Statement of Purpose

Subhojyoti Mukherjee

I want to pursue Ph.D. in Computer Science and I aspire to become a professor in this field. My research interests span the areas of *Reinforcement Learning, Online Optimization, and Recommender systems*. I completed M.S (Research) in Computer Science on July 2018 from **Indian Institute of Technology Madras** where my advisers where *Dr. Balaraman Ravindran* and *Dr. Nandan Sudarsanam*. I am currently a Ph.D. candidate at **University of Massachusetts Amherst** from Fall 2018 but I am looking for Ph.D opportunities in this university as my recruiting faculty *Dr. Akshay Krishnamurthy* has left for Microsoft Research and there are no other faculty at the current institute whose research interest matches with mine.

1 Completed/Ongoing Research

My previous research have mainly focused on the theory aspect of Reinforcement learning (RL). An RL agent employs a sequentially adaptive strategy to select actions based on some environmental feedback that maximizes some long-term performance metric. Within the RL framework, Multi-armed Bandits try to explain one of the fundamental challenges of learning in general called the exploration-exploitation dilemma. The exploration-exploitation trade-off tries to answer the question when should a learning algorithm stop exploring alternative actions and focus on the action that has yielded the maximum payoff till now. This trade-off is a fundamental challenge in many practical and industrial level problems such as recommending items to new users for which no prior information exist; in medical domain trying to administer treatments to patients who are suffering from a disease; in speech and dialogue system in Natural Language Processing, etc. Some of my research works which got published in premiere conferences (or is ongoing) which addresses these issues are mentioned below.

1. Variance aware Bandits: While variance aware bandits have been studied for quite some time now, there are quite a few issues that still needs attention. One of them is to how to use them in pure-exploration setting and how to close some theoretical gaps. In the pure-exploration Thresholding Bandits (Mukherjee et al., 2017) we propose an algorithm which finds the best set of actions all of whose quality is above a particular threshold. This is relatively more complex problem to handle as this best set can be quite large and the candidate set of actions from which to select the best set can be exponentially large. This type of pure exploration problems arises in mobile channel recommendation, online item recommendation where there is a separation between the testing and deploying phase. We propose a novel algorithm to solve this problem and give theoretical bounds on the maximum possible loss suffered by our method. My next work focused on a more fundamental problem regarding exploration-exploitation dilemma. For quite some time there has been a theoretical gap in the lower and upper bound on the loss of a class of strategies called variance aware elimination algorithms. These algorithms uses variance estimates to determine the quality of an action and successively eliminates actions which are deemed to be sub-par based on their estimated qualities. Our proposed algorithm called Efficient-UCB-Variance (Mukherjee et al., 2018) is such a strategy and we established that the maximum possible loss (upper bound) of this strategy matches the lower bound for all class of strategies in all the possible environments satisfying certain properties. These works have been accepted in **IJCAI 2017** and **AAAI 2018** respectively.

2. Changepoint detection and Piecewise i.i.d Bandits: Applying Bandit strategies in environments where the quality of all the actions change abruptly and simultaneously has proved to be quite challenging. This is quite common in *medical domain* where the behavior of a drug-resistant bacteria may change abruptly prompting a different response to all the treatments (actions). We proposed two novel algorithms called UCB-CPD and ImpCPD for this setting with interesting theoretical guarantees and closed some gaps in the existing bounds. This work was jointly done with Dr. Odalric Maillard while I was an intern at INRIA, SequeL Lab,Lille and is currently under review in **AISTATS 2019**.

3. Ranking and Latent Bandits: Online ranking of items personalized for each user based on their historical preference is an active area of research. This problem is challenging because the user-item preference matrix can be very large while it has a low rank structure which can be leveraged. Previous works have mainly focused to approximately reconstruct this user-item matrix with stochastic observations which requires a lot of assumptions and costly matrix

operations. We mainly focused on intelligently exploring a partially observed user-item matrix without reconstructing it by leveraging the prior knowledge of the rank of the matrix. This is an ongoing work with Dr. Branislav Kveton, Dr.Anup Rao while I was an intern at Adobe Research which we hope to publish soon.

2 Future Directions

Although I have mainly focused on immediate Reinforcement Learning settings and mostly theoretical works, I am open to a variety of research areas. A few of the interesting research problems that I intend to work are,

1. Off Policy Reinforcement Learning: Often the complete knowledge of the world is not available to the learner and this forces it to adopt different strategies. Some of the motivating examples of these are, a robot trying to navigate a maze using faulty sensors, or administering medical treatments to patients with only a partial knowledge of the true state of the patient (as it is requires information to an atomic level) or even in recommender systems trying to figure out user preferences from past logs without actually subjecting user to cumbersome exploration strategies. In RL, partial observability is a broad area of research and theoretical guarantees are hard to give because of the high variance in the performance of various strategies due to limited exposure to the state space. This is related to my previous experience as this again boils down to conducting intelligent exploration to reduce variance in performance as well as give a theoretical guarantee on how worse the performance can be.

2. Safe Reinforcement Learning: I am interested in safe RL and it has some of the same challenges posed above. For example, in *medical domain* not only we have to work in a partial observable environment but also give confidence guarantees for the proposed action (medical treatment).This gives confidence to the medical practitioner in a high-risk environment where faulty actions might result in death. This is again directly related to my previous experience as this requires a sound knowledge of concentration inequalities to give the confidence on an action taken. In several areas of robotics safe RL is of extreme importance, for example in autonomous driving, motion planning and control, etc.

3. High Dimensional Reinforcement Learning: There has been several recent successes of RL for large state spaces like the game Go or in autonomous driving. In large state spaces the success of RL has come by employing powerful non-linear function approximation techniques like Neural Networks. But sound theoretical guarantees for RL only exist for small state spaces using linear function approximation techniques. I am interested in bridging this gap between the theoretical and empirical performance of RL as this will directly benefit some of the challenges mentioned before and I can draw my knowledge from concentration measure theory and RL to contribute to this.

3 Program and Faculty of Interest

There are several professors at Carnegie Mellon University such as *Dr. Pradeep Ravikumar* and *Dr. Aaditya Ramdas* whose projects are especially appealing to me. Dr. Pradeep Ravikumar works in theoretical machine learning and its applications to high dimensional real-life data-sets which is an area I am interested to work in. Specifically, I want to work with him in the theoretical areas of learning theory and find common solutions where bandit theory can be applied. Also, I am interested to work with Dr. Aaditya Ramdas in the area of online learning and optimization. I want to collaborate with him in several of his areas of interest such as automated discovery and efficient data processing in applied sciences including health-sciences and robotics, where there are several bandit variants which can be applied.

References

Mukherjee, S.; Kolar Purushothama, N.; Sudarsanam, N.; and Ravindran, B. 2017. Thresholding bandits with augmented ucb. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2515–2521.

Mukherjee, S.; Kolar Purushothama, N.; Sudarsanam, N.; and Ravindran, B. 2018. Efficient-UCBV: An almost optimal algorithm using variance estimates. In *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence (to appear)*.