Incentivizing Exploration with Heterogeneous Utilities

January 15, 2018

1 Introduction

In this paper, we study incentivizing exploration with heterogeneous agent preferences. In this problem, arms have unknown multivariate attributes, and agents have heterogeneous linear utility functions that map these attribute vectors onto utilities. Agents see noisy observations of attributes of arms pulled by all previous agents, and estimate an arms' attribute vector by the simple average of these observations. Agents are selfish, and pull the arm with the largest estimated utility summed with an optional arm-specific incentivizing payment chosen by the principle. We study strategies for choosing such incentive payments that seek to maximize the total utility derived by agents, subject to a limitation on the total incentive payment. To accomplish this goal, a strategy must induce sufficient exploration to reveal arms' attributes, while still letting agents select myopically and according to their preferences sufficiently often that high-utility arms are chosen and incentive payments are kept small.

Our problem setting models online review aggregators like Amazon, Yelp, and Tripadvisor that host crowdsourced reviews. Users of these websites wish to use the reviews hosted there to choose the product / restaurant / vacation (generically referred to as an "item") that is best according to their preferences. These reviews provide not just cardinal feedback, but also a description of items' attributes that a user may consider together with their personal preferences to select their preferred item. An item with few reviews might have inaccurate attribute estimates, leading users to avoid it even though it may actually be their best choice. Without incentives, this situation may persist and decrease welfare for the platform's user base. By offering incentives, either through price reductions by Amazon or coupons from Yelp or Tripadvisor, a platform may induce more reviews of unexplored items and provide more value over the long term.

Our problem setting also applies to crowd science platforms like eBird [1, 7]. EBird guides birding enthusiasts through a website to explore and report their findings to the birding community. Each user report contains information about when, where and how they go birding and what birds they see and hear. EBird may wish to incentivize enthusiasts to explore less-explored birding locations

and provide more accurate reports on these locations. Each enthusiast may have different preference over a location's attributes such as the diversity of bird species, weather, distance and safety. By offering enthusiasts incentivies to explore, eBird can create a more accurate body of reports and provide better value to the birding community.

In this problem context, we study a simple policy that usually exploits, incentivizing agents to pull an arm only when the set of agent utility functions that would pull this arm without incentives has probability below a time-varying threshold. In our paper, we assume all arms are some agents' best arm. Under this assumption, we prove that with $O(N^2)$ payment budget, this policy has $O(N^2+M(\log(T))^2)$ cumulative expected regret where M is an upper bound on the limiting marginal probability density of agent utilities that are nearly indifferent between their best and the second best arm. If all agents' utility difference between their best and second best arm is bounded below by a positive number, which typically happens when the agent utility distribution is discrete, this policy achieve constant cumulative expected regret $O(N^2)$. The key difference between our problem setting and both the homogenous preference setting and the traditional multi-armed bandits setting is that we must incentivize agents to try suboptimal arms much less often, since all arms are some agents' best arm. Essentially, heterogeneity provides free exploration. These results suggest that heterogeneous agent preferences reduce but do not eliminate the need to incentive exploration, in relation to single-preference settings.

We broadly categorize the relevant previous literature into two categories based on whether there is money transfer. With money transfer, [1] considers a problem setting where the principal pays agents money to explore. This work assumes all agents have equal value for money and provide a complete characterization of achievable reward with a fixed budget. [3] generalizes this framework to include agents with heterogeneous value for money, and to allow an external signal to provide partial information about this valuation. Under this setting, this work proves a bound on achievable reward as a function of the budget and the signal scheme.

Without money transfer but using information asymmetry, [4] considers a simple model where agents arrive to the principal one by one and there are only two actions at each time. [5] generalizes [4] by allowing a finite number of actions at each time. [6] considers a problem setting where there are multiple agents at each time and agents may interact with each other. In these papers, the principal provides each agent a recommendation at each time that is Bayesian incentive-compatible. They prove the principal can achieve constant regret when utilities are deterministic and logarithmic regret when utilities are stochastic.

We structure our paper as follows: Section 2 formulates our problem; Section 3 states our algorithm and proves that we can achieve $O(N^2+M(\log(T))^2)$ regret with $O(N^2)$ incentive budget; Section 4 constructs an example showing regret is $\Omega(\log(T))$ in the worst case, regardless of incentive budget.

2 Problem Setting

We have N arms. Arm i has a fixed but unknown attribute vector $u_i \in \mathbb{R}^m$. A stream of myopic selfish agents come to our system. Agent t has linear preferences over attributes described by a vector $\theta_t \in \mathbb{R}^m$ that is unknown to the principal and drawn i.i.d. from a known distribution $F(\cdot)$ with compact support. We refer synonomously to an agent and that agent's preference vector: when we say "an agent θ", we mean "an agent with preference vector θ."

Each agent t chooses an arm to pull A_t, according to a process described below, and obtains utility $\theta_t \cdot u_{A_t}$. The principal and all agents then see a noisy observation of the attribute vector of the pulled arm of the form $O_t = u_{A_t} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2 I_m)$ is independent normally distributed noise, and I_m denotes an m-dimensional identity matrix. Although we assume a common variance across attributes for simplicity of presentation, our theoretical results hold if the variance differs.

At each time t, for each arm i, we (the principal) offer a non-negative payment $c_{i,t} \geq 0$ based on previous observations. We assume that agent t chooses to pull the arm that myopically maximizes the sum of this payment and an estimate of the utility obtained $\theta_t \cdot u_{i,t}$ where $u_{i,t}$ denotes the simple average of O_s over all previous pulls of arm i. In this paper, we assume all arms have been pulled once at time $t = 0$ and $u_{i,0}$ denotes a random draw from the arm attribute vector. For $t > 0$, denote $u_{i,t} = \frac{\sum_{s<t} O_s 1\{A_s=i\}+u_{i,0}}{\sum_{s<t} 1\{A_s=i\}+1}$ and $A_t = \text{argmax}_i\{\theta_t \cdot u_{i,t} + c_{i,t}\}$, breaking ties in favor of the arm with the highest incentive. We use $c_t = c_{A_t,t}$ to denote the actual incentive payment at time t.

This behavior may be recovered if agents are Bayesian and share a common non-informative prior distribution that is constant over $\mathbb{R}^m$ and know σ^2. In this case, the posterior distribution on u_i at time t is multivariate normal with mean $u_{i,t}$, and the expected value of $\theta_t \cdot u_i$ under this posterior conditioned on θ_t is $\theta_t \cdot u_{i,t}$ (see equation 2.13 in section 2.5, [2]). Alternatively, one may simply take our assumption that agents use the average as their estimate of an attribute value directly without such a Bayesian justification.

We define the regret at time t as $r(t) = \max_i\{\theta_t \cdot u_i\} - \theta_t \cdot u_{A_t}$, and the cumulative regret up to time T as $R(T) = \sum_{t=1}^{T} r(t)$. Define the cumulative payment up to time T similarly as $C(t) = \sum_{t=1}^{T} c(t)$. As the principal, we want to find a strategy $\mathcal{A}$ under which both the cumulative expected regret $\mathbb{E}_{\mathcal{A}}[R(T)]$ and the cumulative expected payment $\mathbb{E}_{\mathcal{A}}[C(T)]$ are small.

To support later development, we define some additional notation. We let $B(\theta)$ and $\hat{B}(\theta)$ refer to the index of the arm that is best and second best for an agent with preference vector θ, $B(\theta) \in \text{argmax}_i\, \theta \cdot u_i$ and $\hat{B}(\theta) = \text{argmax}_{i \neq B(\theta)}\, \theta \cdot u_i$, breaking ties uniformly at random. We let $N(i,t)$ denote the number of pulls of arm i at times up to and including t plus 1 (because of the initial pull), i.e. $N(i,t) = \sum_{s<t} 1\{A_s = i\} + 1$. We call time $t_n = \min_t\{\forall i, N(i,t) \geq n\}$ the *starting point of the n^{th} round*. We call the set of times $[t_n, t_{n+1})$ the *n^{th} round*.

3 Algorithm and Upper Bound

In this section, we propose a simple policy that mostly exploits, and occasionally incentivizes exploration when the probability of an arm would be pulled by all agent types below a time-varying threshold given the current posterior. We prove that with the help of heterogeneous preferences, we can get a certain amount of exploration for free via heterogeneity.

3.1 Our Algorithm

Our algorithm incentivizes pulling an arm i at a time t in round n if and only if both of the following criteria are met:

- the probability of pulling arm i would be below n^{-1} without incentives;
- arm i has not been played previously in the current round.

Ties are broken randomly. This algorithm does not need to know the horizon T in advance.

If our algorithm decides to incentivize an arm i, it uses the "pay whatever it takes" strategy in which the payment offered is $\max_{\theta,j} \theta \cdot (u_{j,t} - u_{i,t})$. This maximum over θ is taken over the support of F, which we recall is assumed compact. (We use this "pay whatever it takes" strategy for its simplicity, and in Section 3.4 we provide an alternate and smaller incentive payment that achieves the same payment budget bound and regret bound).

We describe our algorithm in detail as follows:

Algorithm 1 Algorithm: Incentivizing Exploration

Set n = 1 to denote the round number; Let $V = \emptyset$ be the set of arms that were pulled in the current round;
for $t = 1, 2, 3, \cdots$ **do**
  Let $S = \{i : P(\theta \cdot u_{i,t} > \theta \cdot u_{j,t}\ \forall j \neq i | u_{j,t}\ \forall j) < n^{-1}\}$ be the set of arms with unincentivized probability of being pulled below n^{-1}.
  if $S \setminus V$ is non-empty **then**
    Choose an arm i uniformly at random from $S \setminus V$
    Pay whatever it takes to incentivize pulling arm i, i.e., offer payment $c_{i,t} = \max_{\theta,j} \theta \cdot (u_{j,t} - u_{i,t})$ and $c_{j,t} = 0$ for $j \neq i$.
  else
    Let agents play myopically, i.e., offer payment $c_{j,t} = 0$ for all j
  end if
  Denote A_t as the pulled arm, update $V = V \cup \{A_t\}$, $u_{A_t,t}$ and $N(A_t, t)$
  if $n \neq \min_i N(i,t)$ **then**
    $V = \emptyset$
  end if
  Update the round number, $n = \min_i N(i,t)$
end for

3.2 Assumptions

In this section, we state several assumptions assumed by our analysis. First define

$$\Omega_{i,j} = \{\theta : B(\theta) = i, \hat{B}(\theta) = j\},$$

which is the set of agent preferences whose best arm is arm i and second best arm is arm j. With this definition, our analysis makes the following assumptions:

Assumption 1. *Let $F_{i,j}(y)$ be the marginal cumulative density function (or cumulative mass function if $F(\cdot)$ is a discrete distribution) of $(u_i - u_j) \cdot \theta$ conditioned on $\theta \in \Omega_{i,j}$. We assume $F_{i,j}(y) \leq My$ for all $y \in R^+$, $\forall i, j$.*

As we can see later in our proof, we only need $\max_{i,j} \limsup_{y \to 0^+} \frac{F_{i,j}(y)}{y}$ to be finite. Intuitively, assumption 1 states that there are not many agents who are indifferent between their best arm and the second best arm.

Assumption 2. *We assume F has a compact support set. Without loss of generality, we assume $\theta \in [0, W]^m$.*

We use $R = \max_{\theta,i,j} \theta \cdot (u_i - u_j)$ to denote the maximum regret that can be incurred at each time. Assumption 2 shows that $R < \infty$.

Assumption 3. *Denote $p = \min_i P(\{\theta : B(\theta) = i\})$. We assume $p > 0$.*

Assumption 3 means each arm i has a strictly positive proportion of users for which that arm is best.

3.3 General Results

In this section, we prove Algorithm 1 achieves $O(N^2 + M(\log(T))^2)$ cumulative regret with $O(N^2)$ payment budget. This is stated in the following pair of theorems, which together constitute our main results.

Theorem 1. *The payment budget for Algorithm 1 is bounded above by $O(N^2)$.*

Theorem 2. *The cumulative regret for Algorithm 1 is bounded above by $O(N^2 m + Mm^2(\log(T))^2)$.*

Before we prove these two theorems, we must first introduce two additional pieces of notation, which will be used in preliminary lemmas. Let $S(\delta)$ be the proportion of users whose utility difference between their best and second best arm is less than δ. Formally, $S(\delta) = P(\theta : \theta \cdot u_{B(\theta)} - \theta \cdot u_{\hat{B}(\theta)} \leq \delta)$. Then, let $p(\delta) = \min_i P(\{\theta : B(\theta) = i, \theta \cdot u_{B(\theta)} - \theta \cdot u_{\hat{B}(\theta)} > \delta\})$. We know $p(0) = p$.

With this additional notation, we now prove several lemmas. First, based on Assumption 1, we have the following bound for $S(\delta)$.

Lemma 1. *$S(\delta) \leq M\delta$.*

Proof.

$$\begin{aligned} S(\delta) &= \sum_{i,j} P(\theta \cdot (u_i - u_j) \leq \delta | \theta \in \Omega_{i,j}) P(\theta \in \Omega_{i,j}) \\ &\leq \sum_{i,j} M\delta \times P(\theta \in \Omega_{i,j}) \\ &= M\delta. \end{aligned}$$

□

The following lemma bounds the probability of making a mistake if we let the agents play myopically in the n^{th} round, given that the utility difference between his/her best and second best arm is bounded below by a constant.

Lemma 2. *Define τ to be any stopping time that is almost surely between t_n and $t_{n+1} - 1$ with respect to the filtration $\mathcal{F}_t = \sigma(A_1, \cdots, A_t, c_1, \cdots, c_t, O_1, \cdots, O_t)$, we have*

$$P(\arg\max\{\theta_\tau \cdot u_{i,\tau}\} \neq B(\theta_\tau) | \theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \leq 24Nm \exp\left(-\frac{1.8n\lambda^2}{16\sigma^2}\right),$$

for $n \geq n_0 = \max\{50, \frac{92.16\sigma^4}{\lambda^4}\}$.

We need the following lemma in order to prove Lemma 2.

Lemma 3. *For $n \geq n_0 = \max\{50, \frac{92.16\sigma^4}{\lambda^4}\}$, we have*

$$\frac{n\lambda}{4\sigma} \geq \sqrt{0.6n \log(\log_{1.1}(n) + 1)}.$$

Proof. First, we observe that

$$\begin{aligned} &\frac{n\lambda}{4\sigma} \geq \sqrt{0.6n \log(\log_{1.1}(n) + 1)} \\ \iff &\frac{n}{\log(\log_{1.1}(n) + 1)} \geq \frac{9.6\sigma^2}{\lambda^2}. \end{aligned}$$

Since $\log(x) \leq x - 1$ for $x > 0$, we know

$$\log(\log_{1.1}(n) + 1) = \log\left(\frac{\log(n)}{\log(1.1)} + 1\right) \leq \log(11\log(n) + 1) \leq \log(11n) \leq 3 + \log(n).$$

Thus, we know

$$\frac{n}{\log(\log_{1.1}(n) + 1)} \geq \frac{n}{3 + \log(n)}.$$

To prove the lemma, we just need to show for $n \geq n_0$, we have

$$\frac{n}{3 + \log(n)} \geq \frac{9.6\sigma^2}{\lambda^2}. \tag{1}$$

Inequality (1) is true because of the following two observations:

- for $n \geq 50$, we have $\frac{n}{3+\log(n)} \geq n^{0.5}$;
- for $n \geq \frac{92.16\sigma^4}{\lambda^4}$, we have $n^{0.5} \geq 9.6\sigma^2\lambda^2$.

Thus, we know our lemma is true.

□

To prove Lemma 2, we also need to use an adaptive concentration inequality due to [8]. For reference, we state it here as a Lemma.

Lemma 4 (Corollary 1 in [8])**.** *Let X_i be zero mean 1/2-subgaussian random variables. $\{S_n = \sum_{i=1}^n X_i, n \geq 1\}$ be a random walk. Let J be any stopping time with respect to $\{X_1, X_2, \cdots\}$. We allow J to take the value of ∞ where $P(J = \infty) = 1 - \lim_{n\to\infty} P(J \leq n)$. If*

$$f(n) = \sqrt{0.6n \log(\log_{1.1}(n) + 1) + bn},$$

then

$$Pr[\{S_J \geq f(J)\} \cap \{J < \infty\}] \leq 12e^{-1.8b}.$$

We now prove Lemma 2.

Proof of Lemma 2. In the n^{th} round, we know all arms have been pulled at least n times. For all the agents θ whose utility difference between their best and second best arm is greater than $2mW\lambda$, denote $K(\theta) = \max_{i\neq B(\theta)}\{\theta \cdot u_{i,t}\}$. If $|u_{i,t}^j - u_i^j| \leq \lambda$ for all i, j, then

$$\begin{aligned}
&\theta \cdot (u_{B(\theta),t} - u_{K(\theta),t}) \\
\geq &\theta \cdot (u_{B(\theta),t} - u_{B(\theta)}) + \theta \cdot (u_{K(\theta)} - u_{K(\theta),t}) + \theta \cdot (u_{B(\theta)} - u_{K(\theta)}) \\
> &- Wm\lambda - Wm\lambda + 2Wm\lambda = 0,
\end{aligned}$$

which means their myopic action would incur no regret.

Define $\epsilon_{i,\tau} = u_{i,\tau} - u_i$ and $\epsilon_{i,\tau}^j$ to be the j^{th} component of $\epsilon_{i,\tau}$. Thus, we have

$$\begin{aligned}
&P(\arg\max\{\theta_\tau \cdot u_{i,\tau}\} \neq B(\theta_\tau) | \theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\
\leq &P(\exists i, \exists j, |u_{i,\tau}^j - u_i^j| \geq \lambda | \theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\
\leq &\sum_i \sum_j P(|u_{i,\tau}^j - u_i^j| \geq \lambda | \theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\
= &\sum_i \sum_j P(|\epsilon_{i,\tau}^j| \geq \lambda | \theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda). \qquad (2)
\end{aligned}$$

To bound equation (2), we use Lemma 4. Define

$$S_{N(i,\tau)}^{i,j} = \frac{\epsilon_{i,\tau}^j}{2\sigma}.$$

Based on Lemma 3, for $n_0 = \max\{50, \frac{92.16\sigma^2}{\lambda^2}\}$ and $n \geq n_0$, we have

$$\frac{n\lambda}{4\sigma} \geq \sqrt{0.6n\log(\log_{1.1}(n)+1)}.$$

Thus, if we set $b = \frac{n\lambda^2}{16\sigma^2}$ in Lemma 4, for any $N(i,\tau) \geq n \geq n_0$, we have

$$\begin{aligned}\frac{N(i,\tau)\lambda}{2\sigma} \geq& \sqrt{0.6N(i,\tau)\log(\log_{1.1}(N(i,\tau))+1)} + \frac{\lambda}{4\sigma}\sqrt{nN(i,\tau)} \\ \geq& \sqrt{0.6N(i,\tau)\log(\log_{1.1}(N(i,\tau))+1) + bN(i,\tau)},\end{aligned}$$

where the last inequality is because $\sqrt{x}+\sqrt{y} \geq \sqrt{x+y}$. Thus, we have

$$\begin{aligned}&P(\epsilon^j_{i,\tau} \geq \lambda|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\ =&P\left(S^{i,j}_{N(i,\tau)} \geq \frac{N(i,\tau)\lambda}{2\sigma}\middle|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda\right) \\ \leq&P\left(S^{i,j}_{N(i,\tau)} \geq \sqrt{0.6N_{i,\tau}\log(\log_{1.1}(N(i,\tau))+1) + bN(i,\tau)}\middle|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda\right) \\ \leq&12\exp(-1.8b) = 12\exp\left(\frac{-1.8n\lambda^2}{16\sigma^2}\right).\end{aligned}$$

Similarily, we can bound

$$\begin{aligned}&P(\epsilon^j_{i,\tau} \leq -\lambda|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\ =&P(-\epsilon^j_{i,\tau} \geq \lambda|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\ \leq&12\exp\left(\frac{-1.8n\lambda^2}{16\sigma^2}\right).\end{aligned}$$

Therefore, we know $P(|\epsilon^j_{i,\tau}| \geq \lambda|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \leq 24\exp\left(\frac{-1.8n\lambda^2}{16\sigma^2}\right)$. Thus, we know

$$\begin{aligned}&\sum_i\sum_j P(|\epsilon^j_{i,\tau}| \geq \lambda|\theta_\tau \cdot (u_{B(\theta_\tau)} - u_{\hat{B}(\theta_\tau)}) > 2Wm\lambda) \\ \leq&24Nm\exp\left(\frac{-1.8n\lambda^2}{16\sigma^2}\right).\end{aligned}$$

□

Before we start analyzing the cumulative regret, we first prove the following lemma which bounds the expected length of each round.

Lemma 5. *Using our algorithm, we have* $\mathbb{E}[t_{n+1} - t_n] \leq Nn$, $\forall n \geq 1$.

Proof. A round completes when each arm is pulled at least once in that round. Let X_i be the number of agents who come to the system between the time after

the $(i-1)^{th}$ unique arm was pulled, up to and including the time when the i^{th} unique arm was pulled. Then we know

$$\mathbb{E}[t_{n+1} - t_n] = \sum_{i=1}^{N} E[X_i].$$

Fix i. In bounding X_i, we think of agents as "trials", where each trial can result in a new unique arm being pulled (which we call a "successful" trial), or not. There are two ways a trial can be successful:

- If there is at least one arm that has not been pulled and the probability of an agent utility function that would pull this arm without incentives is less than n^{-1}, then the principal will offer an incentive that causes this arm to be pulled (or one of these arms if there is more than one). In this case, the probability that the trial is succesful is 1.
- The probability of an agent utility function that would pull each un-pulled arm without incentives is at least n^{-1}. In this case, the probability that the trial is successful is at least n^{-1}.

Thus, X_i is stochastically dominated below by a geometric random variable with success probability n^{-1}, the expected number of trials up to and including the first success, $E[X_i]$, is bounded above by n. Thus,

$$E[t_{n+1} - t_n] \leq Nn.$$

□

We also need the following lemma in part of the proof of Theorem 1.

Lemma 6. *For all $n \geq 1$, we have*

$$0.9n^{5/6} \geq \sqrt{0.6n \log(\log_{1.1}(n) + 1)}.$$

Proof.

$$\begin{aligned} & 0.9n^{5/6} \geq \sqrt{0.6n \log(\log_{1.1}(n) + 1)} \\ \iff & 0.81n^{5/3} \geq 0.6n \log(\log_{1.1}(n) + 1) \\ \iff & \frac{81}{60} n^{2/3} \geq \log(\log_{1.1}(n) + 1) \end{aligned}$$

Denote $f(x) = \frac{81}{60}x^{2/3} - \log(\log_{1.1}(x) + 1)$. It's easy to compute $f'(x) = 0$ has a unique solution $x_0 = e^{2/3w(\frac{20e^{20000/314763}}{27}) - \frac{10000}{104921}}$ (here $w(\cdot)$ is the Lambert W-Function) and it is the global minimum. Since $f(x_0) \approx 0.0252 > 0$, we know $f(x) > 0$ for all $x \geq 1$. Thus, our lemma holds true. □

Now we are ready to prove our first main result, Theorem 1.

Proof. Denote $\epsilon_{i,t} = u_{i,t} - u_i$ to be the estimation error for the attribute vector u_i at time t. Denote $\epsilon_{i,t}^j$ to be the j^{th} component of $\epsilon_{i,t}$. Denote ω to be a sample path and $n(t,\omega)$ to be the round number for sample path ω at time t. For a fixed time t, define

$$L^{'}[l](t) = \{\omega : |\epsilon_{i,t}^j(\omega)| \leq g(n(t,\omega), l), \forall i, j\}$$

where $g(n,l)$ is a function which we will define later. Define $L[1](t) = L^{'}[1](t)$ and $L[i](t) = L^{'}[i](t) \setminus L^{'}[i-1](t)$ for $i \geq 2$. We call $L[l](t)$ the l^{th} envelope at time t. We often simplify the notation and use $L[l]$ instead of $L[l](t)$ without confusion.

In the calculation below, we omit the dependency on ω when refering to variables $c(t)$, $\epsilon_{i,t}^j$ and t_n. Based on the definition of $L[l]$, we know if $\omega \in L[l]$, the maximum payment we need to offer at time t is bounded above by

$$\begin{aligned}
&\max_i \theta_t \cdot u_{i,t} - \min_j \theta_t \cdot u_{j,t} \\
=&\max_i \theta_t \cdot (\epsilon_{i,t} + u_i) - \min_j \theta_t \cdot (\epsilon_{j,t} + u_j) \\
\leq&\max_i \theta_t \cdot u_i - \min_j \theta_t \cdot u_j + \max_i \theta_t \cdot \epsilon_{i,t} - \min_j \theta_t \cdot \epsilon_{j,t} \\
\leq&R + 2Wmg(n,l).
\end{aligned}$$

Based on the above notations, we can rewrite the cumulative payment as follows:

$$\begin{aligned}
&\sum_{t=1}^{\infty} c(t) \\
=&\sum_{l=1}^{\infty}\sum_{t=1}^{\infty} c(t)\mathbb{1}\{\omega \in L[l]\} \\
=&\sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\infty} c(t)\mathbb{1}\{\omega \in L[l]\}\mathbb{1}\{t \in [t_n, t_{n+1})\}.
\end{aligned}$$

Set $g(n,l)$ to be $\frac{2\sigma l}{n^{1/6}}$. Since if $|u_i^j - u_{i,t}^j| \leq \lambda$ is true $\forall i$, $\forall j$, then we know for those $\theta \in \{\theta : \theta \cdot u_{B(\theta)} - \max_{j \neq B(\theta)}\{\theta \cdot u_j\} > 2Wm\lambda\}$, they will correctly identify their best arm. Thus, if $|u_i^j - u_{i,t}^j| \leq \frac{2\sigma l}{n^{1/6}} \leq \frac{p^{-1}(\frac{p}{2})}{2Wm}$ $\forall i$ and $\forall j$, then the probability that an unincentivized agent would pull arm i is at least $\frac{p}{2}$. Further, if time t is in a round n that satisfies $n^{-1} \leq p/2$, then our algorithm will not incentivize pulling any arms. Denote $a_0 = \frac{4Wm\sigma}{p^{-1}(\frac{p}{2})}$. In order to have $\frac{2\sigma l}{n^{1/6}} \leq \frac{p^{-1}(\frac{p}{2})}{2Wm}$, it is sufficient to have $n \geq \lceil (a_0 l)^6 \rceil$. In order to have $n^{-1} \leq \frac{p}{2}$, we need $n \geq \frac{2}{p}$. Denote $n_2 = \frac{2}{p}$. Thus, we know we can only incur regret for sample paths ω in the l^{th} envelope in the first $\max\{n_2, \lceil (a_0 l)^6 \rceil\}$ rounds.

Thus,

$$\sum_{t=1}^{\infty} c(t) = \sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2, \lceil (a_0 l)^6 \rceil\}} c(t)\mathbb{1}\{\omega \in L[l]\}\mathbb{1}\{t \in [t_n, t_{n+1})\}.$$

Therefore,

$$
\begin{aligned}
&E\left[\sum_{t=1}^{\infty} c(t)\right] \\
=&\sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil(a_0 l)^6\rceil\}} E[c(t)|\omega\in L[l], t\in[t_n,t_{n+1}))P(\omega\in L[l], t\in[t_n,t_{n+1})) \\
=&\sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil(a_0 l)^6\rceil\}} \left[R+2Wm\frac{2\sigma l}{n^{1/6}}\right] P(\omega\in L[l]|t\in[t_n,t_{n+1}))P(t\in[t_n,t_{n+1})) \\
=&\sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil(a_0 l)^6\rceil\}} [R+4Wm\sigma l]\, P(\omega\in L[l]|t\in[t_n,t_{n+1}))P(t\in[t_n,t_{n+1})).
\end{aligned}
$$

We now bound $P(\omega\in L[l]|t\in[t_n,t_{n+1}))$ for $n\geq n_0$ and $l\geq 2$.

$$
\begin{aligned}
&P(\omega\in L[l]|t\in[t_n,t_{n+1})) \\
=&P(\omega\in L^{'}[l]|t\in[t_n,t_{n+1})) - P(\omega\in L^{'}[l-1]|t\in[t_n,t_{n+1})) \\
\leq&1-P\left(|\epsilon_{i,t}^j|<\frac{2\sigma(l-1)}{n^{1/6}},\forall i,j|t\in[t_n,t_{n+1})\right) \\
=&P\left(\exists i,j,s.t.|\epsilon_{i,t}^j|\geq\frac{2\sigma(l-1)}{n^{1/6}}|t\in[t_n,t_{n+1})\right) \\
\leq&\sum_{i,j}P\left(|\epsilon_{i,t}^j|\geq\frac{2\sigma(l-1)}{n^{1/6}}|t\in[t_n,t_{n+1})\right)
\end{aligned}
$$

Define $S_{i,t}^j=\frac{N(i,t)\epsilon_{i,t}^j}{2\sigma}$, then we know $S_{i,t}^j$ is a summation of $1/2$ gaussian random numbers. Therefore,

$$
\begin{aligned}
&\sum_{i,j}P\left(|\epsilon_{i,t}^j|\geq\frac{2\sigma(l-1)}{n^{1/6}}\middle|t\in[t_n,t_{n+1})\right) \\
=&\sum_{i,j}P\left(|S_{i,t}^j|\geq\frac{N(i,t)(l-1)}{n^{1/6}}\middle|t\in[t_n,t_{n+1})\right) \\
\leq&\sum_{i,j}P(|S_{i,t}^j|\geq N(i,t)^{5/6}(l-1)|t\in[t_n,t_{n+1})).
\end{aligned}
$$

Based on Lemma 6, we know

$$
\begin{aligned}
&N(i,t)^{5/6}(l-1) \\
=&0.9N(i,t)^{5/6}+N(i,t)^{5/6}(l-1.9) \\
\geq&\sqrt{0.6N(i,t)\log(\log_{1.1}(N(i,t))+1)}+\sqrt{(l-1.9)^2N(i,t)} \\
\geq&\sqrt{0.6N(i,t)\log(\log_{1.1}(N(i,t))+1)+(l-1.9)^2N(i,t)}.
\end{aligned}
$$

Based on Lemma 4, we know

$$\begin{aligned}&\sum_{i,j} P(|S_{i,t}^j| \geq N(i,t)^{5/6}(l-1)|t \in [t_n, t_{n+1}))\\ \leq &\sum_{i,j} 24e^{-1.8(l-1.9)^2} = 24Nme^{-1.8(l-1.9)^2}.\end{aligned}$$

Thus,

$$\begin{aligned}&\sum_{t=1}^{\infty} c(t)\\ \leq &\sum_{l=1}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0 l)^6\rceil\}} \left[R + 2Wm\frac{2\sigma l}{n^{1/6}}\right] P(\omega \in L[l]|t \in [t_n, t_{n+1}))P(t \in [t_n, t_{n+1}))\\ \leq &\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0)^6\rceil\}} [R + 4Wm\sigma]\, P(\omega \in L[1]|t \in [t_n, t_{n+1}))P(t \in [t_n, t_{n+1}))\\ + &\sum_{l=2}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0 l)^6\rceil\}} [R + 4Wm\sigma l]\, P(\omega \in L[l]|t \in [t_n, t_{n+1}))P(t \in [t_n, t_{n+1}))\\ \leq &\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0)^6\rceil\}} [R + 4Wm\sigma]\, P(t \in [t_n, t_{n+1}))\\ + &\sum_{l=2}^{\infty}\sum_{t=1}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0 l)^6\rceil\}} [R + 4Wm\sigma l]\, 24Nme^{-1.8(l-1.9)^2} P(t \in [t_n, t_{n+1}))\\ \leq &\sum_{n=1}^{\max\{n_2,\lceil (a_0)^6\rceil\}} [R + 4Wm\sigma]\, Nn + \sum_{l=2}^{\infty}\sum_{n=1}^{\max\{n_2,\lceil (a_0 l)^6\rceil\}} [R + 4Wm\sigma l]\, 24Nme^{-1.8(l-1.9)^2} Nn\\ \leq &[R + 4Wm\sigma]\, N(\max\{n_2, \lceil (a_0)^6\rceil\})^2 + \sum_{l=2}^{\infty} 24N^2 m[R + 4Wml\sigma](\max\{n_2, \lceil (a_0 l)^6\rceil\})^2 e^{-1.8(l-1.9)^2}\\ = &O(N^2).\end{aligned}$$

□

Lemma 7. *The expected number of payments for Algorithm 1 is bounded above by $O(N^2)$.*

Proof. If $|u_i^j - u_{i,t}^j| \leq \lambda$ is true $\forall i, \forall j$, then we know for those $\theta \in \{\theta : \theta \cdot u_{B(\theta)} - \max_{j \neq B(\theta)}\{\theta \cdot u_j\} > 2mW\lambda\}$, they will correctly identify their best arm. Thus we know, in the n^{th} round, if $|u_i^j - u_{i,t}^j| \leq \frac{p^{-1}(\frac{p}{2})}{2Wm}$ $\forall i$ and $\forall j$, and $n^{-1} \leq p/2$, we do not need to incentivize any arms. In order to have $n^{-1} \leq \frac{p}{2}$, we need $n \geq \frac{2}{p}$. Denote $n_1 = \max\{n_0, \frac{2}{p}\}$. Denote $\delta_0 = p^{-1}(\frac{p}{2}) > 0$ (because of Assumption 1).

Define τ_n^i to be the first time we pull arm i in the n^{th} round. Then

$$\sum_{t=1}^{\infty} \mathbb{1}\{c(t) > 0\} = \sum_{n=1}^{\infty} \sum_{i=1}^{N} \mathbb{1}\{c(\tau_n^i) > 0\}.$$

The cumulative expected number of payments is bounded above by:

$$\begin{aligned}
&E\left[\sum_{t=1}^{\infty} \mathbb{1}\{c(t) > 0\}\right] \\
=&\sum_{n=1}^{\infty} \sum_{i=1}^{N} P(c(\tau_n^i) > 0) \\
\leq &\sum_{n=n_1}^{\infty} \sum_{i=1}^{N} P\left(\exists i, j : |u_i^j - u_{i,\tau_n^i}^j| > \frac{p^{-1}(\frac{p}{2})}{2Wm}\right) + \sum_{n=1}^{n_1} N \\
\leq &\sum_{n=n_1}^{\infty} \sum_{i=1}^{N} 24Nm \exp\left(\frac{-1.8n\delta_0^2}{64W^2m^2\sigma^2}\right) + \sum_{n=1}^{n_1} N \\
\leq &\sum_{n=n_1}^{\infty} 24Nm \exp\left(\frac{-1.8n\delta_0^2}{64W^2m^2\sigma^2}\right) \times N + \sum_{n=1}^{n_1} N \\
\leq &24N^2m \frac{1}{\exp(\frac{1.8\delta_0}{64W^2m^2\sigma^2}) - 1} + Nn_1,
\end{aligned}$$

Thus, we know the expected number of payments is bounded above by $O(N^2)$.

□

Now we are ready to prove our second main result, Theorem 2.

Proof. For regret incurred in the first n_0 round, it is bounded above by $\sum_{n=1}^{n_0} NRn$.

For regret incurred after the first n_0 round, it has two different components: the regret incurred when we let the agents play myopically and the regret incurred when we incentivize the agents. Using Lemma 7, the expected regret incurred when we incentivize the agents is bounded above by: $\left[24N^2m \frac{1}{\exp(\frac{1.8\delta_0}{64W^2m^2\sigma^2}) - 1} + Nn_1\right] R$.

For the regret incurred when we let the agents play myopically at time $t \geq t_{n_0}$, it consists of the following two components:

- For those users whose utility difference between their best and the second best arm is greater than $f(t)$: we define a sequence of stopping time τ_n^k to be the k^{th} time period in the n^{th} round. For $k > t_{n+1} - t_n$, we define $\tau_n^k = \infty$. For $\tau_n^k = t$, the probability of these users making a mistake is bounded above by $24Nm \exp\left(-\frac{1.8nf(\tau_n^k)^2}{64W^2m^2\sigma^2}\right)$ and the expected regret is bounded above by $24Nm \exp\left(-\frac{1.8nf(\tau_n^k)^2}{64W^2m^2\sigma^2}\right) \times R$. We denote the regret incurred by these agents as $r_1(\tau_n^k)$. For $k > t_{n+1} - t_n$, we define $r_1(\tau_n^k) = 0$.

- For those user whose utility difference between their best and the second best arm is smaller than $f(t)$: this happens with probability $S(f(t))$ at each time and regret is bounded above by $S(f(t)) \times f(t) = Mf(t)^2$. We denote the regret incurred by these agents as $r_2(t)$.

Thus, the cumulative expected regret incurred up to time T when we let the agent play myopically is bounded above by:

$$\begin{aligned}
&E\left[\sum_{t=1}^{T} r(t)\right] \\
=&E\left[\sum_{t=1}^{t_{n_0}} r(t) + \sum_{t=t_{n_0}}^{T} (r_1(t) + r_2(t))\right] \\
\leq&\sum_{n=1}^{n_0} NRn + E\left[\sum_{n=n_0}^{T} \sum_{t=t_n}^{t_{n+1}-1} r_1(t)\right] + E\left[\sum_{t=1}^{T} r_2(t)\right] \\
=&\sum_{n=1}^{n_0} NRn + E\left[\sum_{n=n_0}^{T} \sum_{k=1}^{\infty} r_1(\tau_n^k)\right] + E\left[\sum_{t=1}^{T} r_2(t)\right]. \qquad (3)
\end{aligned}$$

Since

$$\begin{aligned}
&E\left[\sum_{n=n_0}^{T} \sum_{k=1}^{\infty} r_1(\tau_n^k)\right] \\
=&\sum_{n=n_0}^{T} \sum_{k=1}^{\infty} E[r_1(\tau_n^k)] \\
=&\sum_{n=n_0}^{T} \sum_{k=1}^{\infty} (E[r_1(\tau_n^k)|\tau_n^k < \infty] \times P(\tau_n^k < \infty) + E[r_1(\tau_n^k)|\tau_n^k = \infty] \times P(\tau_n^k = \infty)) \\
=&\sum_{n=n_0}^{T} \sum_{k=1}^{\infty} E[r_1(\tau_n^k)|\tau_n^k < \infty] \times P(\tau_n^k < \infty),
\end{aligned}$$

we have

$$\begin{aligned}
(3) =& \sum_{n=1}^{n_0} NRn + \sum_{n=n_0}^{T} \sum_{k=1}^{\infty} E[r_1(\tau_n^k)|\tau_n^k < \infty] \times P(\tau_n^k < \infty) + E\left[\sum_{t=1}^{T} r_2(t)\right] \\
\leq& \sum_{n=1}^{n_0} NRn + \sum_{n=n_0}^{T} \left[\sum_{k=1}^{\infty} 24Nm \exp\left(-\frac{1.8nf(\tau_n^k)^2}{64W^2m^2\sigma^2}\right) R \times P(\tau_n^k < \infty)\right] + \sum_{k=1}^{T} Mf(t)^2 \\
\leq& \sum_{n=1}^{n_0} NRn + \sum_{n=1}^{T} 24Nm \exp\left(-\frac{1.8nf(n)^2}{64W^2m^2\sigma^2}\right) R \times Nn + \sum_{t=1}^{T} Mf(t)^2. \qquad (4)
\end{aligned}$$

Thus the cumulative regret at time T is bounded above by

$$\sum_{n=1}^{n_0} NRn + \sum_{n=1}^{T} 24Nm \exp\left(-\frac{1.8nf(n)^2}{64W^2m^2\sigma^2}\right) \times R \times Nn + \sum_{t=1}^{T} Mf(t)^2$$
$$+24N^2m\frac{1}{e^{\frac{1.8\delta_0}{64W^2m^2\sigma^2}}-1}R + N\left(\max\left\{n_0, \frac{2}{p}\right\}\right)R.$$

For a fixed T, we only need to minimize the following two terms since all others are constant:

$$\sum_{n=1}^{T} 24Nm \exp\left(-\frac{1.8nf(n)^2}{64W^2m^2\sigma^2}\right) \times R \times Nn + \sum_{t=1}^{T} Mf(t)^2. \tag{5}$$

If we set $f^2(t) = \frac{2\log(T)\times 64W^2m^2\sigma^2}{1.8t}$, then

$$\begin{aligned}
&\sum_{n=1}^{T} 24Nm \exp\left(-\frac{1.8nf(n)^2}{64W^2m^2\sigma^2}\right) \times R \times Nn + \sum_{t=1}^{T} Mf(t)^2 \\
\leq &\sum_{n=1}^{T} 24N^2mnR\exp\left(-2\log(T)\right) + \frac{128W^2m^2\sigma^2M\log(T)}{1.8}\sum_{t=1}^{T}\frac{1}{n} \\
\leq &24N^2mR\frac{T(T-1)}{2T^2} + 71.12W^2m^2\sigma^2M\log(T)(\log(T)+1) \\
\leq &12N^2mR + 71.12W^2m^2\sigma^2M\log(T)(\log(T)+1).
\end{aligned}$$

Thus, the cumulative expected regret is bounded by $O(N^2m + Mm^2\log(T))$. □

Corollary 1. *If $\exists\delta > 0$ such that $F_{i,j}(\delta) = 0$ for all i, j, then the cumulative expected regret is bounded by $O(N^2)$.*

Proof. The proof of this corollary is similar to the proof of Theorem 2. If we set $f^2(t) = \frac{2\log(T)\times 64W^2m^2\sigma^2}{1.8t}$, then there exists a t_0 such that for $t > t_0$, $S(f(t)) = 0$. Thus, similar to equation (**??**), we know the cumulative expected regret when we let the agents play myopically is bounded above by

$$\sum_{n=1}^{n_0} NRn + \sum_{n=1}^{T} 24Nm \exp\left(-\frac{1.8nf(n)^2}{64W^2m^2\sigma^2}\right) \times R \times Nn + \sum_{t=1}^{t_0} Mf(t)^2$$

Therefore, based on the same analysis of Theorem 2, we know the cumulative regret is bounded by $O(N^2)$. □

3.4 Practical Issues

In Algorithm 1, we use "pay whatever it takes" strategy when we decide to incentivize the agent. However, "pay whatever it takes" only shows up in the proof of Lemma 5. Without loss of generality, suppose we want to incentivize arm i at time t at the n^{th} round. Based on the proof of Lemma 5, as long as we offer a payment $c_{i,t}$ such that arm i has at least n^{-1} probability being pulled at time t, our results still hold true. We could compute this $c_{i,t}$ dynamically based on $F(\cdot)$ as well as our current estimate $u_{i,t}$. Here is the revised algorithm which would work well in practice:

Algorithm 2 Algorithm: Incentivizing Exploration

Set n = 1 to denote the round number; Let $V = \emptyset$ be the set of arms that were pulled in the current round;
for $t = 1, 2, 3, \cdots$ **do**
  Let $S = \{i : P(\theta \cdot u_{i,t} > \theta \cdot u_{j,t}\ \forall j \neq i | u_{j,t}\ \forall j) < n^{-1}\}$ be the set of arms with unincentivized probability of being pulled below n^{-1}.
  if $S \setminus V$ is non-empty **then**
    Choose an arm i uniformly at random from $S \setminus V$
    Offer payment $c_{i,t} = \inf\{c : P(\theta \sim F : \theta \cdot u_{i,t} + c > max_j \theta \cdot u_{j,t}) > n^{-1}\}$
  else
    Let agents play myopically, i.e., offer payment $c_{j,t} = 0$ for all j
  end if
  Denote A_t as the pulled arm, update $V = V \cup \{A_t\}$, $u_{A_t,t}$ and $N(A_t, t)$
  if $n \neq \min_i N(i, t)$ **then**
    $V = \emptyset$
  end if
  Update the round number, $n = \min_i N(i, t)$
end for

The same proof would work and we can get the exact same results as Algorithm 1.

4 Lower Bound $\Omega(\log(T))$

In this section, we assume θ follows a continuous distribution $F(\cdot)$. We provide an example to show the best possible lower bound is $\Omega(\log(T))$ regardless of the incentivizing strategy.

Suppose we have two arms. Arm 1 has attribute vector $(0, 0)$ and arm 2 has attribute vector $(0, 1)$. We assume the users' preference are uniformly distributed on the unit circle. If the user knows the exact attribute vectors for both arms, then the users with preference on the bottom half circle will choose arm 1 and the users with preference on the top half circle will choose arm 2.

Consider the following algorithm: at each step, let the agents play myopically; however, they are going to see the noisy rewards for both arms.

To lower bound the regret, we assume that the agents already know the true attribute vector for arm 1. Without loss of generality, denote $u_{2,t} = (0,1) + (z_{t,1}, z_{t,2}) = (0,1) + (N(0,1/t), N(0,1/t))$ to be the estimate attribute vector for arm 2 (Without loss of generality, we assume the variance for the noise is 1).

Since $(z_{t,1}, z_{t,2})$ is symmetric around $(0,0)$, we know

$$\begin{aligned}
&E[r(t)] \\
=&E[r(t)|z_{t,1}>0, z_{t,2}>0] \times P(z_{t,1}>0, z_{t,2}>0) + E[r(t)|z_{t,1}>0, z_{t,2}<0] \times P(z_{t,1}>0, z_{t,2}<0) \\
&+ E[r(t)|z_{t,1}<0, z_{t,2}>0] \times P(z_{t,1}<0, z_{n,2}>0) + E[r(t)|z_{t,1}<0, z_{t,2}<0] \times P(z_{t,1}<0, z_{t,2}<0) \\
\geq&0.25 \times E[r(t)|z_{t,1}>0, z_{t,2}>0].
\end{aligned}$$

Given $z_{t,1}>0$ and $z_{t,2}>0$, we know users whose preference vector between $(-1,0)$ and $\left(\frac{-1-z_{t,2}}{\sqrt{z_{t,1}^2+(1+z_{t,2})^2}}, \frac{z_{t,1}}{\sqrt{z_{t,1}^2+(1+z_{t,2})^2}}\right)$ as well as users whose preference vector between $(1,0)$ and $\left(\frac{1+z_{t,2}}{\sqrt{z_{t,1}^2+(1+z_{t,2})^2}}, \frac{-z_{t,1}}{\sqrt{z_{t,1}^2+(1+z_{t,2})^2}}\right)$ will make a mistake. The regret is the absolute value of the second coordinate of the user's preference vector. Thus, we know

$$\begin{aligned}
&E[r(t)|z_{t,1}>0, z_{t,2}>0] \\
=&4 \times 2 \int_0^\infty \int_0^\infty \int_0^{\arctan\left(\frac{z_{t,1}}{1+z_{t,2}}\right)} \frac{\sin(\theta)}{2\pi} d(\theta) \frac{e^{-\frac{t \times z_{t,1}^2}{2}}\sqrt{t}}{\sqrt{2\pi}} d(z_{t,1}) \frac{e^{-\frac{t \times z_{t,2}^2}{2}}\sqrt{t}}{\sqrt{2\pi}} d(z_{t,2}) \\
=&\frac{2}{\pi^2} \int_0^\infty \int_0^\infty t \times \left[1 - \frac{1+z_{t,2}}{\sqrt{z_{t,1}^2+(1+z_{t,2})^2}}\right] e^{-\frac{t \times z_{t,1}^2}{2}} e^{-\frac{t \times z_{t,2}^2}{2}} d(z_{t,1}) d(z_{t,2}) \\
=&\frac{2}{\pi^2} \int_0^\infty \int_0^\infty \left[1 - \frac{\sqrt{t}+z_{t,2}}{\sqrt{z_{t,1}^2+(\sqrt{t}+z_{t,2})^2}}\right] e^{-\frac{z_{t,1}^2}{2}} e^{-\frac{z_{t,2}^2}{2}} d(z_{t,1}) d(z_{t,2})
\end{aligned}$$

Below, we want to show

$$\lim_{t\to\infty} \frac{E[r(t)|z_{t,1}>0, z_{t,2}>0]}{t} = O(1),$$

and use the fact that $\sum_{n=1}^{T} \frac{1}{n} = O(\log(T))$ to show the regret is at least $\Omega(\log(T))$.

Denote $d(t) = t\left[1 - \frac{\sqrt{t}+z_{t,2}}{\sqrt{z_{t,1}^2+(\sqrt{t}+z_{t,2})^2}}\right]$. Since

$$d'(t) = \frac{-z_{t,1}^2(2z_{t,2}+3\sqrt{t}) - 2(z_{t,2}+\sqrt{t})^3}{2(z_{t,1}^2+(z_{t,2}+\sqrt{t})^2)^{3/2}} + 1.$$

and $\lim_{t\to\infty} d^{'}(t) = 2$, we know for t large enough, $d(t)$ is a increasing function in terms of t. Thus, based on the Monotone Convergence Theorem, we have

$$\begin{aligned}
&\lim_{t\to\infty} \frac{E[r(t)|z_{t,1}>0, z_{t,2}>0]}{t} \\
=&\frac{2}{\pi^2}\int_0^\infty \int_0^\infty \lim_{t\to\infty} \left[\left[1 - \frac{\sqrt{t}+z_{t,2}}{\sqrt{z_{t,1}^2+(\sqrt{t}+z_{t,2})^2}}\right] e^{-\frac{z_{t,1}^2}{2}} e^{-\frac{z_{t,2}^2}{2}}\right] d(z_{t,1}) d(z_{t,2})
\end{aligned}$$

Based on our calculation, we know

$$\lim_{t\to\infty} d(t) = \frac{z_{t,1}^2}{2}.$$

Thus,

$$\begin{aligned}
&\lim_{t\to\infty} \frac{E[r(t)|z_{t,1}>0, z_{t,2}>0]}{t} \\
=&\frac{2}{\pi^2}\int_0^\infty \int_0^\infty \lim_{t\to\infty} \left[\frac{z_{t,1}^2}{2} e^{-\frac{z_{t,1}^2}{2}} e^{-\frac{z_{t,2}^2}{2}}\right] d(z_{t,1}) d(z_{t,2}) = \frac{1}{2\pi}.
\end{aligned}$$

Thus, the cumulative expected regret is at least $\Omega(\log(T))$.

5 Conclusion

In this paper, we study the incentivizing exploration problem with heterogeneous user preferences, which generalize the problem setting studied by [1] and [3]. We propose a simple policy that mostly exploits and occasionally incentivizing exploration, which can achieve $O(N^2 + M(\log(T))^2)$ cumulative expected regret with $O(N^2)$ payment budget.

References

[1] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22. ACM, 2014.

[2] A.B. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. CRC Press, second edition, 2004.

[3] Li Han, David Kempe, and Ruixin Qiang. Incentivizing exploration with heterogeneous value of money. In *International Conference on Web and Internet Economics*, pages 370–383. Springer, 2015.

[4] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the wisdom of the crowd. *Journal of Political Economy*, 122(5):988–1012, 2014.

[5] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582. ACM, 2015.

[6] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. *arXiv preprint arXiv:1602.07570*, 2016.

[7] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[8] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. In *Advances In Neural Information Processing Systems*, pages 1343–1351, 2016.