THE UNIVERSITY OF HONG KONG

Application for admission to the
Master of Philosophy (MPhil) and Doctor of Philosophy (PhD)
programmes in the Faculty of Science
Research Proposal

Du Junye (junyedu2-c@my.cityu.edu.hk)

1. Applicant's name

Du Junye

2. Title of proposed research project

Neural Contextual Bandit under Assortment Selection settings

3. Summary of research

Maximum 250 words, understandable by a non-specialist.

The 'Contextual Bandits Problem' is an extension of classic bandit problem where the decision-maker is provided with contextual information when making choices. The topic has a wide range of applications in fields like online advertising and retailing, and has been extensively studied in the field of machine learning. In this scenario, at each round, the agent chooses between K actions which will result a reward and goal of it is to maximize the expected cumulative rewards over T rounds. Previous research mainly focuses on situations under assumptions like linear reward function, which may fail to hold in practice.

Thanks to the development of deep neural networks(DNN), which exhibit great generalization and representation power, we are now capable of expressing the underlying non-linear reward function. However, in most realistic situations, we are faced with assortment selection problem, the decision maker needs to select a subset $S \in N$, after which it could get a response dependent of the elements in S. This setting resembles real life scenario when we browse online shopping sites. The bandit problem could be seen as a special case of assortment selection since under naïve approach, we could treat each $size - k$ subset of N products as an independent arm of actions, but this will lead to a computationally inefficient algorithm with regret exponential in K. My research hopes to develop an algorithm which leverages the power of deep neural networks to approximate the probability of each item being selected into the subset and finally give a near-optimal regret guarantee.

4. Introduction

Provide the relevant background information and explain the motivation for your study.

The stochastic contextual bandit problem, an extension of Multi-armed bandits(MAB), falls under the category of online-decision making problems where the agent repeatedly interacts with the environment, aiming to maximize the expected cumulative reward. During round $t \in \{1, 2, \dots T\}$, the agent is presented with K actions, each represented by a d-dimensional action feature vector. The reward received after each choice is typically generated from an unknown function of the contextual variables plus random noise. The challenge of making the choice in contextual bandit problem is how to balance between the exploration and exploitation, the common balancing strategies for addressing this issue could be classified into three categories: ϵ-greedy approach [3], upper confident bound(UCB) approach [2] and Thompson sampling(Ts) [6] approach. The simplest method is ϵ-greedy strategy, a widely used approach in reinforcement learning. With probability of ϵ, the agent will make

a random selection. The UCB methods employs uncertainty to guide exploration where it calculates an upper confidence bound for every action basing on the confidence interval and the action with highest UCB will be selected. The Thompson Sampling method approaches the bandit problem from the standpoint of statistics, which assumes a prior distribution over the possible rewards for each action.

Most research deal with the situations when the reward function is linear, the renowned algorithm – LinUCB [5], proposed by Yahoo Lab, made a great success in the field of news article recommendation. However, the assumption of linear reward function is often too strong to apply to the realistic problems. In a gesture to solve this, the deep neural networks(DNNs) have been introduced to approximate the underlying reward function. Given that DNNs have strong representation power, the input to the last layer could be viewed as the lower dimensional feature map. Building on this insight, Zhou et al. (2020) developed NeuralUCB [10] algorithm, which leveraged a neural network-based random feature mapping to construct an upper confidence bound(UCB) for the reward. Thanks to the neural tangent kernel(NTK) [4] methods by Jaco et al. (2018), under standard assumption, the NeuralUCB algorithm could achieve $O(\tilde{d}\sqrt{T})$ regret where T is the round number. Since $\tilde{d}$ is the effective dimension of the NTK matrix, it could scale with $O(TK)$. As such, Xu et al. proposed Neural-LinUCB [9] algorithm, which decoupled the feature learning and exploration. In Xu's work, after the feature mapping, the regret bound of Neural-LinUCB achieved the same complexity as LinUCB.

However, the real-world application situation is much more complicated. For an online shopping platform with N product offerings, due to the constraints of display, a more practical strategy for the platform is to offer customers an assortment denoted as $S \in N$ of products at a time and the customer can choose whether to select one item or not. Since we could only observe whether the customer picks a specific item or not making a purchase, the principle of the assortment problem is to maximize the expected revenue of the corresponding assortment S, the objective is given by:

$$\max_{S \in S} R(S) = \max_{S \in S} \sum_{i \in S} r_i p_i(S) = \max_{S \in S} \sum_{i \in S} \frac{r_i v_i}{v_0 + \sum_{j \in S} v_j} \tag{1}$$

where the denominator term v_0 represents the chance that the customer buys nothing and v_i represents the likelihood that product i is selected in assortment S. A straightforward translation of this bandit-MNL (Multinomial Logit) problem into MAB bandits is to treat $\binom{N}{K}$ subsets as independent arms and apply standard UCB method, but this will naturally yields a combinatorial-bound constraint. Several methods have developed to solve the problem, Shipra et al. (2020) came up with an efficient exploration and exploitation algorithm [1] under no contextual information. Min-hwan et al. proposed several approaches to optimize the revenue under multinomial logit condition, like UCB-MNL [7] algorithm and TS-MNL with Optimistic Sampling[6] , which utilized the MLE to update parameter θ. Similar to MAB, Wang et al. (2023) utilized the deep neural networks and Residual structures to optimize feature-based assortment problem with their Res-Assort-Net [8]. However, this method could not learn the model and make decisions at the same time, thereby precluding its classification as a bandit problem. Up until now, there is no effective algorithm to solve neural contextual bandits under assortment settings problems, so the potential of applying DNNs and NTK tools in assortment problems deserves to be explored more deeply.

5. Objectives and hypotheses to be tested

List (using bullet points) the aims/objectives of the study and, where relevant, specify the hypotheses to be tested.

- My research hopes to propose an online neural network based algorithm to solve the bandit problem under assortment settings, which will have the capability to make decisions and concurrently update its parameters in real-time
- Since the utility measure v_i in my reaseach is represented by DNN, the expected regret on real-world datasets is supposed to be significantly reduced than under linear assumptions.
- Incorporate the decoupling mechanism into the deep neural network architecture, which could empower the last of layer of DNN to construct a UCB, facilitating a balanced approach between exploration and exploitation.
- Leverage recent advancements of optimiztaion and generalization in deep neural networks, hopefully, the NTK tools could be utilized to give a $\tilde{O}(\sqrt{NT})$ regret bound under standard assumption.

6. Literature review

Discuss the key publications in the proposed research area.

"Neural Contextual Bandits with UCB-based Exploration" by Zhou et al.(2020)

This foundamental paper studied the neural contextual bandits problems. Leveraging the strong representational capabilities of DNNs, the NeuralUCB algorithm employs a random feature mapping based on neural networks to establish an upper confidence bound(UCB). This paper stands out as the pioneering work to utilize the NTK tools to give a near-optimal regret guarantee with $\tilde{O}(\sqrt{T})$ regret under standard assumptions.

"Neural Contextual Bandits with Deep Representation and Shallow Exploration" by Xu et al.(2020)

This paper further explores the topic of neural contextual bandits, providing theoretical validation for the decoupling of representation learning and exploration. Building upon the NeuralUCB algorithm from Zhou's article, Xu's Neural-LinUCB algorithm performs a UCB-type exploration over the last layer of DNN. The paper proved a sublinear regret by virtue of NTK tools and demonstrated a better computational efficiency over NeuralUCB algorithm.

"Thompson Sampling for Multinomial Logit Contextual Bandits" by Min-hwan et al.(2019)

This paper presents a solution to the assortment selection problem under MNL conditions. By assuming that the feedback is given by a multinomial logit choice model, Min-hwan et al. proposed two Thompson sampling methods. The first attains a Bayesian regret of $\tilde{O}(d\sqrt{T})$ over T rounds while the other one achieves $\tilde{O}(d^3/2\sqrt{T})$ worst-case regret bound.

"A Neural Network Based Choice Model for Assortment Optimization" by Wang et al.(2023)

This paper introduces a neural network architecture to optimize the assortment revenue. Apart from this, the author devised an assortment formulation that can be efficiently solved by off-the-shelf integer programming solvers. Their Res-Assort-Net framework serves as an efficient feature encoder, which outperforms other optimization heuraistics when the underlying model is complex.

"Multinomial logit contextual bandits: Provable optimality and practicality" by Min-hwan et al.(2021)

This paper centers on the sequential assortment selection problem, where user choices are given by a multinomial logit(MNL) choice model. Given d-dimensional contextual information, the agent could present a size-K assortment to the user. The author proposed a UCB-MNL algorithm, which optimizes the revenue through solving the Maximum Likelihood Estimation(MLE) equation and achieves a $\tilde{O}(d\sqrt{T})$ regret over T rounds.

7. Materials and methods

Explain how you intend to undertake the research. There is no need to describe experimental apparatus or analyses in great detail: an indication of the approach to be adopted (with citation of sources) is adequate in most cases.

My research aims to build on the framework of Neural-LinUCB [9] framwork and extend the Multi-armed Banit(MAB) task to assortment selection. Similar to neural contextual bandit, the input of the neural network is feature vector of contexual information and the initialization of the network weight could refer to NeuralUCB [10]. However, in contrast to MAB where the agent selects action with highest upper confidence bound, my research will sort the item profit first and incrementally expanding the assortment till the turning revenue point. The UCB measure could refer to the UCB-MNL [7] algorithm, which updates a A_t matrix with $\sum_{i \in S_t} x_{ti} x_{ti}^T$ and ensures the exploration capability.

One notable difference between the MAB and assortment lies in the optimization of the linear parameter θ. The TS-MNL with Optimistic Sampling [6] presents a solution of minimizing the regularized MLE. In terms of

the regret analysis, my research wants to construct a NTK matrix [4], which makes no assumption of the utility function. The rough algorithm framework is outlined below:

Algorithm 1 Neural-Assort-LinUCB

Data: initial weight w_0 of neural network ϕ, regularization parameter λ, total steps T, episode length H, exploration parameter α_t

for $t = 1, ..., T$ **do**

Receive contextual feature vectors $\{x_{t,1}, ..., x_{t,N}\}$ and item profit $r_i, ..., r_N$

Sort the feature vectors according to decreasing profit

Initialize subset $S = \varnothing$

for $j = 1, ...N$ **do**

$v_j = \theta_{t-1}^T \phi(x_{t,j}; w_{t-1}) + \alpha_t ||\phi(x_{t,j}; w_{t-1})||_{A_{t-1}^{-1}}$

Add item index to S_t until $\sum_{i \in S_t} \frac{r_i v_i}{v_0 + \sum_{j \in S_t} v_j}$ decreases

Present assortment S_t and observe the purchase vector

Update A_t with:

$A_t = A_{t-1} + \sum_{i \in S_t} \phi(x_{t,i}; w_{t-1}) \phi(x_{t,i}; w_{t-1})^T$

Compute the regularized MLE θ_t by minimizing:

$-\sum_{i=1}^{t} \sum_{i \in S}^{n} y_{\tau i} \log p_{\tau i}(S_\tau, \theta) + \frac{\lambda}{2} ||\theta||^2$

if $mod(t, H) = 0$ **then**

backpropagation of neural network ϕ and update w_t

return w_t, θ

8. Anticipated outcome and value of the research

Explain the likely significance of your research and how it will impact the research field.

The assortment selection shows promising application prospects in various fields like retailing and E-Commerce. It seems that, up until now, there has been no article in the literature database that addresses assortment problems under the MNL condition using neural networks. If fortunately my work could perform well and show favorable regret bound, it could contribute to the progress of this domain.

References

[1] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 599–600, 2016.

[2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[3] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[4] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020.

[5] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[6] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9205–9213, 2021.

[8] Hanzhao Wang, Zhongze Cai, Xiaocheng Li, and Kalyan Talluri. A neural network based choice model for assortment optimization. *arXiv preprint arXiv:2308.05617*, 2023.

[9] Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration, 2020.

[10] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11492–11502. PMLR, 13–18 Jul 2020.