Incentivizing Exploration By Heterogeneous Users

Bangrui Chen[1], Peter I. Frazier[1] & David Kempe[2]

[1] School of Operations Research and Information Engineering, Cornell University; [2] Department of Computer Science, USC

Motivation

Amazon wants users to ***explore***
Each customer only wants to buy one good item

An alternative that seems worse initially may remain unexplored because customers have no incentives to explore it!

Literature Review

Two recent lines of work have shown that effecting a societally near-optimal outcome in this setting requires explicitly inducing exploration:

- **Without Money Transfer** Kremer et al. (2014) and Mansour et al. (2015, 2016, 2018) assume that the principal has an informational advantage in being the only one to observe the results of past arm pulls (as in driving route recommendations). The principal can use her advantage to induce exploration by recommending apparently sub-optimal arms, as long as agents cannot do better on their own.
- **With Money Transfer** Frazier et al. (2014) and Han et al. (2015) instead assume that the results of all past arm pulls are publicly observable (as on a review-sharing site). They suppose that the principal can incentivize exploration by offering arm-specific reward payments.

In this work, we present the first algorithm/analysis for incentivizing exploration when users have heterogeneous preferences over arms and we prove **"heterogeneity provides free exploration"**.

Model

- N bandit arms provide vector-valued outcomes equal to an unknown arm-specific attribute vector $u_i \in \mathbb{R}^d$, perturbed by independent noise.
- Agents arrive sequentially and observe current estimates of the arms' attribute vectors $\hat{u}_{i,t}$, which are averages of other agents' past pulls.
- Agents have heterogeneous linear preferences over arm attributes.
- Agent t has preference vector θ_t drawn from a known distribution.
- A principal knows only the distribution from which agents' preferences are drawn but not the specific draws.
- The principal can offer arm-specific incentive payments $c_{t,i}$ to encourage agents to explore underplayed arms.
- Agents are myopic and choose the arm $i_t = \arg\max_i\{\theta_t \cdot \hat{u}_{i,t} + c_{t,i}\}$.
- The regret at time t is $r_t = (\max_i \theta_t \cdot u_i) - \theta_t \cdot u_{i_t}$ and the payment is $c_t = c_{t,i_t}$.
- The principal seeks to minimize the total expected cumulative regret while also making a small expected cumulative payment.

Key Assumptions

- **(Every arm is someone's best)** We use p to denote the minimum (over all arms) fraction of users that prefer any particular arm.
- **(Not too many near-ties)** Let $q(z)$ be the cumulative distribution function of those agents whose utility difference between their best and second best arm is less than or equal to z, then there exists a $\hat{z} > 0$, L such that $q(z) \leq L \cdot z$ for all $z \leq \hat{z}$.
- **(Compact Support)** θ has a compact support set contained in $[0, D]^d$.

Main Results

Theorem 1

With the previously stated assumptions, there is a policy that achieves expected cumulative regret $O(Ne^{2/p} + LN\log^3(T))$, using expected cumulative payments of $O(N^2 e^{2/p})$.

In particular, when agents who are close to tied between two arms have measure 0, both the expected regret and expected payment are bounded by constants (with respect to T).

Algorithm

Notation:

- Phase: Phase s starts when each arm has been pulled at least s times. $m_{t,i}$ denotes the number of pulls for arm i up to time t.
- Payment-eligible: An arm i is *payment-eligible* at time t (in phase s) if both of the following hold:
  - i has been pulled at most s times up to time t, i.e., $m_{t,i} \leq s$.
  - The conditional probability of pulling arm i is less than$1/\log(s)$ given the current estimates $\hat{u}_{t,i'}$ of the arms' attribute vectors.

Our Algorithm:

Set the current phase number $s = 1$. {Each arm is pulled once initially "for free."}
for time steps $t = 1, 2, 3, \ldots$ **do**
  if $m_{t,i} \geq s + 1$ for all arms i **then**
    Increment the phase $s = s + 1$.
  end if
  if there is a payment-eligible arm i **then**
    Let i be an arbitrary payment-eligible arm.
    Offer payment $c_{t,i} = \max_{\theta, i'} \theta \cdot (\hat{\mu}_{t,i'} - \hat{\mu}_{t,i})$ for pulling arm i (and payment 0 for all other arms).
  else
    Let agent t play myopically, i.e., offer payments 0 for all arms.
  end if
end for

Proof Sketch

The key technical lemma in our proof is a Hoeffding-like concentration inequality that holds for a random, adaptively chosen number of samples (Zhao et al. 2016).

Payment Proof:

- For the early phases, we crudely bound the number of payment by N for each phase;
- For the later phases, we uses our technical lemma to rule out any incentives unless large misestimates of the arm locations occur, which is exponentially unlikely as the phase advances.

Regret Proof:

- Regret incurred when an agent was incentivized to pull a sub-optimal arm: the analysis here is very similar to the payment proof;
- Regret incurred when an agent myopically pulled a suboptimal arm: in this case, we define a phase-dependent cutoff $\gamma(s(t))$ to distinguish agents based on their regret.
  - Agents with $r(t) \geq \gamma(s(t))$:
    - this requires severe misestimates of arm locations and such misestimates are exponentially unlikely to occur;
    - since θ_t has a compact support, the maximum regret is bounded by a constant;
  - Agents with $r(t) \leq \gamma(s(t))$:
    - this requires agents to be almost tied in their preference for the best arm, but there are not so many agents have near-ties preferences;
    - the maximum regret is bounded above by $\gamma(s(t))$;

References

[1] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". Journal of Political Economy, 122(5):988–1012, 2014.
[2] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In Proceedings of the 16th ACM Conference on Economics and Computation (EC), pages 565–582. ACM, 2015.
[3] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian explo- ration: Incentivizing exploration in bayesian games. In Proceedings of the 17th ACM Conference on Economics and Computation (EC), 2016.
[4] Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In Proceedings of the 9th Innovations in Theoretical Computer Science (ITCS) conference, 2018.
[5] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In Proceedings of the 15th ACM conference on Economics and Computation (EC), pages 5–22. ACM, 2014.
[6] Li Han, David Kempe, and Ruixin Qiang. Incentivizing exploration with heterogeneous value of money. In Proceedings of the 11th International Conference on Web and Internet Economics (WINE), pages 370–383. Springer, 2015.
[7] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequal- ities for sequential decision problems. In Proceedings of the 30th Advances In Neural Information Processing Systems (NIPS), pages 1343–1351, 2016.