Thompson Sampling Based Multi-Armed-Bandit Mechanism Using Neural Networks

by

Manisha Padala, Sujit Prakash Gujar

in

International Foundation for Autonomous Agents and Multiagent Systems AAMAS2019

Report No: IIIT/TR/2019/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2019

Thompson Sampling Based Multi-Armed-Bandit Mechanism Using Neural Networks

Extended Abstract

Padala Manisha
International Institute of Information Technology
Hyderabad, Telangana
manisha.padala@research.iiit.ac.in

Sujit Gujar
International Institute of Information Technology
Hyderabad, Telangana
sujit.gujar@iiit.ac.in

ABSTRACT

In many practical applications such as crowd-sourcing and online advertisement, use of mechanism design (auction-based mechanisms) depends upon inherent stochastic parameters which are unknown. These parameters are learnt using multi-armed bandit (MAB) algorithms. The mechanisms which incorporate MAB are referred to as Multi-Armed-Bandit Mechanisms. While most of the MAB mechanisms focus on frequentist approaches like upper confidence bound algorithms, recent work has shown that using Bayesian approaches like Thompson sampling results in mechanisms with better regret bounds; although lower regret is obtained at the cost of the mechanism ending up with a weaker game theoretic property i.e. Within-Period Dominant Strategy Incentive Compatibility (WP-DSIC). The existing payment rules used in the Thompson sampling based mechanisms may cause negative utility to the auctioneer. In addition, if we wish to minimize the cost to the auctioneer, it is very challenging to design payment rules that satisfy WP-DSIC while learning through Thompson sampling.

In our work, we propose a data-driven approach for designing MAB-mechanisms. Specifically, we use neural networks for designing the payment rule which is WP-DSIC, while the allocation rule is modeled using Thompson sampling. Our results, in the setting of crowd-sourcing for recruiting quality workers, indicate that the learned payment rule guarantees better cost while maximizing the social welfare and also ensuring reduced variance in the utilities to the agents.

KEYWORDS

Mechanism Design; MAB; Neural Networks

ACM Reference Format:
Padala Manisha and Sujit Gujar. 2019. Thompson Sampling Based Multi-Armed-Bandit Mechanism Using Neural Networks. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the real world we often encounter situations where we have to chose among competing and strategic agents to achieve a specific goal. These agents hold private information which is crucial to the decision. Misreporting of the private information may lead to a sub-optimal outcome. Hence, we need to design a mechanism that ensures truthful reporting of the private information, [12] chap. 9.

Designing appropriate mechanism involves designing an *allocation rule* and a *payment rule*. There are many settings like crowd-sourcing, online advertisement etc, where auction design relies on environmental parameters which neither the agent nor the hiring agency is sure about. For example, in crowd-sourcing, the actual quality of the agent is known to neither the agent nor the auctioneer. Such parameters are not deterministic but are subject to various environmental conditions, hence are stochastic or could even be adversarial. In this paper, we restrict to stochastic settings. In such a setting, it becomes necessary to figure out the average values of these parameters through exploration and at the same time ensure that the agents do not misreport their cost. However in the presence of such learning algorithms, the strategic agents have more freedom to manipulate. Hence it is required to design novel mechanisms that also learn the environmental parameters. Such mechanisms are referred to as *Multi-Armed-Bandit (MAB) Mechanisms* [2, 4, 7, 8, 10, 11, 13].

In MAB, we consider each of the agents or the advertisements as an arm. The auctioneer repeatedly selects an arm in order to observe its performance and get an estimate of the expected reward from that arm. The performance of an MAB algorithm is captured through the notion of *regret*, which is the difference between the expected reward from the optimal arm and the expected reward from the algorithm. There are two popular algorithms in MAB, one algorithm is based on the frequentist approach called as Upper Confidence Bound (UCB) algorithm [1]. The other technique follows the Bayesian approach and is called Thompson Sampling [14]. Thompson sampling is known to achieve lower regret than the other algorithms. When designing an MAB based mechanism, we impose restrictions on the allocation rule to ensure truthfulness of the payment rule.This in turn affects the regret of the algorithm. The payment rule could be 1) Deterministic, which leads to high regret in social welfare [4, 6], or 2) Randomized, which achieves low regret but higher variance in the utilities of the agents [3, 5].

In this paper, our goal is to design MAB based mechanisms which ensure truthful reporting of the strategic values and achieve *Allocative Efficiency (AE)*. We consider the problem of selecting high quality service providers (agents) such that the welfare obtained by the hiring agency (auctioneer) is maximized at minimal cost. The welfare is dependent on the Quality of Service (QoS) provided by the agent and is a stochastic quantity. This is a reverse auction setting where the auctioneer pays the selected agent for its service. The auctioneer wants to minimize the payments to the cost

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

Payment	TSM-D	TSM-R	TSM-NN
Type	Deterministic	Randomized	Randomized
WP-DSIC	No	Yes	Yes
EPIR	No	Yes	Yes
Variance in utility of the optimal agent	3 (highest)	2	1 (lowest)
Cost Index	3 (highest)	2	1 (lowest)

Table 1: Summarizing properties satisfied by the three mechanisms

optimal agents at each round (AE is satisfied). Note that, this is different from Myerson's optimal auction design, which in our setting would be same as minimizing the payments to the agents without guaranteeing AE. In order to evaluate the payments made by our mechanism, we introduce the notion of *Cost Index (CI)*. It is the expected value of the ratio of payments made by the mechanism to the optimal payments. In our setting, we desire CI should to be as low as possible ideally near one.

Ghalme et al. [9] propose two Thompson sampling based MAB mechanisms TSM-D and TSM-R for solving the above problem of crowd-sourcing. The primary aim in their paper is to achieve low regret for the auctioneer while ensuring reduced variance in the utilities of the agents. The lower the regret achieved by the learning algorithm, the more likely it is for the mechanism to achieve AE. As discussed in [9], ensuring ex-post dominant strategy incentive compatibility (DSIC) [4, 8] is difficult as it unlikely for the agents to have full knowledge of the future events. Instead their mechanism ensures a weaker notion of truthfulness, called Within-Period DSIC (WP-DSIC). The payment rules in TSM-D and TSM-R are designed just to ensure WP-DSIC, but the auctioneer's payments to the agents are not minimized. The mechanisms also ignore the possibility of the payment exceeding the welfare to the auctioneer. Our analysis shows TSM-D pays very high as compared to the welfare and there is non-zero probability of the payments being higher than welfare in TSM-R. Analytically coming up with payment rules in Thompson sampling based MAB settings is challenging. With these shortcomings of TSM-D and TSM-R in sight, we propose a data-driven mechanism which learns the optimal payment rule to minimize the payments while ensuring high social welfare. We also ensure this mechanism is WP-DSIC and Ex-post Individual Rationality (EPIR). What we propose, is a *neural network and multi-armed based mechanism design* (NN). Refer to Table 1 for a comparison among the different payment rules. *Contributions*: i) Data-driven approach for learning the payment rule in stochastic setting. ii) The payment rule is learned to minimize the total payment while maximizing welfare. iii) The payment rule enjoys the desirable properties of within-period DSIC and ex-post IR. iv) The payment is ensured not to exceed the welfare. v) The variance in the utility to the agents decreases with time.

2 RESULTS

In this section, we discuss the different experiments conducted for comparison with the existing approaches, TSM-R and TSM-D. In the extended version we explicitly define the properties the mechanism satisfies. We discuss the network design required for the same and other training details.

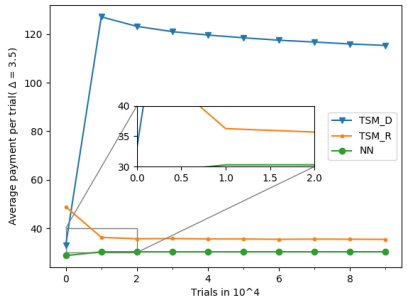

Figure 1: Average payments Vs Trials

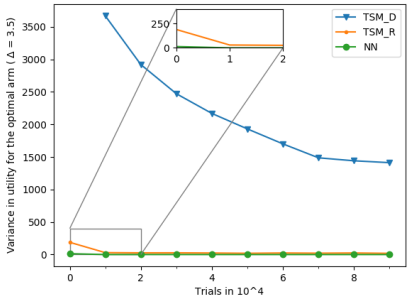

Figure 2: Variance in Utility Vs Trials

In all the experiments, we fix the bids to $\{30.0, 35.0\}$. Figure 1 shows that the average payments by the NN are consistently low, although higher than 30 to maintain EPIR leaving the first 10^4 trials. Figure 2 shows the variance in utility to the optimal agent across the 1000 iterations for a fixed bid.

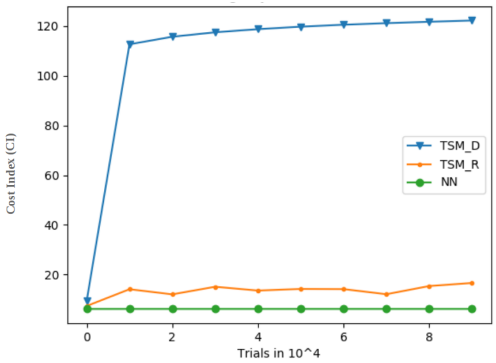

Figure 3: Cost Index Vs Trials

From the plot in Figure 3, it is clearly indicative that NN has the least CI in all the rounds t. TSM-D has the highest CI whereas TSM-R has considerable value although higher than NN.

REFERENCES

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 2-3 (May 2002), 235–256. https://doi.org/10.1023/A:1013689704352

[2] Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. 2012. Dynamic pricing with limited supply. In *Thirteenth ACM Conference on Electronic Commerce*. ACM, 74–91. https://doi.org/10.1145/2229012.2229023

[3] Moshe Babaioff, Robert D. Kleinberg, and Aleksandrs Slivkins. 2010. Truthful mechanisms with implicit payment computation. In *Eleventh ACM Conference on Electronic Commerce*. ACM, 43–52.

[4] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. 2009. Characterizing truthful multi-armed bandit mechanisms: extended abstract. In *Tenth ACM Conference on Electronic Commerce*. ACM, 79–88.

[5] Satyanath Bhat, Shweta Jain, Sujit Gujar, and Yadati Narahari. 2015. An Optimal Bidimensional Multi-Armed Bandit Auction for Multi-unit Procurement. In *Fourteenth International Conference on Autonomous Agents and Multiagent Systems, AAMAS'15*. 1789–1790.

[6] Nikhil R. Devanur and Sham M. Kakade. 2009. The price of truthfulness for pay-per-click auctions. In *Tenth ACM Conference on Electronic Commerce*. 99–106.

[7] Ghalme Ganesh, Jain Shweta, Gujar Sujit, and Narahari Y. 2016. A Deterministic MAB Mechanism for Crowdsourcing with Logarithmic Regret and Immediate Payments. In *Proceedings of Fifteenth International Confrence on Autonomous Agents Multi-Agent Systems (AAMAS'16)*. To Appear.

[8] Nicola Gatti, Alessandro Lazaric, and Francesco Trovò. 2012. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Thirteenth ACM Conference on Electronic Commerce*. 605–622.

[9] Ganesh Ghalme, Shweta Jain, Sujit Gujar, and Y. Narahari. 2017. Thompson Sampling Based Mechanisms for Stochastic Multi-Armed Bandit Problems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 87–95. http://dl.acm.org/citation.cfm?id=3091125.3091143

[10] Shweta Jain, Sujit Gujar, Onno Xoeter, and Y. Narahari. 2014. A Quality Assuring Multi-Armed Bandit Crowdsourcing Mechanism with Incentive Compatible Learning. In *Thirtheenth International Conference on Autonomous Agents and Multiagent Systems*. 1609–1610.

[11] Shweta Jain, Balakrishnan Narayanaswamy, and Y. Narahari. 2014. A Multiarmed Bandit Incentive Mechanism for Crowdsourcing Demand Response in Smart Grids. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*. 721–727. http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8355

[12] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA.

[13] Akash Das Sharma, Sujit Gujar, and Y. Narahari. 2012. Truthful multi-armed bandit mechanisms for multi-slot sponsored search auctions. *Current Science* Vol. 103 Issue 9 (2012), 1064–1077.

[14] William R. Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 3/4 (1933), pp. 285–294. http://www.jstor.org/stable/2332286