Research Statement — Khaled Nakhleh

Recent multi-agent reinforcement learning (MARL) algorithms have been utilized to solve certain Markov games that are known to converge to the optimal solution. These Markov games include the special case where agents obtain the same cost or reward (i.e. identical-interest Markov games) [10], and the case where two agents act adversarially (i.e. zero-sum Markov game setting) [5]. However, for the more realistic setting of general-sum Markov games, where agents have different cost or reward functions, finding a solution concept was recently shown to be PPAD complete [1]. This intractability result mainly stems from when agents interact in environments that are dynamic and uncertain, along with having incomplete information about other agents' policies.

One of the key observations that I made in my research thus far is that in order to search for tractable solutions, one must examine a certain class of general-sum Markov games, exploit the latent structure in the considered class, and then develop solvers that run in polynomial time using either general-purpose or high-performance computing platforms. Hence, in order to design and implement tractable MARL algorithms for real-world cases, it is necessary to develop special purpose MARL algorithms that apply only to the considered class of general-sum Markov games. However, despite the promise of specialized MARL algorithms providing tractable solutions, convergence guarantees and finite-time analysis for such algorithms appear as challenges that require further research. Such challenges lead to ask the following question:

How can we design provably efficient multi-agent reinforcement learning algorithms for special classes of general-sum Markov games, and ensure their scalable deployment onto high-performance computing systems?

My research aims to address these challenges by developing novel MARL algorithms within the theoretical framework of Markov games. Specifically, I focus on learning methods that leverage game-theoretic insights to achieve scalable and sample-efficient solutions. My research has led to publications that propose and implement multi-agent reinforcement learning algorithms for restless multi-armed bandits (RMABs), general-sum Kullback-Leibler (KL) controlled Markov games, and identical-interest multi-agent KL controlled systems. I also proposed the first deep reinforcement learning algorithm for learning the Whittle index of RMABs [8]. Below, I summarize my research achievements so far in three main categories.

1. Kullback-Leibler controlled multi-agent reinforcement learning:

In [7], I investigated a new class of multi-agent MDPs called Kullback-Leibler (KL) controlled MDPs, where agents are able to choose a new policy from their policy space given a base policy that prevents prohibitive transitions. By using KL-divergence as a regularizer, I was able to show that the optimal joint policy has a closed-form that only depends on the base policy and the agent's current value function estimate. I have also proposed a new optimistic policy iteration scheme that iteratively updates the value function and joint policy estimates, and proved the scheme's convergence to the optimal value function and an optimal policy.

In [6], I developed a new fictitious play variant for KL-controlled Markov games, and proved, given standard conditions on the learning rates, that the learning dynamics asymptotically converge to a near stationary Nash equilibrium. In each learning stage, each agent maintains beliefs on all other agents, and calculates their best response to their current beliefs. Agents then transition to a new state, and each agent updates their beliefs and value function estimates given the realized new state. Interestingly, since this is a general-sum Markov games' setting, each agent receives a different cost (or reward) value, and the convergence proof follows from the potential-game-like property of the KL-controlled Markov games' class that I proposed. Moreover, agents independently learn their best responses without coordinating with other agents.

2. Index-based multi-agent reinforcement learning:

Finding the optimal policy for RMABs was proven to be PSPACE hard [4] due to the exponential dimensionality increase in the state-action space with more arms. Previous work would only consider special instances of restless bandits, and they would find the Whittle index under such assumptions. In this research project, I circumvented this issue by training each arm's neural network independently of other arms. I then developed two deep multi-agent reinforcement learning algorithms for learning index policies for RMABs which were trained using the high-performance computing clusters at Texas A&M University.

In [8], I presented the first index-based deep reinforcement learning algorithm, NeurWIN, that learns the Whittle index function (and policy) for virtually all RMAB problems. I also designed and implemented the arms as separate training environments, and found that the training converges to the optimal control policy. The arms are required to be differentiable with respect to the neural network parameters. To solve this issue, I modified the framework to include a differentiable activation function (sigmoid function) when calculating the index from the neural network. Since no testing environments existed publicly for Whittle index training and testing, I made my implementation of NeurWIN and the baselines available online.

In [9], I presented a new deep MARL algorithm called DeepTOP: Deep Threshold Optimal Policy. DeepTOP learns a broader set of control policies, called threshold policies, than the Whittle index policy for RMABs. I exploited an important feature of threshold policies called the monotone property of actions. If an agent assigns a value to each state in the state space and compares those values against a certain threshold, then the agent would pick the same binary action for a given state and all states with a larger value. The monotone property led to finding a gradient which is tractable to compute, hence minimizing the number of samples required to learn the optimal threshold policy. The implementation code was made available online as well.

3. Research contributions in wireless networks' optimization:

In [2] and [3], we introduced a new framework of second-order wireless network optimization for new performance metrics, such as the Age-of-Information (AoI) and timely throughput. For clients being served by an access point (AP), second-order optimization entails finding the mean and temporal variance of each client subject to a set of constraints on the wireless channel. The model incorporates the random processes associated with wireless transmissions such as channel qualities and packet deliveries.

In the two publications, I characterized the Gilbert-Elliot channel model that determines when a client is able to receive packets from the AP. We also found a closed-form expression of the channel's mean value over all clients, and the channel's temporal variance. Our framework then solves a linear optimization problem given the channel and clients' constraints. In the simulation results, I have shown that our scheduling policy, Variance-Weighted Deficit (VWD), outperforms all state of the art network scheduling policies. In addition, we validate our framework's theoretical results by showing how the empirical and theoretical AoI values are virtually the same for one client. I have made my implementation publicly available so that other researchers can contribute to this wireless networks' framework.

Future research and plans:

In my postdoctoral research, my short-term goal is to consider other Markov games' classes and obtain new tractable solutions that can be deployed onto HPC and networked systems. This includes addressing the challenges of integrating advanced MARL and ML models with large-scale computing clusters, while ensuring that these models are both computationally efficient and capable of high-speed inference. Due to my experience with network programming at my current research group, particularly through the integration of Software-Defined Networks (SDNs) and Named Data Networking (NDN), I have gained experience in developing custom networking protocols for my proposed schemes. Furthermore, I have gained extensive knowledge of essential tools such as Kubernetes, Docker, and SLURM when using the HPC cluster at Texas A&M University for my research work. The overarching goal of my research is to transition into a full-time research position and continue developing my research agenda. I am confident that through interdisciplinary collaborations and a focus on efficient and scalable solutions, I will be able to advance the state of the art in MARL and also in making impactful contributions to the MARL and HPC research communities.

References

[1] Xiaotie Deng, Ningyuan Li, David Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *National Science Review*, 10(1):nwac256, 2023.

[2] Daojing Guo, **Khaled Nakhleh**, I-Hong Hou, Sastry Kompella, and Clement Kam. A theory of second-order wireless network optimization and its application on aoi. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 999–1008, 2022.

[3] Daojing Guo, **Khaled Nakhleh**, I-Hong Hou, Sastry Kompella, and Clement Kam. Aoi, timely-throughput, and beyond: A theory of second-order wireless network optimization. *IEEE/ACM Transactions on Networking*, pages 1–15, 2024.

[4] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*, pages 318–322, 1994.

[5] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized q-learning in zero-sum markov games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18320–18334. Curran Associates, Inc., 2021.

[6] **Khaled Nakhleh**, Sarper Aydin, Ceyhun Eksin, and Sabit Ekin. Ficitious play in kullback-leibler controlled markov games. *Preprint*, 2025.

[7] **Khaled Nakhleh**, Ceyhun Eksin, and Sabit Ekin. Simulation-based optimistic policy iteration for multi-agent mdps with kullback-leibler control cost. *Submitted to the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025.

[8] **Khaled Nakhleh**, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

[9] **Khaled Nakhleh**, I Hou, et al. Deeptop: Deep threshold-optimal policy for mdps and rmabs. *Advances in Neural Information Processing Systems*, 35:28734–28746, 2022.

[10] Onur Unlu and Muhammed O. Sayin. Episodic logit-q dynamics for efficient learning in stochastic teams. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1985–1990, 2023.