Reinforcement Learning

Current Trends and Future Directions

Book of Abstracts – Day 1 (Feb 26)

1. Shivaram Kalyanakrishnan, IITB, Mumbai

Title: On Designing a Winning Agent for Reconnaissance Blind Chess

Timings: 26 Feb, 9 - 10 am

Abstract: Reconnaissance Blind Chess (RBC) is a variant of Chess in which players can only sense a 3 x 3 grid within the 8 x 8 Chess board before they play their move. This restriction makes RBC a game of imperfect information, with many new challenges to address. In this talk, I present Fianchetto, our agent that won the NeurIPS 2021 RBC competition by a large margin, and finished runner-up at NeurIPS 2022. Fianchetto builds on the publicly-available code base of StrangeFish (the 2019 winner), and incorporates changes such as a faster board evaluation module, Bayesian belief updating, and incentives for strategic RBC moves. I discuss how elements of our solution may generalise to other games of imperfect information, and outline topics for future research.

2. Sridhar Mahadevan, UMass, Adobe, USA

Title: Universal Imitation Games: Generative AI Beyond Deep Learning

Timings: 26 Feb, 10 – 11 am

Abstract: In this talk, we propose a categorical framework for generative AI called GAIA that extends beyond deep learning. GAIA is based on learning in simplicial complexes, a construction from higher-order category theory. Unlike compositional deep learning, GAIA is intrinsically hierarchical and builds on many insights from category theory. We illustrate the design of GAIA by first showing how backpropagation can be viewed as a functor, and then show how to generalize backpropagation from compositional graphs to simplicial complexes.

3. Praneeth Netrapalli, Google Research, Bengaluru

Title: Second Order Methods for Bandit Optimization and Control

Timings: 26 Feb, 11:30 – 12:30 pm

Abstract: Bandit convex optimization (BCO) is a general framework for online decision making under uncertainty. While tight regret bounds for general convex losses have been established, existing algorithms achieving these bounds have prohibitive computational costs for high dimensional data. In this talk, we will describe a simple and practical BCO algorithm inspired by the online Newton step algorithm. We show that our algorithm achieves optimal (in terms of horizon) regret bounds for a large class of convex functions that we call κ-convex. This class contains a wide range of practically relevant loss functions including linear, quadratic, and generalized linear models. In addition to optimal regret, this method is the most efficient known algorithm for several well-studied applications including bandit logistic regression. Furthermore, we investigate the adaptation of our second-order bandit algorithm to online convex optimization with memory. We show that for loss functions with a certain affine structure, the extended algorithm attains optimal regret. This leads to an algorithm with optimal regret for bandit LQR/LQG problems under a fully adversarial noise model, thereby resolving an open question posed in (Gradu et al. 2020) and (Sun et al. 2023). Finally, we show that the more general problem of BCO with (non-affine) memory is harder. We derive a $\tilde{\Omega}(T^{2/3})$ regret lower bound, even under the assumption of smooth and quadratic losses.

4. M Vidyasagar, IITH, Hyderabad

Title: Block Asynchronous Stochastic Approximation: Convergence and Rates

Timings: 26 Feb, 2 – 3 pm

Abstract: Stochastic Approximation (SA) is a probabilistic algorithm for solving equations when only noisy measurements are available. In the traditional formulation, every component of the argument is updated at each iteration, which might be called "Synchronous SA (SSA)." In some Reinforcement Learning (RL) applications, only one component of the argument is updated at each time, which might be called "Asynchronous SA (ASA)." In-between lies the topic of this talk, namely "Block ASA (BASA)," in which at each iteration, some but not necessarily all components of the argument are updated. In the talk, convergence results are presented for BASA under quite general conditions, along with estimates of the rates of convergence. These bounds are then applied to block updating in nonconvex optimization, and to simplify some RL algorithms.

5. Volkan Cevher, EPFL, Switzerland

Title: Advancing Infinite Horizon Imitation Learning: Efficiency Guarantees and Assumption-free Exploration

Timings: 26 Feb, 3 - 4 pm

Abstract: In this talk, we describe new algorithms with efficiency guarantees for infinite horizon imitation learning (IL) with linear function approximation.

We begin with the minimax formulation of the problem and then outline how to leverage classical tools from optimization, in particular, the proximal-point method (PPM) and dual smoothing, for online and offline IL, respectively. If a certain exploration assumption is met, we bound the number of MDP interactions needed to compute a policy ϵ-suboptimal compared to the expert. Thanks to PPM, we avoid nested policy evaluation and cost updates for online IL appearing in the prior literature. In particular, we do away with the conventional alternating updates by the optimization of a single convex and smooth objective over both cost and Q-functions. These features allow us to obtain convincing empirical performance for both linear and neural network function approximation.

Finally, we present a second approach which allows us to drop the exploration assumption required for the first result. This new technique is based on an interesting connection between imitation learning and online learning in MDP with time changing rewards.

6. Siddhartha Gadgil, IISc, Bengaluru

Title: AlphaGeometry and friends: AI for Mathematics

Timings: 26 Feb, 4:30 – 5:30 pm

Abstract: Recently researchers at Google developed a system AlphaGeomety that can solve geometry problems from the International Mathematical Olympiad (IMO) at close to Gold Medal level. This was based on algorithmic (i.e., rule based) deduction together with a Language Model ("Generative AI") to generate auxiliary constructions. To train the language model "synthetic data" was generated.

This work follows what are becoming common patterns for the use of AI in mathematics, in particular using Generative AI to obtain useful candidates paired with Deductive Systems, including Interactive Theorem Provers (ITPs), to check correctness, complete proofs, evaluate results etc. Essentially, Generative AI is used for "intuitive" aspects of reasoning and Algorithms/Symbolic AI/ITPs are used for the "logical" aspects of reasoning.

In this talk I will begin with discussing AlphaGeometry, including general lessons. I will then discuss a few other systems for AI for mathematics, including "FunSearch" which proved a result giving an improved bound for the so-called CapSet problem. I will also discuss the design of possible systems for going beyond the present systems, and discuss experiments with GPT-4 showing its powers and its limitations relevant to this quest.

No knowledge of AI or Machine learning will be assumed.

7. Avishek Ghosh, IITB, Mumbai

Title: Competing Bandits in Non-Stationary Matching Markets

Timings: 26 Feb, 5:30 – 6:30 pm

Abstract: Understanding complex dynamics of two-sided online matching markets, where the demand-side agents compete to match with the supply-side (arms), has recently received substantial interest. To that end, in this paper, we introduce the framework of decentralized two-sided matching market under non stationary (dynamic) environments. We adhere to the serial dictatorship setting, where the demand-side agents have unknown and different preferences over the supply-side (arms), but the arms have fixed and known preference over the agents. We propose and analyze an asynchronous and decentralized learning algorithm, namely Non-Stationary Competing Bandits (NSCB), where the agents play (restrictive) successive elimination type learning algorithms to learn their preference over the arms. The complexity in understanding such a system stems from the fact that the competing bandits choose their actions in an asynchronous fashion, and the lower ranked agents only get to learn from a set of arms, not dominated by the higher ranked agents, which leads to forced exploration. With carefully defined complexity parameters, we characterize this forced exploration and obtain sub-linear (logarithmic) regret of NSCB. Furthermore, we validate our theoretical findings via experiments.