therefore $i < j$) but $\mu_i < \mu_j$. Let k be the largest index of an arm with $b_k = b_i$. We know from Lemma 6 that if a first-best mechanism explores explores a_i, it must explore a_i before a_j and recommend a_i exactly once within the first k agents (with probability 1). Crucially, the first-best mechanism must NOT explore a_j within the first i agents. Therefore, when an agent $t \leq i$ is recommended a_i, they know that with probability 1, both a_i and a_j unexplored. But then it is strictly better for the agent to pull a_j and the mechanism cannot be IC. Therefore, the first condition is necessary for an IC mechanism achieving first-best to exist. Suppose the second condition does not hold but the first does. Let i and j be indices of arms that violate the property, that is $i < j$ but $E_{r_i \sim D_i}[r_i | r_i < b_j] > \mu_j$. Consider the case where no arm explored before a_j obtains a realization of at least b_j, and so any first-best mechanism must explore a_j. Let k be the largest index of an arm with $b_k = b_j$. We know from Lemma 6 that (again, with probability 1) a first-best mechanism must recommend a_j exactly once to the first k agents if it recommends a_j at all. In particular, this must occur for any value of arm i with $r_i < b_j$. Therefore, if an agent $t \leq k$ is recommended a_j, the know almost surely that they are the first agent to be recommended a_j. But then the recommendation is not IC since $E_{r_i \sim D_i}[r_i | r_i < b_j] > \mu_j$. Hence, both conditions are necessary for the existence of an IC mechanism than achieves first-best.

Observation 3. *Our mechanism,* IPM*, is equivalent to the first-best mechanism in this case.*

Proof. Notice that in this case, for each k-phase, s^{t_k} is well defined, and $s^{t_k} = b_k$. Furthermore, $s^{t_k} = s^{t_k+1}$. So, for each arm $k \in \{2, \ldots, m\}$, there is a single interval in each phase, where the arm k is recommended to agent j if and only if $r_{\max} < b_k$. Therefore, IPM is identical to the mechanism described in Lemma 5.

Remark 1. There are some salient special cases of distributions for which the optimal IC mechanism can achieve the first-best outcome. One such case is when every distribution D_i has the same mean. Another is the setting of nested intervals, where each D_i is a uniform distribution over some interval, each interval is a subset of the previous one, and the means are non-increasing.

6 Future Work

It may be possible to extend our results to other settings. One promising line of enquiry would be to extend our work to the case of heterogeneous agents. [7] has an efficient, constant regret mechanism for this problem in the case of discrete distributions. The techniques from this paper may allow us to extend that to continuous distributions as well. Another direction would be to extend our mechanism to settings where the reward distributions of the arms are correlated rather than independent. Our technique of independent phases, which allows us to conceal all information about what is learned in previous phases from the agents, could be key in this setting. In particular, it might allow for some sort of reduction from correlated to independent distributions.

References

[1] Guy Aridor et al. *Competing Bandits: The Perils of Exploration Under Competition*. 2021. arXiv: 2007.10144 [cs.GT].

[2] Dirk Bergemann and Stephen Morris. "Information Design: A Unified Perspective". In: *Journal of Economic Literature* 57.1 (Mar. 2019), pp. 44–95. DOI: 10.1257/jel.20181489. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20181489.

[3] Matteo Castiglioni et al. "Online Bayesian Persuasion". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 16188–16198. URL: https://proceedings.neurips.cc/paper/2020/file/ba5451d3c91a0f982f103cdb Paper.pdf.

[4] Bangrui Chen, Peter Frazier, and David Kempe. "Incentivizing Exploration by Heterogeneous Users". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 798–818. URL: http://proceedings.mlr.press/v75/chen18a.html.

[5] Lee Cohen and Yishay Mansour. "Optimal Algorithm for Bayesian Incentive-Compatible Exploration". In: *CoRR* abs/1810.10304 (2018). arXiv: 1810.10304. URL: http://arxiv.org/abs/1810.10304.

[6] Peter Frazier et al. "Incentivizing Exploration". In: *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. EC '14. Palo Alto, California, USA: Association for Computing Machinery, 2014, pp. 5–22. ISBN: 9781450325653. DOI: 10.1145/2600057.2602897. URL: https://doi.org/10.1145/2600057.2602897.

[7] Nicole Immorlica et al. "Bayesian Exploration with Heterogeneous Agents". In: *CoRR* abs/1902.07119 (2019). arXiv: 1902.07119. URL: http://arxiv.org/abs/1902.07119.

[8] Nicole Immorlica et al. *Incentivizing Exploration with Selective Data Disclosure*. 2020. arXiv: 1811.06026 [cs.GT].

[9] Emir Kamenica. "Bayesian Persuasion and Information Design". In: *Annual Review of Economics* 11.1 (2019), pp. 249–272. DOI: 10.1146/annurev-economics-080218-025739. eprint: https://doi.org/10.1146/annurev-economics-080218-025739. URL: https://doi.org/10.1146/annurev-economics-080218-025739.

[10] Emir Kamenica and Matthew Gentzkow. "Bayesian Persuasion". In: *American Economic Review* 101.6 (Oct. 2011), pp. 2590–2615. DOI: 10.1257/aer.101.6.2590. URL: https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

[11] Ilan Kremer, Yishay Mansour, and Motty Perry. "Implementing the Wisdom of the Crowd". In: vol. 122. June 2013, pp. 605–606. DOI: 10.1145/2482540.2482542.

[12] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. "Bayesian Incentive-Compatible Bandit Exploration". In: *CoRR* abs/1502.04147 (2015). arXiv: 1502.04147. URL: http://arxiv.org/abs/1502.04147.

[13] Yishay Mansour et al. *Bayesian Exploration: Incentivizing Exploration in Bayesian Games*. 2021. arXiv: 1602.07570 [cs.GT].

[14] Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva. "Crowdsourcing Exploration". In: *Management Science* 64.4 (2018), pp. 1727–1746. DOI: 10.1287/mnsc.2016.2697. eprint: https://doi.org/10.1287/mnsc.2016.2697. URL: https://doi.org/10.1287/mnsc.2016.2697.

[15] Mark Sellke and Aleksandrs Slivkins. *The Price of Incentivizing Exploration: A Characterization via Thompson Sampling and Sample Complexity*. 2021. arXiv: 2002.00558 [cs.GT].

[16] Aleksandrs Slivkins. "Incentivizing Exploration via Information Asymmetry". In: *XRDS* 24.1 (Sept. 2017), pp. 38–41. ISSN: 1528-4972. DOI: 10.1145/3123744. URL: https://doi.org/10.1145/3123744.

[17] Siwei Wang and Longbo Huang. *Multi-armed Bandits with Compensation*. 2018. arXiv: 1811.01715 [cs.LG].