Research Statement

Qingyun Wu (qw2ky@virginia.edu)
Department of Computer Science, University of Virginia

The past decades have witnessed a prominent trend of adopting intelligent systems, such as recommendation systems and smart homes, into ordinary people's daily life. One key characteristic of such systems is the need of *online sequential decision making*: decisions have to be made when the learning agent only has incomplete knowledge about the world/environment. The consequences of such decisions will, in turn, contribute to the data the agent can collect, forming an interactive feedback loop between the agent and the world/environment. This makes conventional offline training based machine learning methods incompetent, i.e., the so-called explore-exploit dilemma. It urges us to move from the passive learning paradigm to a more interactive and proactive one. My research vision is to develop efficient interactive online machine learning solutions that can continuously learn from and interact with sophisticated real-world environments, where humans are involved. And my current research has built solid foundation for me to realize my vision. In a long-term, my research can impact a broad spectrum of applications that were impossible or immature before, including conversational recommendation systems, interactive online education systems, human-in-the-loop cyber-physical systems, and many more.

1 Research Achievements

As an important part of my long-term research vision, my current research builds upon the insight that interactive online learning agents are often situated in a dynamically changing and potentially collaborative or structured environment. Based on these insights, themes of my research to date include 1) Sample-efficient interactive online learning, specifically multi-armed bandit learning, in collaborative and structured environments; 2) Dynamic online decision making in non-stationary environments. My research has generate impactful contributions in various application scenarios, including personalized online recommendation [5, 6, 9–11], online education [2], online learning to rank [4], and online influence maximization [8].

1.1 Efficient online learning in collaborative and structured environments

Real-world environments can be highly structured. For example, users targeted by a recommendation system can be connected by social networks, which reveal potential affinities or similarity among the users; low-rank structure can thus be exploited due to the similarity between users or network assortativity. Such information can be effectively leveraged to reduce the sample complexity of online learning algorithms.

Efficient online learning from explicitly collaborative environments. I proposed collaborative contextual bandit learning solutions [5, 10], which successfully leverage the structural information into online decision making through collaborative learning. The key insight of my solutions is to capitalize on the information propagation among learning agents to reduce uncertainty and expedite the convergence of the online learning process. These work provide a theoretical understanding of the relationship between the available structural information and the sample efficiency of an online learning agent. Under this proposed framework, an up to $O(\sqrt{T}\log N)$ regret reduction can be achieved, where T is the total number of interactions and N is the number of users served in the system. This sheds light on the need and benefits of leveraging structure information into online decision making.

Efficient online learning from implicitly structured environments. In many cases, the structural information may not be explicitly available to the learners, indicating an implicitly structured environment. To address this problem, I proposed solutions [2, 5, 6, 8] that can infer the structure from interactions with the environment and further leverage it to improve online decision making. The key insights come from two sources: Firstly, domain knowledge, such as human behavior modeling, network assortativity, in different application scenarios can and should be leveraged to help regulate the collaborative effect or structural information in the learning environment; Secondly, interactive feedback from the environment can help to maintain an online estimate of the structural information. With these two insights, unique properties of different environments can be leveraged to reduce sample complexity.

1.2 Dynamic online decision making in non-stationary environments

In many real-world systems, the only constant is the forever changing user intent and preferences. Adjustments in decision making must be made in accordance with the changes in the environment; otherwise, sub-optimal decisions will result constantly.

Dynamic bandit learning in non-stationary environments. I proposed a suite of non-stationary contextual bandit learning solutions [7, 11], where instead of maintaining only one learning agent, a higher level learner is introduced to maintain a dynamic set of base learning agents. Insights from statistical hypothesis tests are used to adaptively add a learner, remove a learner, and form learner ensembles to fit the current environment. With these solutions, a near-optimal regret bound is achieved when learning in environments with abrupt changes; otherwise I prove that a linear regret is inevitable.

An unified approach through statistical hypothesis tests. The application of our proposed solutions are not limited to the non-stationarity detection in the time horizon, and they can be generalized multi-agent/environment cases, where the non-stationarity can be considered as the heterogeneity across different environments. I developed more general solutions based on online hypothesis tests to unify the online change detection in non-stationary environments and online clustering of learners. Through this unified approach with online hypothesis tests, much more flexible methodologies can be developed to efficiently handle environments where both non-stationarity and collaborative structure exist.

1.3 Research impact

My research has generated impacts on both improved theoretical guarantees and pioneering practical deployments. As a proof of the former, my theoretical analysis on the relationship between the structural information and sample efficiency in online learning [5, 10] sheds light on the need of the collaborative online learning and has inspired many follow-up studies. As proofs of the latter, my research in [7, 8] made the pioneering effort in effectively handling users' changing preferences in various real-world commercial recommendation systems, such as news recommendation in Yahoo, and lens recommendation in Snapchat. And one of our solutions [2] was successfully deployed on the largest MOOC platform in China, XuetangX, to recommend pop-up quiz questions for improving student engagement, where positive learning outcome was achieved. As a proof of both contributions, in [4] we were able to significantly improve the dueling bandit gradient descent based online learning to rank methods' convergence by carefully designed variance reduction techniques during online exploration. For its unique contribution to the Information Retrieval community, our work [4] has received the **SIGIR'2019 Best Paper Award**.

2 Future Research Plan

My current research has prepared me with deep working experience in efficient online learning for building intelligent systems, and proves an exciting potential of its impact on the next generation of artificial intelligence. Looking forward, I plan to conduct fundamental research to build intelligent systems that can benefit humans and society.

2.1 Long-term goals

Human-in-the-loop interactive machine learning. One long term goal of my research is to develop interactive online learning systems where humans are an integral part. The involvement of human in such interactive systems brings in both challenges and opportunities. Firstly, **ethical considerations**, such as privacy protection, fairness constraints, become a critical issue when designing the online learning agents. Secondly, the interactions need to be efficient because acquiring feedback from humans can be expensive. Thirdly, because of human intelligence, richer forms of information acquisition in different application scenarios can be adopted to improve the interaction efficiency. As an example of such attempts, recently my collaborators and I developed a reinforcement learning based conversational recommendation system Estimation-Action-Reflection [1], which is able to augment traditional recommendation systems with a conversational component. Despite such promising attempts and growing interest, there is far more to be done especially in terms of **interaction efficiency, systematic bias and robustness**, when humans are in the loop. I will continue my effort in such endeavors to develop interactive online learning agents and systems that can better serve human needs.

Human-out-of-the-loop trial and error. Design choices are pervasive in both scientific and industrial endeavors: for example, engineers and scientists design machines or programs to execute tasks more efficiently, and pharmaceutical researchers design new drugs to fight disease. Such design choices inevitably involve extensive human experts' trial and error to find the right ones. Any significant advances in automating such processes can result in immediate productivity improvement and innovation in a wide area of domains. Another direction of my long-term research goals is to use interactive online learning techniques to move humans out of those tedious trial and error processes. Such a long-term goal is not an unachievable tale, and promising progress has been made in areas such as automated machine learning (AutoML), and autonomous systems using machine learning and reinforcement learning techniques. However, such innovations are not mature enough to be widely adopted in real-world scenarios. A lot of challenges are yet to be addressed. For example, achieving this goal requires the online learning agent to be **efficient, robust, and accountable**. As an important attempt toward this long-term goal, my collaborator and I developed a fast and light-weight hyperparameter optimization method and an AutoML solution FLAML [3], which showed superior empirical performance over state-of-the-art hyperparameter optimization baselines and AutoML libraries. In the future, I will continue my collaboration with researchers from different areas to develop intelligent agents that can enhance human productivity.

2.2 Short-term plans

To achieve my long-term goals, a series of research questions need to be answered, base on which I developed a plan for my future research in the incoming 3-6 years.

From sample-efficient to interaction-efficient interactive online learning. Deep reinforcement learning has achieved great success in recent years, notably in playing video games and strategic board games. However, training those agents, especially model-free ones, usually requires

a huge amount of samples. This directly limits its application in many important real-world problems, especially the ones where humans are involved. One of my future research plans is to improve the sample efficiency of online learning agents from the perspective of meta-learning and experiences sharing. Another important research plan is to extend the concept of sample efficiency to **interaction efficiency**. Interaction efficiency can be a fundamentally more suitable optimization target in many application scenarios, where richer forms of interactions between the learning agent and the environment are feasible.

Detection of model misspecifications during online learning. Model-based approaches are generally more sample-efficient than model-free ones, but they are prone to model misspecifications. Model misspecification directly leads to serious systematic bias, such as representational bias and learning bias. Unfortunately, model misspecification is largely ignored in model-based approaches, which greatly hinders the application of such solutions to real-world scenarios. I plan to develop principled solutions to address this limitation and make model-based online learning approaches more robust. My insight is that statistical hypothesis tests, especially online hypothesis tests, are needed as an indispensable component in most model-based online learning approaches. My expertise in online hypothesis testswill help me develop principled solutions to address this important challenge.

Robust, accountable, private and fair interactive online learning. **Robustness.** In many real-world scenarios, the environment with which the learning agent is interacting is not always benign. Various attacks, such as data poisoning attack and extract attack, exist in many real-world situations. Thus it is necessary to consider the online learning process in an adversarial setting. In my future research, I plan to harden the interactive online learning solutions for adversarial robustness and study the theoretical effects of such solutions. **Accountability.** The performance of an interactive online learning algorithm, such as multi-armed bandit and reinforcement learning, can vary drastically during online learning, for example because of exploration. Conventional online learning algorithms provide little information about the quality of their current policies before deployment, which directly limits their usage in high-stake applications like healthcare. To conquer this severe limitation, I plan to develop accountable online learning frameworks whose performance can be quantified and guaranteed during online learning. Furthermore, I plan to extend theoretical understandings about online learning algorithms to a more holistic view: variance and stability of an online learning algorithm need to be studied in addition to sample-complexity. **Privacy.** Privacy concerns have been raised on online learning algorithms. Real-world privacy breaches have been reported in Amazon and Facebook's recommendation systems, where an adversary extracts private information about a user solely based on the system's recommendation sequence. In my future research, I plan to harden the interactive online learning solutions with privacy protection under different threat models. **Fairness.** Fairness is another important ethical constraint for online learning algorithms, especially when humans are in the loop, and the online decisions have important consequences on people's lives, such as hiring, policing, and even criminal sentencing. In my future work, I would like to study fairness in online decision making from the following perspectives: Firstly, because of the online learning agents' uncertainty about the environment, the definition of 'fairness' needs to be carefully studied, and it need to be defined in a problem-dependent manner in the online learning setting. Secondly, due to the sequential nature of many online learning solutions, such as reinforcement learning, the impact of decisions may be delayed, which requires us to study their delayed impact on fairness. Thirdly, it is necessary to study the trade-off and reconciliation between fairness and utility maximization during online decision making.

3 Summary

I believe the research I am planning to conduct will greatly contribute to both algorithmic deployment and theoretical understanding of interactive online learning in both human-in-the-loop and human-out-of-the-loop scenarios. It not only can contribute to the next generation of artificial intelligence, but also can bring in broader impacts on education and the next generation of the industrial revolution. The resolution of these important issues necessarily requires fundamental academic research and interdisciplinary collaborations, which I would be excited to contribute to. I believe the extensive experience in both theoretical and applied research has prepared me well for this mission.

References

[1] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. Estimation–action–reflection: Towards deep interaction between conversational and recommender systems. In *13th ACM International WSDM Conference, WSDM '20.*

[2] Yi Qi, Qingyun Wu, Hongning Wang, Jie Tang, and Maosong Sun. Bandit learning with implicit feedback. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2018.

[3] Chi Wang and Qingyun Wu. Flo: Fast and lightweight hyperparameter optimization for automl. *arXiv 1911.04706*, arXiv 2019.

[4] Huazheng Wang, Sonwoo Kim, Eric McCord-Snook, Qingyun Wu, and Hongning Wang. Variance reduction in gradient exploration for online learning to rank. In *42nd ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 835–844. ACM, 2019.

[5] Huazheng Wang, Qingyun Wu, and Hongning Wang. Factorization bandits for interactive recommendation. In *The 31th AAAI Conference on Artificial Intelligence, AAAI '17*.

[6] Huazheng Wang, Qingyun Wu, and Hongning Wang. Hidden feature leaning for contextual bandit. In *The 25th ACM International Conference on Information and Knowledge Management, CIKM '16.*

[7] Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, pages 495–504, New York, NY, USA, 2018. ACM.

[8] Qingyun Wu, Zhige Li, Huazheng Wang, Wei Chen, and Hongning Wang. Factorization bandits for online influence maximization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 636–646, New York, NY, USA, 2019. ACM.

[9] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1927–1936, New York, NY, USA, 2017. ACM.

[10] Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538. ACM, 2016.

[11] Qingyun Wu, Huazheng Wang, Yanen Li, and Hongning Wang. Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *The World Wide Web Conference*, WWW '19, pages 2080–2090, New York, NY, USA, 2019. ACM.