Incentivizing Exploration by Heterogeneous Users

COLT 2018

Bangrui Chen, Peter Frazier

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

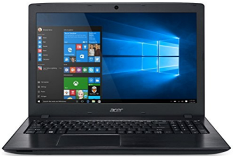
David Kempe

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

July 8, 2018

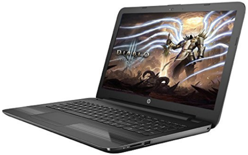
Motivation

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



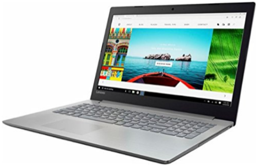
2,202
3.7 out of 5 stars

5 star	51%
4 star	16%
3 star	7%
2 star	6%
1 star	20%



508
3.6 out of 5 stars

5 star	49%
4 star	15%
3 star	8%
2 star	6%
1 star	22%



1
2.0 out of 5 stars

5 star	0%
4 star	0%
3 star	0%
2 star	100%
1 star	0%

Previous Work

Without Money Transfer:

- Implementing the "Wisdom of the Crowd", Kremer et al. 2014;
- Bayesian incentive-compatible bandit exploration, Mansour et al. 2015;
- $\cdots$

With Money Transfer

- Incentivizing exploration, Frazier et al. 2014;
- Incentivizing exploration with heterogeneous value of money, Han et al. 2015;
- $\cdots$

Our Contribution

- First algorithm and analysis for incentivizing exploration when users have heterogeneous preferences over arms;
- We proved **"heterogeneity provides free exploration"**.

Problem Setting

N **arms**

- Each arm is associated with a feature vector $u_i \in R^d$;
- Both the agents and the principal can observe the current estimate of u_i, denoted as $\hat{u}_{i,t}$, which equals to the average of all past observations;

Myopic Agents

- Agents arrive sequentially and their preference $\theta_t \in R^d$ follows known distribution $F(\cdot)$.
- Without any incentives, agent θ_t would choose the arm $i_t = \arg\max_i \{\theta_t \cdot \hat{u}_{i,t}\}$.

Problem Setting

Agents behavior and feedback

- Principal chooses payment $c_{t,i}$ for arm i at time t;
- Agent θ_t would choose the arm $i_t = \arg\max_i \{\theta_t \cdot \hat{u}_{i,t} + c_{t,i}\}$;
- Each pull provides vector-valued outcomes equal to u_i, perturbed by independent noise.

Principal's goal

- Regret $r_t = (\max_i \theta_t \cdot u_i) - \theta_t \cdot u_{i_t}$ and payment $c_t = c_{t,i_t}$;
- Incentivize to minimize the cumulative regret while making a small 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 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 $O(N^2 e^{2/p})$.

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).

Notations

Phase

- Phase s starts when each arm has been pulled at least s times.

Payment-eligible

- Arm i has been pulled at most s times up to time t;
- The conditional probability of pulling arm i is less than $\frac{1}{\log(s)}$ given the current estimates $\hat{u}_{i,t}$.

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

Key technical lemma: an adaptive concentration inequality (Zhao et al. 2016);

Early Phases: bound by N;

Later Phases: exponentially unlikely as the phases advances;

Regret Analysis

When principal incentivizes: similar to the payment proof;

When agents pull myopically: We define a phase-dependent cutoff $\gamma(s(t))$ to further distinguish agents based on the regret.

- $r(t) \geq \gamma(s(t))$:
  - ⋆ this happens with exponentially decreasing probability;
  - ⋆ since θ_t has a compact support, the maximum regret is bounded by a constant;
- $r(t) \leq \gamma(s(t))$:
  - ⋆ not so many agents have near-ties preferences;
  - ⋆ the maximum regret is bounded above by $\gamma(s(t))$;

Question?

Thanks for your time!