Decision Market Based Learning For Multi-agent Contextual Bandit Problems

Wenlong Wang
Massey University
Auckland, New Zealand
stan.wenlong.wang@gmail.com

Thomas Pfeiffer
Massey University
Auckland, New Zealand
T.Pfeiffer@massey.ac.nz

ABSTRACT

Information is often stored in a distributed and proprietary form, and agents who own this information are often self-interested and require incentives to reveal it. Suitable mechanisms are required to elicit and aggregate such distributed information for decision-making. In this study, we use simulations to investigate the use of decision markets as mechanisms in a multi-agent learning system to aggregate distributed information for decision-making in a contextual bandit problem.

KEYWORDS

Multi-agent systems, Prediction markets, Federated learning

ACM Reference Format:
Wenlong Wang and Thomas Pfeiffer. 2024. Decision Market Based Learning For Multi-agent Contextual Bandit Problems. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In many decision-making tasks, relevant information is distributed over multiple parties. To optimise decision-making, multi-agent learning systems are required to obtain, aggregate and learn from such distributed information. When the agents' information is private but their objectives are aligned, this problem has been extensively researched as Federated Learning [13, 18, 21]. However, when their objectives are not aligned (i.e., the agents are selfish and optimise their own independent objective functions), incentives may be required to induce the agents to reveal their information. For efficient multi-agent learning in such a situation, the rewards must be designed so that as agents maximise their rewards in the training phase, the system's overall performance is also optimised.

Consider, for example, a recommendation system that aims to optimise advertisement targeting by using information from multiple sources (e.g., Google, Facebook and Amazon). Such information could involve the companies' different user profile data for the targeted person, which the companies have no interest to reveal. The system, therefore, needs to elicit information in a form that is agreeable to the information source (e.g., recommendations for the task at hand, rather than complete user profiles) and needs to provide fair rewards for these contributions. Such rewards can be monetary but need to be designed such that each information source can learn from the realised rewards and while maximising its rewards, the performance of the recommendation system improves as well.

In this work, we develop a multi-agent learning system that provides agents with rewards that align the agents' objectives with the system's objectives. We test the system in simulations of learning in a multi-armed Bandit problem where contextual information is distributed over multiple agents. A full description of the approach, the methods used for the simulations, and our results is provided in [22].

Our approach is based on decision markets. Decision markets are extensions of prediction markets that allow to elicit forecasts from self-interested agents for the purpose of decision making [11]. Similar to prediction markets, they employ scoring rules to define incentives for self-interested agents to reveal their information [10, 12]. However, in addition, they use decision rules to make decisions based on the elicited forecasts. The core challenge in such a situation is quantifying scores for counterfactual outcomes contingent on actions not taken. Stochastic decision rules, proposed by Chen et al., which select actions stochastically with strictly positive probabilities that depend on forecasts, have been shown to allow to define proper scores for incentive-compatible elicitation from selfish agents [6, 8]. This is in contrast to deterministic rules, such as max decision rules, which deterministically select the action with the most desirable outcomes. Deterministic rules have been shown to potentially induce strategic manipulation of forecasts by selfish agents [7, 19].

2 ALGORITHM

We study a multi-agent multi-armed contextual Bernoulli bandit problem, where one agent (referred to as the principal) decides between multiple alternative actions and receives a corresponding reward that evaluates the quality of the decision. The context, however, is distributed over multiple self-interested agents. In the system we investigate here, the principal uses a decision market to sequentially elicit probabilistic reports for the Bernoulli outcomes of the available actions from the agents (see Figure 1). In each time step, the principal receives an initial set of prior probability distributions for the outcomes of each action. It then selects an agent to alter this report. The agent will be scored for this altered report using a decision scoring rule. The principal then adopts this report and selects the next agent to alter it, and this process is repeated until the last agent has been selected. Once all agents have been queried, the principal uses the final report (from the last agent) and a decision rule to select an action. When the selected action is executed and the outcome is observed, the scores for all agents can be calculated, and the time step concludes.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

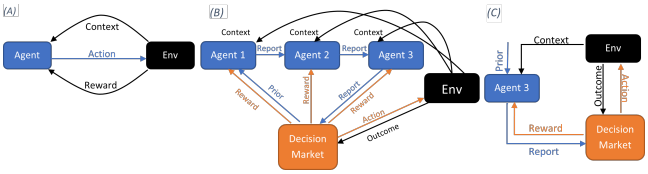

Figure 1: Decision markets based multi-agent bandit system. (A) shows a diagram for a regular contextual bandit problem. (B) shows a multi-agent contextual bandit problem with a decision market, which is the main design of this paper. An action is selected by a decision market, which aggregates distributed posterior probabilities reported from agents. The decision market assigns a reward to each agent based on the quality of their reports. (C) shows a contextual bandit problem with a continuum-arm space in agent 3's perspective.

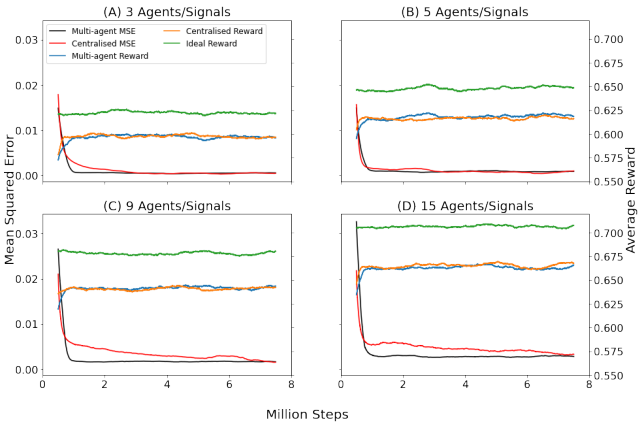

Figure 2: System Performance Comparison in Multi-Agent and Centralised Systems. Panels (A)-(D) depict simulations with 3, 5, 9, and 15 signals. In the multi-agent system, each agent receives one signal, while in the centralised counterpart, a single agent receives all signals. The black and red lines represent running averages of mean squared errors for multi-agent and centralised systems, respectively. The green line indicates the average received reward for a principal using a correct Bayesian model with all available information. The blue line displays the actual reward for multi-agent systems, while the orange line represents the reward for centralised systems. Notably, errors and rewards between centralised and distributed systems exhibit striking similarities. Although rewards are lower than the Bayesian model, the discrepancy arises from employing a stochastic decision rule in both multi-agent and centralised systems.

Note that while the principal faces a contextual Bernoulli bandit problem with discrete arm space, every other agent faces a continuous contextual bandit problem, where the agent's action is its probabilistic report to the principal (see Figure 1). To clearly distinguish between these two contextual bandit problems, we refer to the context of the Bernoulli bandit problem as the system's context, and the context in the continuous bandit problem of the individual agents as the agent's context. The agent's context consists of the signals it receives from the system's environment, and the previous report it receives from the principal or the previous agent. The system's context consists of all signals that are received by the agents from the environment, including the priors that the principal receives from the environment. We want to emphasise that the principal in this system is not a learning agent but an entity that employs decision markets for decision making. However, the agents can learn to use the context to generate reports that maximise the score they receive. We test if, in such a system, the agents can efficiently learn such that the principal's performance in the Bernoulli bandit problem improves.

3 SIMULATION RESULTS

We compare a multi-agent system with a centralised agent. In the multi-agent system, signals are distributed across the individual agents, while in the centralised system, there is a single agent that receives all signals. In both cases, a stochastic decision rule is used. The results show that the multi-agent contextual bandit system performs as well as the centralised system.

As shown in Figure 2, we observe that the mean square error (MSE) of the final report decreases rapidly and stabilises close to zero in both multi-agent and centralised systems. The MSE declines faster in the multi-agent system, compared to the centralised counterpart when the agent or signal number is high. Once converged, the average rewards for both systems are very similar, with the reward being defined as one when the selected action leads to the preferred Bernoulli outcome, and zero when it is not. Note that the gap between the actual reward and the ideal reward is due to the nature of stochastic decision rules, which assigns a positive probability to select a sub-optimal action. The performance will be close to the ideal reward if we account for the disadvantage of the stochastic decision rules.

We also use simulations to investigate strategies learned under deterministic decision rules. We observe strategic manipulation of the probabilistic reports in simulations with single and multiple agents, which aligns with expectations from theory. For a detailed description of these results, see [22].

4 CONCLUSION

We explore the use of decision markets for contextual bandit learning in a multi-agent system. In this system, contextual information is distributed among several self-interested agents, each possessing exclusive ownership of their information. These agents require incentives to disclose and learn to interpret the contextual data.

Our simulations demonstrate that the decision market-based multi-agent system can effectively train self-interested agents, achieving a performance on a par with a centrally trained counterpart that has access to all pieces of the same contextual information.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Marsden Fund (Grant nr. MFP-MAU1710).

REFERENCES

[1] Alekh Agarwal, John Langford, and Chen-Yu Wei. 2020. Federated residual learning. *arXiv preprint arXiv:2003.12880* (2020).

[2] Pragnya Alatur, Kfir Y Levy, and Andreas Krause. 2020. Multi-player bandits: The adversarial case. *Journal of Machine Learning Research* 21 (2020), 77.

[3] Andreea Bobu, Dexter RR Scobee, Jaime F Fisac, S Shankar Sastry, and Anca D Dragan. 2020. Less is more: Rethinking probabilistic models of human behavior. In *International Conference on Human-Robot Interaction*. 429–437.

[4] Nicolò Cesa-Bianchi, Tommaso Cesari, and Claire Monteleoni. 2020. Cooperative online learning: Keeping your neighbors updated. In *Algorithmic learning theory*. PMLR, 234–250.

[5] Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. 2016. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*. PMLR, 605–622.

[6] Yiling Chen, Ian Kash, Mike Ruberry, and Victor Shnayder. 2011. Decision Markets with Good Incentives. In *Internet and Network Economics*. 72–83. https://doi.org/10.1007/978-3-642-25510-6

[7] Yiling Chen and Ian A. Kash. 2011. Information elicitation for decision making. In *10th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 1. 161–168.

[8] Yiling Chen, Ian A. Kash, Michael Ruberry, and Victor Shnayder. 2014. Eliciting predictions and recommendations for decision making. *ACM Transactions on Economics and Computation* 2, 2 (2014), 1–27. https://doi.org/10.1145/2556271

[9] Zhongxiang Dai, Yao Shu, Arun Verma, Flint Xiaofeng Fan, Bryan Kian Hsiang Low, and Patrick Jaillet. 2022. Federated Neural Bandit. *arXiv e-prints* (2022), 1–27.

[10] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (3 2007), 359–378. https://doi.org/10.1198/016214506000001437

[11] Robin D. Hanson. 1999. Decision Markets. *IEEE Intelligent Systems* 14, 3 (1999), 16–19.

[12] Robin D. Hanson. 2003. Combinatorial information market design. *Information Systems Frontiers* 5, 1 (2003), 107–119. https://doi.org/10.1023/A:1022058209073

[13] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint* (2016), arXiv:1610.05492.

[14] Nathan Korda, Balazs Szorenyi, and Shuai Li. 2016. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*. PMLR, 1301–1309.

[15] Tan Li, Linqi Song, and Christina Fragouli. 2020. Federated Recommendation System via Differential Privacy. In *2020 IEEE International Symposium on Information Theory (ISIT)*. 2592–2597. https://doi.org/10.1109/ISIT44484.2020.9174297

[16] R Duncan Luce. 1977. The choice axiom after twenty years. *Journal of mathematical psychology* 15, 3 (1977), 215–233.

[17] R Duncan Luce. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

[18] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[19] Abraham Othman and Tuomas Sandholm. 2010. Decision rules and decision markets. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol. 1. 625–632.

[20] Clémence Réda, Sattar Vakili, and Emilie Kaufmann. 2022. Near-optimal collaborative learning in bandits. *Advances in Neural Information Processing Systems* 35 (2022), 14183–14195.

[21] Chengshuai Shi and Cong Shen. 2021. Federated Multi-Armed Bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9603–9611. https://doi.org/10.1609/aaai.v35i11.17156

[22] Wenlong Wang and Thomas Pfeiffer. 2022. Decision Market Based Learning For Multi-agent Contextual Bandit Problems. *arXiv preprint arXiv:2212.00271* (2022).

[23] Jialin Yi and Milan Vojnovic. 2023. Doubly adversarial federated bandits. In *International Conference on Machine Learning*. PMLR, 39951–39967.

[24] Jialin Yi and Milan Vojnovic. 2023. On Regret-optimal Cooperative Nonstochastic Multi-armed Bandits. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1329–1335.

[25] Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. 2021. Federated bandit: A gossiping approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 5, 1 (2021), 1–29.