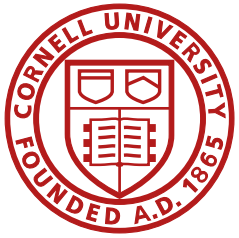

# Incentivizing Exploration by Heterogeneous Users

COLT 2018

Bangrui Chen, Peter Frazier

Cornell University
Operations Research and Information Engineering
bc496@cornell.com, pf98@cornell.edu

David Kempe

University of Southern California
Department of Computer Science
david.m.kempe@gmail.com

July 8, 2018

# Problem Setting

- We consider the problem of incentivizing exploration with heterogeneous agents.
- In this problem, $N$ 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 with preference vector $\theta_t$ and choose arms to pull based on their own private and heterogeneous linear utility functions over attributes and the estimates of the arms' attribute vectors $\hat{u}_{i,t}$, derived from observations of other agents' past pulls.
- Agents are myopic and selfish and thus would choose the arm with maximum estimated utility, i.e. they will pull arm $i_t = \arg\max_i \{\theta_t \cdot \hat{u}_{i,t}\}$.

# Problem Setting

- A principal, knowing only the distribution from which agents' preferences are drawn, but not the specific draws, can offer arm-specific incentive payments $c_{t,i}$ to encourage agents to explore underplayed arms.
- We define the regret at time $t$ as $r_t = (\max_i \theta_t \cdot u_i) - \theta_t \cdot u_{i_t}$ and payment $c_t = c_{t,i_t}$.
- The principal seeks to minimize the total expected cumulative regret incurred by agents relative to their best arms, while also making a small expected cumulative payment.

# Previous Work

- Kremer et al. (2014) and Mansour et al. (2015, 2016, 2018) (see Slivkins (2017) for an overview) 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.
- 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.

# Our Contribution

We present the first algorithm and analysis (of which we are aware) for incentivizing exploration when users have heterogeneous preferences over arms, with the following applications:

- Sites that host user-written reviews of products, restaurants, hotels and travel experiences, such as TripAdvisor;
- Citizen science projects such as eBird and Galaxy Zoo.

# Key Assumptions

- (**Every arm is someone's best**) Each arm $i$ has a strictly positive proportion of users for whom $i$ is the best arm. 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**) $\theta$ has a compact support set contained in $[0, D]^d$.

# Main Result

**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 $O(N^2e^{2/p})$.

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 $T$).

# Notations

- Phase: Our algorithm divides time into *phases* $s = 1, 2, 3, \ldots$. Phase $s$ starts when each arm has been pulled at least $s$ times. We use $m_{t,i}$ to denote the number of pulls for arm $i$ up to time $t$ and indicate the start time of phase $s$ by $t_s$.

# Notations

- Phase: Our algorithm divides time into *phases* $s = 1, 2, 3, \ldots$. Phase $s$ starts when each arm has been pulled at least $s$ times. We use $m_{t,i}$ to denote the number of pulls for arm $i$ up to time $t$ and indicate the start time of phase $s$ by $t_s$.
- 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.

# 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 $s = s + 1$.
  **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.

# Payment Analysis

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

- 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 Analysis

**Regret Proof:**

- Regret occurred when an agent was incentivized to pull a sub-optimal arm: the analysis here is very similar to the payment proof;

# Regret Analysis

**Regret Proof:**

- Regret occurred when an agent was incentivized to pull a sub-optimal arm: the analysis here is very similar to the payment proof;
- Regret occurred when an agent myopically pulled a suboptimal arm: in this case, we define a phase-dependent cutoff to further distinguish agents based on their regret.

# Regret Analysis

**Regret Proof:**

- Regret occurred when an agent was incentivized to pull a sub-optimal arm: the analysis here is very similar to the payment proof;
- Regret occurred when an agent myopically pulled a suboptimal arm: in this case, we define a phase-dependent cutoff to further distinguish agents based on their regret.
    - For those agents incurring large regret, which requires severe misestimates of arm locations and such misestimates are exponentially unlikely to occur, we use the following analysis to bound the regret:
        - ⋆ the technical lemma suggests this happens with exponentially decreasing probability;
        - ⋆ based on our compact support assumption, the maximum regret is bounded above by a constant;

# Regret Analysis

**Regret Proof:**

- Regret occurred when an agent was incentivized to pull a sub-optimal arm: the analysis here is very similar to the payment proof;
- Regret occurred when an agent myopically pulled a suboptimal arm: in this case, we define a phase-dependent cutoff to further distinguish agents based on their regret.
  - For those agents incurring large regret, which requires severe misestimates of arm locations and such misestimates are exponentially unlikely to occur, we use the following analysis to bound the regret:
    - ⋆ the technical lemma suggests this happens with exponentially decreasing probability;
    - ⋆ based on our compact support assumption, the maximum regret is bounded above by a constant;
  - For those agents incurring small positive regret, which requires these agents to be almost tied in their preference for the best arm, we use the following analysis to bound the regret:
    - ⋆ there are not so many agents have near-ties preferences;
    - ⋆ the maximum regret is bounded above by the phase-dependent cutoff;

# Question?

Thanks for your time!