Reinforcement Learning

Current Trends and Future Directions

Book of Abstracts – Day 2 (Feb 27)

8. Nan Jiang, UIUC, Chicago, USA

Title: Rethinking the Theoretical Foundation of Reinforcement Learning

Timings: 27 Feb, 9 – 10 am

Abstract: Given two candidate functions, can we identify which one is the true value function of a large Markov decision process (MDP), given a "benign" dataset? Trivial as it might seem, a version of the question was open for 20+ years in reinforcement learning (RL), and the core difficulties are intimately related to the training instability of modern deep RL. In this talk, I will argue that by rethinking fundamental questions like this, RL theory can provide unique perspectives and solutions to practically relevant problems that are critical to the deployment of RL in real-world scenarios. The first part of the talk concerns holdout validation in offline RL, where the aforementioned question naturally arises. I will show how our algorithm, Batch Value-Function Tournament (BVFT), breaks the theoretical barrier and enjoys promising empirical performances. The second part of the talk is about offline training: when we learn policies from a pre-collected dataset, how to reason about policies that would visit states not seen in the data and avoid over-estimation? I will present the Bellman-consistent pessimism framework, whose extension gives a surprising unification of offline RL and imitation learning.

9. Sridhar Mahadevan, UMass, Adobe, USA

Title: Universal Imitation Games: Generative AI as Coalgebras

Timings: 27 Feb, 10 – 11 am

Abstract: In this talk, we propose a categorical model for generative AI based on coalgebras. We show that many common ML frameworks, including reinforcement learning, can be viewed in terms of coinduction over universal coalgebras. We contrast coinduction with the standard paradigm of inductive inference. We illustrate universal constructions in

category theory, such as pullbacks, pushouts, colimits and limits, and the most general Kan extensions. We give a wide range of examples of coalgebras. We show that RL can be viewed as finding a final coalgebra in a category of coalgebras defined by MDPs. We explain the significance of Lambek's lemma, and the Final Coalgebra Theorem of Aczel and Mendler.

10. Shantala M N, Walmart Research, Bengaluru

Title: Multi-Agent Reinforcement Learning For Supply Chains

Timings: 27 Feb, 11:30 – 12:30 pm

Abstract: A significant portion of industrial applications, particularly those rooted in Operations Research, pose inherent NP-hard challenges and revolve around optimal decision-making. Current methodologies heavily lean on heuristics and numerical optimization techniques. The dynamic nature of these problems necessitates tailored solutions to adapt to frequent changes while maintaining optimal outcomes. In this landscape, reinforcement learning emerges as a promising but burgeoning field, especially apt for NP-hard problems entailing sequential decision-making. Real-world dynamic scenarios are characterized by intricate challenges involving multiple independent entities acting as subsystems of large end to end processes. The theoretical approach of extending Reinforcement Learning (RL) into a multi-agent representation seems fitting for addressing such complexities. We have applied MARL to solve one of the most fundamental and complex problems faced in large supply chains such as that operated by the world's largest retailer. We offer insights into the application of MARL to an Inventory Management system along with the learnings and challenges.

11. N. Hemachandra, IITB, Mumbai

Title: Differential Privacy Algorithms for Decentralised Multi-Agent Reinforcement Learning

Timings: 27 Feb, 2 – 3 pm

Abstract: Privacy of user data is an important requirement. We consider data privacy of the agents' data in the setting of decentralised Multi-agent reinforcement learning under the linear function approximation assumption and propose differential privacy preserving algorithms. Our algorithms achieve a sub-linear regret (in number of episodes). In addition to the standard noise adding mechanisms with unbounded support, we also propose noise injecting mechanisms with bounded noise support, Uniform and Bounded Laplace. Such finite support noise mechanisms capture implementations on machines with finite arithmetic. Our algorithm with

Bounded Laplace noise is as good as the mechanisms with unbounded noise support. We also bring out the trade-off between data privacy and the performance of our decentralised RL algos. Our data privacy preserving algos scale well, super-linearly, with the number of agents. We validate our findings on a well known hardest instance.

12. Sandeep Juneja, TIFR, Mumbai

Title: Best arm identification and average treatment effect in multi-armed bandits – optimal algorithms based on fluid analysis

Timings: 27 Feb, 3 – 4 pm

Abstract: We are given finitely many unknown probability distributions that can be sampled from and our aim is, through sequential sampling, to identify the one with the largest mean. This is a classical problem in statistics, simulation and learning theory. Lately, methods have been proposed that identify a sample complexity lower bound that any algorithm providing probabilistic correctness guarantees must satisfy, and algorithms have been developed that asymptotically match these lower bounds even for general sampling distributions, as the probabilistic error guarantees converge to zero. We review these ideas and propose a novel algorithm that relies on exploiting the underlying fluid structure in the evolution of the optimal sampling process and improves upon existing asymptotically optimal algorithms. We also discuss a related and equally important problem of estimating the difference between the means of the best arm and the competing one, and discuss the associated nuances in the analysis and algorithms.

13. D. Manjunath, IITB, Mumbai

Title: Influencing Bandits: Arm Selection for Preference Shaping

Timings: 27 Feb, 4:30 – 5:30 pm

Abstract: We consider a non stationary multi-armed bandit in which the population preferences are positively and negatively reinforced by the observed rewards. The objective of the algorithm is to shape the population preferences to maximize the fraction of the population favouring a predetermined arm. For the case of binary opinions, two types of opinion dynamics are considered—decreasing elasticity (modeled as a Polya urn with increasing number of balls) and constant elasticity (using the voter model). For the first case, we describe an Explore-then-commit policy and a Thompson sampling policy and analyse the regret for each of these policies. We then show that these algorithms and their analyses carry over

to the constant elasticity case. We also describe a Thompson sampling based algorithm for the case when more than two types of opinions are present. Finally, we discuss the case where presence of multiple recommendation systems gives rise to a trade-off between their popularity and opinion shaping objectives.