Multi-armed Bandit Mechanism with Private Histories[*]

(Extended Abstract)

Chang Liu
Alibaba Group
chang.liu181@gmail.com

Qingpeng Cai
Tsinghua University
cqp14@mails.tsinghua.edu.cn

Yukui Zhang
Alibaba Group
yukui.zyk@alibaba-inc.com

1. INTRODUCTION

The fundamental challenge in bandit problem is the trade off between exploration and exploitation. To minimize the regret in a long period, an algorithm has to explore by actually choosing seemingly suboptimal arms so as to gather more information about them. The exploration obviously has higher short-term regrets. In recommendation of new items, the lifecycles of these items are remarkably short. We try to gather information as plenty as possible in an exploration process and expect we can get rewards in the following exploitation, but the gains are tiny and some newer items come in and next exploration should be start. We must increase the intensity of exploration so as to gather information quickly, but this will draw more regrets.

1.1 Our approach

We present a multi-armed bandits with private histories (PH-MAB) model that combines mechanism design and multi-arm bandit algorithm. In this model each seller called agent preserves a private history, and reports their histories to the designer. In each round the mechanism solicits the reports and the record of pulling arms in previous rounds, outputs a randomized arm selection rule. Correspondingly, the platform gets a reward from the selected arm and pays agents. By the well-known revelation principle[13], it is without loss of generality to consider only truthful mechanisms. This model can be viewed as a variant of resource allocation setting[14], and the arm selection rule can be considered as an allocation rule. However, it is different from the traditional model in mainly two aspects. Firstly the utility of an agent is the expected payment she receives conditioned on her private history, which not equals the expected value minus the payment to the designer. Secondly the payment of each round in our setting is not only decided by the reports, but also by the rewards that agents and the platform will observe. We define the consistency, a mechanism is consistent with a MAB algorithm means that the selection rule of arms in the mechanism is the same as the algorithm. We focus on designing truthful mechanisms that are consistent with ϵ-greedy algorithm maximizing the revenue. The revenue equals the sum of rewards minus the sum of payments.

1.2 Related work

We are not the first to combine mechanism design and MAB, the idea is discussed in[15, 7, 5, 6, 11, 12, 1, 2].

We are not the first to study mechanism design in e-commerce and reputation sites either. A line of work [8, 9, 10, 4, 3, 16] study the topic.

2. SETTING

Let there be K arms, $[K] = \{1, 2, ..., K\}$. The reward of arm k, r_k is drawn from a probability distribution, with density function , $f(x, \theta_k)$, where θ_k is a deterministic unknown parameter. The prior distributions of all parameters are i.i.d, and denoted by D. We assume that r_k is drawn from a Bernoulli distribution $B(1, \theta_k)$, and D is a Beta distribution, $Beta(1, 1)$.

Let $H_k = \{r_{k,t}\}$ denote the history of agent $k(r_{k,t}$ means the reward of arm k of $t-$th pulling in the first stage), $H_k^{'}$ denote the reported history of agent k and $H_{-k}^{'}$ denote the reported histories of agents except from agent k. Also, $H = (H_1, ..., H_K)$ and $H^{'} = (H_1^{'}, ..., H_K^{'})$.

Definition 1. *A multi-armed bandits with private histories (PH-MAB) mechanism f is a process of T rounds:*

- *Before round 1, each agent k reports their private pulling history of his own arm to the designer, i.e, $H_k^{'}$.*
- *At each round $t(1 \leq t \leq T)$, the mechanism f solicits the reports $H^{'}$, and the records of the pulling arm and the rewards in previous rounds, denoted by R_{t-1}, outputs a randomized rule selecting the arm with probability distribution $f(t, H^{'}, R_{t-1})$.*

  After receiving the reward of the arm $r_t(a_t)$, each agent k receives the payment from the mechanism

  $p(k, t, H^{'}, R_{t-1}, r_t(a_t))$.

The utility of each agent k with a PH-MAB mechanism f and a report $H^{'}$ is the expectation of the sum of payments the agent gets , conditioned on the history of agent k.

Our aim is to design a PH-MAB mechanism that is individual rational, truthful-in-expectation and consistent with $\epsilon-$greedy algorithm(Using the same randomized rule), in order to maximize the revenue and minimize the regret.

The revenue of a PH-MAB mechanism f, given the truthfully reports H, is the difference between the expected sum of rewards and the expected payments to agents, denoted by

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

$Rev(f, H)$. The regret is the difference between the maximum expectation of the sum of rewards of optimal algorithms and the expected sum of rewards of this mechanism, denoted by $Reg(f, H)$.

3. RESULTS

Given a history of agent i, H_i, the posterior distribution of θ_i is $p(\theta_i|H_i)$. Then we can calculate the expected reward for arm i conditioned on the history H_i, denoted by $R(H_i)$.

Similarly, Let $R(H_{-i})$ denote the maximum of the expected reward of other arms conditioned on their own history except i, i.e, $R(H_{-i}) = Max_{j \neq i} R(H_j)$.

Let $R_{k,t} = \{r_{t'} | a_{t'} = k, t' < t\}$ denote the record of arm k before round t, and the total histories before round t is $H^{all}_{k,t} = \{H^{'}_k, R_{k,t}\}$. We get a class of truthful mechanims.

MECHANISM 1. *We first generate a random variable I drawn from a Bernoulli Distribution $B(1, \epsilon)$. Arm selection:*

$$a_t = \begin{cases} randomly - selection & I = 1 \\ argmax_k(R(H^{all}_{k,t})) & I = 0 \end{cases} \tag{1}$$

Payments: Let $i = argmax_k(R(H^{all}_{k,t}))$,

$$p(k, t, H^{'}, f, R_{t-1}) = \begin{cases} \lambda_4 r_t & I = 1 \\ \lambda_1 r_t - \lambda_2 R(H^{all}_{-i,t}) & k = i \& I = 0 \\ \lambda_3 R(H^{all}_{i,t}) & k \neq i \& I = 0 \end{cases} \tag{2}$$

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are nonnegative parameters, $\lambda_1 = \lambda_2 + \lambda_3$.

We combine the reports of agents $H^{'}_k$(Private History) and the record $R_{k,t}$(Common History) as the whole history.

THEOREM 1. *Mechanism 1 is truthful when the size of each seller's private history and the size of reported history of each seller are the same.*

4. EXPERIMENTS

Taobao App is one of the most popular online shopping sites around the world. There is a scenario named *Daily New Goods* in Taobao App's homepage to exhibit and sale these items. In this scenario, we have a limited opportunity to show these items everyday, and what we concern is the total transactions. So we try to find the best or similar items with the highest conversion rate, and to give them more opportunity to be exhibited. The platform generally uses a MAB algorithm on different human groups to deliver different items. In this section, we simplify the problem as a pure MAB algorithm on single human group. We regard the seller of every item is an agent, also we suppose that every seller has only one item. We use the data in this scenario to implement the following experiments.

To estimate the performance, we focus on two objectives. The first one is regret, an index of the loss of social welfare including platform and all the sellers. The other is the revenue of platform. We assume the probability of an item being conversed is $\mu_k \in (0, 1)$, which is deterministic, but unknown. Hence the reward of selecting this arm is under a Bernoulli distribution $r_k \sim B(1, \mu_k)$. We select 10 popular items from the *Daily New Goods* scenario. Their frequency of exposures in one day are over 10 thousands. Because the parameters of items are unknown, we use the posteriori estimation $\frac{\sharp transaction}{\sharp exposure}$ to be μ_k. We assume the first 50 exposures are private information of the sellers. In our implementation, there are four mechanisms to be compared:

1. ϵ-greedy with $\epsilon = 0.02$ (Ignore the private information, the agents(sellers) are not involved in the mechanism, do exploration from the virgin paper)
2. ϵ-greedy with $\epsilon = 0.01$
3. Mechanism $1(\epsilon = 0.02)$
4. Mechanism $1(\epsilon = 0.01)$

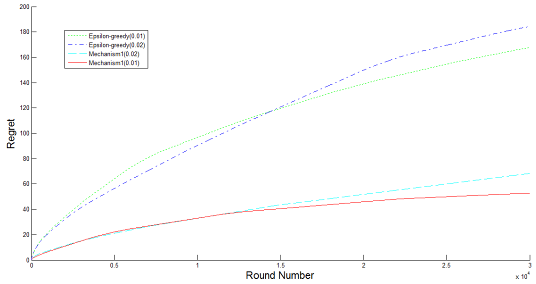

Figure 1: Accumulative regret

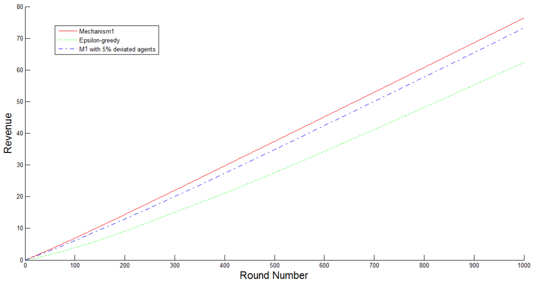

Figure 2: Accumulative platform's revenue

The parameters are set as $\lambda_1 = 0.01$, $\lambda_2 = 0.008, \lambda_3 = 0.002, \lambda_4 = 0$. We plot the accumulative regrets in every round in Fig. 1. We can conclude that the regret of Mechanism 1 is about one third of ϵ-greedy. In the comparison of different ϵ, we can see that larger ϵ will make the exploration process quicker so as to get less regret in short term, however in the other hand, larger ϵ lead into more regret in long term. In Mechanism 1, the short term gap between larger and smaller ϵ is greatly reduced. So we can choose smaller ϵ to earn the benefit in long term. Furthermore, we are also interested in the platform's revenue.We test the expected revenus of Mechanism 1, i.e, $E_H Rev(f, H)$. From Fig. 2, we can see that the revenue of platform exceeds ϵ-greedy significantly. In the end of 1000 rounds, the revenue of Mechanism 1 is 76.32(red line), which is 22.6% greater than that of ϵ-greedy(green line, 62.25).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China Grant 61303077, 61561146398, a Tsinghua University Initiative Scientic Research Grant and a China Youth 1000-talent program.

REFERENCES

[1] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88. ACM, 2009.

[2] G. Bahar, R. Smorodinsky, and M. Tennenholtz. Economic recommendation systems. *arXiv preprint arXiv:1507.07191*, 2015.

[3] Q. Cai, A. Filos-Ratsikas, C. Liu, and P. Tang. Mechanism design for personalized recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 159–166. ACM, 2016.

[4] B. Faltings. Using incentives to obtain truthful information. In *Agents and Artificial Intelligence*, pages 3–10. Springer, 2013.

[5] P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22. ACM, 2014.

[6] L. Han, D. Kempe, and R. Qiang. Incentivizing exploration with heterogeneous value of money. In *International Conference on Web and Internet Economics*, pages 370–383. Springer, 2015.

[7] C.-J. Ho, A. Slivkins, and J. W. Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, 55:317–359, 2016.

[8] S. Johnson, J. W. Pratt, and R. J. Zeckhauser. Efficiency despite mutually payoff-relevant private information: The finite case. *Econometrica: Journal of the Econometric Society*, pages 873–900, 1990.

[9] R. Jurca and B. Faltings. Truthful opinions from the crowds. *ACM SIGecom Exchanges*, 7(2):3, 2008.

[10] R. Jurca, B. Faltings, et al. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209, 2009.

[11] I. Kremer, Y. Mansour, and M. Perry. Implementing the wisdom of the crowd. *Journal of Political Economy*, 122(5):988–1012, 2014.

[12] Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582. ACM, 2015.

[13] R. B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

[14] Y. Shoham and K. Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations.* Cambridge University Press, 2008.

[15] A. Singla and A. Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1167–1178. ACM, 2013.

[16] J. Zhang, R. Cohen, and K. Larson. Combining trust modeling and mechanism design for promoting honesty in e-marketplaces. *Computational Intelligence*, 28(4):549–578, 2012.