Reinforcement Learning for Stochastic Networks, Toulouse

Monday, June 17, 2024 - Friday, June 21, 2024

ENSEEIHT

Program

Table of contents

Monday, June 17, 2024

Coffee - B006 – B007 (10:00 AM - 10:45 AM)

Opening session - Amphi B00 (10:45 AM - 11:00 AM)

Keynote: R. Srikant (University of Illinois Urbana-Champaign) - Amphi B00 (11:00 AM - 12:00 PM)

Learning and Control in Countable State Spaces (11:00 AM)

Abstract: We will consider policy optimization methods in reinforcement learning where the state space is countably infinite. The motivation arises from control problems in communication networks and matching markets. We consider an algorithm called Natural Policy Gradient (NPG), which is a popular algorithm for finite state spaces, and show three results in the context of countable state spaces: (i) in the case where perfect policy evaluation is possible, we show that standard NPG converges with a small modification; (ii) if the error is policy evaluation is within a factor of the true value function, we show that one can obtain bounds on the performance of the NPG algorithms; and (iii) we will discuss the ability of neural network-based function approximations to satisfy the condition in (ii) above.

Lunch - C101-C103 (12:00 PM - 1:30 PM)

Parallel session: Challenges and progress in statistical reinforcement learning - A002 (1:30 PM - 3:00 PM)

Organizer and chair: Odalric-Ambrym Maillard

Exploiting Structure in Undiscounted Reinforcement Learning in Markov Decision Processes (1:30 PM)

Presenter: ORTNER, Ronald (MontanUniversitat Leoben)

This talk considers reinforcement learning in Markov decision processes (MDPs) under the undiscounted reward criterion. In this setting the so-called regret is a natural performance measure that compares the accumulated reward of the learner to that of an optimal policy. Usually the regret depends on the size (number of states and actions) of the underlying MDP as well as its transition structure. We will examine structures of the underlying MDP that allow to give improved bounds on the regret.

Exploration in reward machines with near-optimal regret (2:00 PM)

Presenter: TALEBI, Mohammad Sadegh (University of Copenhagen)

We study reinforcement learning for decision processes with Markovian dynamics but non-Markovian rewards, in which high-level knowledge in the form of a finite-state automaton is available to the learner. Such an automaton, often called Reward Machine (RM) (Toro Icarte et al., 2018), generates rewards based on its internal state as well as events that are detected at various states in the environment. The associated decision processes is called an MDPRM, and we focus on average-reward MDPRMs in the regret setting. For a given MDPRM, there is an equivalent cross-product MDP, to which one can apply provably efficient off-the-shelf algorithms obliviously to the structure induced by the MDPRM. However, this would lead to a large regret in view of the large state-space of the cross-product MDP. We establish a first regret lower bound for MDPRMs and present a model-based algorithm that efficiently exploits the structure in MDPRMs, and analyze its regret non-asymptotically. Like the lower bound, our bound is independent of Q, the number of RM states. Further, it improves over regret bound of the existing baselines (e.g., UCRL2 (Jaksch et al., 2010) applied to the cross-product MDP) by up to a factor of Q^{3/2}. Our regret bound makes appear a notion of diameter in MDPRMs, where we show that it can be smaller by a factor of Q than conventional diameter thereof. Finally, we report numerical experiments that demonstrate the superiority of the proposed algorithm over existing baselines in practice.

Provably Efficient Offline Reinforcement Learning in Regular Decision Processes (2:30 PM)

Presenter: JONSSON, Anders (University Pompeu Fabra)

We study the problem of offline (or batch) Reinforcement Learning (RL) in episodic Regular Decision Processes (RDPs). RDPs are the subclass of Non-Markov Decision Processes where the dependency on the history of past events can be captured by a finite state automaton. We consider a setting where the automaton that underlies the RDP is unknown, and a learner strives to learn a near-optimal policy using pre-collected data, in the form of non-Markov sequences of observations, without further exploration. We present RegORL, an algorithm that suitably combines automata learning techniques and state-of-the-art algorithms for offline RL in MDPs. RegORL has a modular design allowing one to use any off-the-shelf offline RL algorithm in MDPs. We report a non-asymptotic high-probability sample complexity bound for RegORL to yield an ε-optimal policy, which makes appear a notion of concentrability relevant for RDPs. Furthermore, we present a sample complexity lower bound for offline RL in RDPs. To our best knowledge, this is the first work presenting a provably efficient algorithm for offline learning in RDPs.

Parallel session: Policy gradient methods: optimization and convergence - A001 (1:30 PM - 3:00 PM)

Organizer and chair: Isaac Grosof

Computing the bias of stochastic approximation with constant step-size via Stein's method. (1:30 PM)

Presenter: GAST, Nicolas (Inria, Univ. Grenoble Alpes)

Stochastic approximation algorithms are quite popular in reinforcement learning notably because they are powerful tools to study the convergence of algorithms based on stochastic gradient descent (like Q-learning of policy gradient). In this talk, I will focus on constant step-size stochastic approximation and present tools to compute its asymptotic bias, which is non-zero (both for Martingale noise or Markovian noise), contrary to the case of decreasing step-size. The analysis is based on a fine comparison of the generators of the stochastic system and its deterministic counterpart. It is similar to Stein's method.

Convergence for Natural Policy Gradient on Infinite-State Average-Reward Markov Decision Processes (2:00 PM)

Presenter: GROSOF, Isaac (University of Illinois, Urbana-Champaign; Northwestern University)

Infinite-state Markov Decision Processes (MDPs) are essential in modeling and optimizing a wide variety of engineering problems. In the reinforcement learning (RL) context, a variety of algorithms have been developed to learn and optimize these MDPs. At the heart of many popular policy-gradient based learning algorithms, such as natural actor-critic, TRPO, and PPO, lies the Natural Policy Gradient (NPG) algorithm. Convergence results for these RL algorithms rest on convergence results for the NPG algorithm. However, all existing results on the convergence of the NPG algorithm are limited to finite-state settings. We prove the first convergence rate bound for the NPG algorithm for infinite-state average-reward MDPs, proving a $O(1/\sqrt{T})$ convergence rate, if the NPG algorithm is initialized with a good initial policy. Moreover, we show that in the context of a large class of queueing MDPs, the MaxWeight policy suffices to satisfy our initial-policy requirement and achieve a $O(1/\sqrt{T})$ convergence rate. Key to our result are state-dependent bounds on the relative value function achieved by the iterate policies of the NPG algorithm.

On the Global Convergence of Policy Based Methods in Average Reward Problems (2:30 PM)

Presenter: MURTHY, Yashaswini (University of Illinois Urbana Champaign)

In the context of average reward Markov Decision Processes (MDPs), traditional approaches for obtaining performance bounds based on discounted reward formulations fail to provide meaningful bounds due to their dependence on the horizon. This limitation arises because average reward problems can be viewed as discounted reward problems, with the discount factor approaching 1, effectively extending the horizon to infinity. Consequently, theoretical convergence guarantees in the discounted reward framework scale unfavorably with the horizon length, yielding unbounded performance estimates. Therefore, obtaining meaningful convergence bounds for widely employed algorithms in the context of average reward MDPs has been an open problem. In this study, we progress on two classes of algorithms tailored to the average reward objective. First, we examine policy-based reinforcement learning (RL) algorithms, which can be viewed as instances of approximate policy iteration (API). We provide finite time performance bounds of API and show that the asymptotic error goes to zero in the limit as policy evaluation and policy improvement errors tend to zero. We further cast several RL algorithms in the API framework to obtain their overall performance bounds. Second, we study the global convergence analysis of policy gradient algorithms in tabular ergodic average reward MDPs. We obtain a sublinear rate of convergence of the iterates to the globally optimal policy. Unlike discounted reward problems, where the discount factor acts as a source of contraction aiding convergence analysis, average reward problems lack this property. To tackle these challenges, we employ new methods of analysis to prove the global convergence of both classes of algorithms.

These findings shed light on the convergence behavior of policy-based RL algorithms and pave the way for their practical application in average reward scenarios.

Coffee break - B006-B007 (3:00 PM - 3:30 PM)

Parallel session: Online learning in stochastic networks - A002 (3:30 PM - 5:00 PM)

Organizers and chairs: Lei Ying and Weina Wang

Online Learning and Optimization for Queues with Unknown Demand Curve and Service Distribution (3:30 PM)

Presenter: CHEN, Xinyun (The Chinese University of Hong Kong, Shenzhen)

We investigate an online learning and optimization problem in a queueing system having unknown arrival rates and service-time distribution. The service provider's objective is to seek the optimal service fee p and service capacity μ so as to maximize the cumulative expected profit (the service revenue minus the capacity cost and delay penalty). We develop an online learning algorithm is able to effectively evolves the service provider's decisions utilizing real-time data including customers' arrival and service times, without needing the information of the arrival rate or service-time distribution. Effectiveness of the online learning algorithm is substantiated by (i) theoretical results including the algorithm convergence and analysis of the regret, i.e. the cost to pay over time for the algorithm to learn the optimal policy, and (ii) engineering confirmation via simulation experiments of a variety of representative examples. This is joint work with Yunan Liu from NCSU and Guiyu Hong from CUHK-Shenzhen.

Recent Advances in Average-Reward Restless Bandits (4:00 PM)

Presenter: WANG, Weina (Carnegie Mellon University)

We consider the infinite-horizon, average reward restless bandit problem. For this problem, a central challenge is to find asymptotically optimal policies in a computationally efficient manner in the regime where the number of arms, N, grows large. Existing policies, including the renowned Whittle index policy, all rely on a uniform global attractor property (UGAP) assumption to achieve asymptotic optimality, which is a complex and difficult-to-verify assumption. In this talk, I will present new, sampling-based policy designs for restless bandits. One of our proposed policies breaks the long-standing UGAP assumption for the first time; then our subsequent policies eliminate the need for the UGAP assumption to achieve asymptotic optimality entirely. Our techniques offer new insights into guaranteeing convergence (avoiding undesirable attractors or cycles) in large stochastic systems.

Scalable Learning in Weakly Coupled Markov Decision Processes (4:30 PM)

Presenter: YAN, Chen

We explore a general reinforcement learning framework within a Markov decision process (MDP) consisting of a large number N of independent sub-MDPs, linked by global constraints. In the non-learning scenario, when the model meets a specific non-degenerate condition, efficient algorithms (i.e., polynomial in N) exist, achieving a performance gap smaller than $\sqrt{N}$ relative to the linear program upper bound. Analyzing the learning scenario in relation to this upper bound forms the central topic of this work.

Parallel session: Reinforcement learning in continuous time - A001 (3:30 PM - 5:00 PM)

Organizer and chair: Harsha Honnappa

Optimized Decision Making via Active Learning of Stochastic Hamiltonians (3:30 PM)

Presenter: BAJAJ, Chandrajit (UT Austin)

A Hamiltonian represents the energy of a dynamical system in phase space with coordinates of position and momentum. The Hamilton's equations of motion are obtainable as coupled symplectic differential equations. In this talk I shall show how optimized decision making (action sequences) can be obtained via a reinforcement learning problem wherein the agent interacts with the unknown environment to simultaneously learn a Hamiltonian surrogate and the optimal action sequences using Hamilton dynamics, by invoking the Pontryagin Maximum Principle. We use optimal control theory to define an optimal control gradient flow, which guides the reinforcement learning process of the agent to progressively optimize the Hamiltonian while simultaneously converging to the optimal action sequence. Extensions to stochastic Hamiltonians leading to stochastic action sequences and the free-energy principle shall also be discussed. This is joint work with Harsha Honnappa Taemin Heo, Minh Nguyen

Simulation Optimization of Reflected Diffusion Processes (4:00 PM)

Presenter: HONNAPPA, Harsha (Purdue University)

Recent work by Ata, Harrison and Si (2023) introduced a simulation-based computational method for stochastic optimal drift control of multidimensional reflected Brownian motion (RBM). The main objective of their work is to compute an optimal “closed loop” stationary Markov control policy. In this talk, I will present our recent results on computing optimal “open loop” controls for finite horizon control of reflected diffusion processes. Our methodology is also simulation-based, but the unique structure of open loop control problems allows us to pose the control problem as a stochastic optimization problem with an infinite dimensional feasible set. Our main results are rates of convergence and consistency results for the estimated control function. Time permitting, I will also discuss connections and implications of our methodology for training neural ordinary differential equation (NODE) models in deep learning. This is joint work with Zihe Zhou and Raghu Pasupathy at Purdue University.

Symphony of experts: orchestration with adversarial insights in reinforcement learning (4:30 PM)

Presenter: MIGNACCO, Chiara (Université Paris-Saclay)

Structured reinforcement learning leverages policies with advantageous properties to reach better performance, particularly in scenarios where exploration poses challenges. We explore this field through the concept of orchestration, where a (small) set of expert policies guides decision-making; the modeling thereof constitutes our first contribution. We then establish value-functions regret bounds for orchestration in the tabular setting by transferring regret-bound results from adversarial settings. We generalize and extend the analysis of natural policy gradient in Agarwal et al. [2021, Section 5.3] to arbitrary adversarial aggregation strategies. We also extend it to the case of estimated advantage functions, providing insights into sample complexity both in expectation and high probability. A key point of our approach lies in its arguably more transparent proofs compared to existing methods. Finally, we provide simulations for a stochastic matching toy model.

Tuesday, June 18, 2024

Keynote: Éric Moulines (École Polytechnique) - Amphi B00 (9:30 AM - 10:30 AM)

Finite Sample analysis of linear stochastic approximation and TD learning (9:30 AM)

Abstract: In this talk, we consider the problem of obtaining sharp bounds for linear stochastic approximation. We then apply these results to temporal difference (TD) methods with linear functional approximation for policy evaluation in discounted Markov decision processes. We show that a simple algorithm with a universal and instance-independent step size together with Polyak-Ruppert tail averaging is sufficient to obtain near-optimal variance and bias terms. We also provide the respective sample complexity bounds. Our proof technique is based on refined error bounds for linear stochastic approximation together with the novel stability result for the product of random matrices that arise from the TD-type recurrence. We will also discuss how these results extend to the distributed / federated setting.

Coffee break - B006-B007 (10:30 AM - 11:00 AM)

Keynote: Christina Lee Yu (Cornell University) - Amphi B00 (11:00 AM - 12:00 PM)

Exploiting Structure In Reinforcement Learning (11:00 AM)

Abstract: While reinforcement learning has achieved impressive success in applications such as game-playing and robotics, there is work yet to be done to make RL truly practical for optimizing policies for real-world systems. In particular, many systems exhibit clear structure that RL algorithms currently don't know how to exploit efficiently. As a result, domain-specific heuristics often dominate RL both in system performance and resource consumption. In this talk we will discuss types of structures that may arise in real-world systems, and approaches to incorporate such structure in the design of RL algorithms. Examples of structured MDPs include models that exhibit linearity with respect to a low dimensional representation, models that exhibit smoothness in the parameters or the trajectories, and models whose dynamics can be decomposed into exogenous versus endogenous components.

Lunch - C101-C103 (12:00 PM - 1:30 PM)

Parallel session: Reinforcement learning for combinatorial problems - A002 (1:30 PM - 3:00 PM)

Session chair: Daniel Mastropietro

The Traveling Salesman Problem: Novel Approaches Grounded in Evolutionary Reinforcement Learning (1:30 PM)

Presenter: LAYEB, SAFA BHAR (LR-OASIS National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia)

Deep Reinforcement Learning (DRL) has showcased remarkable achievements across various domains, such as image recognition and automation. Nevertheless, its potential in the realm of logistics and transportation, particularly in tackling routing challenges, remains mostly untapped. On the contrary, Evolutionary Algorithms (EA) have enjoyed widespread adoption for solving combinatorial optimization problems. Surprisingly, the amalgamation of EA and DRL techniques for addressing combinatorial optimization problems has received limited attention within existing literature. Motivated by these research gaps, this study introduces an innovative approach known as Evolutionary Reinforcement Learning (ERL) to tackle the Traveling Salesman Problem (TSP). To enhance the policy generated by a deep neural network, we leverage the synergy between the EA and DRL frameworks. Notably, the weights associated with the actor component play a crucial role, especially in non-policy methods. Harnessing the power of EA, we create a population of weights and seamlessly incorporate them into the DRL framework, with the overarching objective of significantly enhancing TSP outcomes. We employed the Genetic Algorithm (GA) as the EA and introduced a novel ERL-based approach, namely the ERL-GA. The conducted computational experiments have demonstrated that the ERL-GA surpasses the basic DRL framework in terms of performance.

Scalable Policies for the Dynamic Traveling Multi-Maintainer Problem with Alerts (2:00 PM)

Presenter: VERLEIJSDONK, Peter (Eindhoven University of Technology)

Downtime of industrial assets such as wind turbines and medical imaging devices is costly. To avoid such downtime costs, companies seek to initiate maintenance just before failure, which is challenging because: (i) Asset failures are notoriously difficult to predict, even in the presence of real-time monitoring devices that signal degradation; and (ii) Limited resources are available to serve a network of geographically dispersed assets. In this work, we study the dynamic traveling multi-maintainer problem with alerts (K-DTMPA) under perfect condition information with the objective to devise scalable solution approaches to maintain large networks with K maintenance engineers. Since such large-scale K-DTMPA instances are computationally intractable, we propose an iterative deep reinforcement learning (DRL) algorithm optimizing long-term discounted maintenance costs. The efficiency of the DRL approach is vastly improved by a reformulation of the action space (which relies on the Markov structure of the underlying problem) and by choosing a smart, suitable initial solution. The initial solution is created by extending existing heuristics with a dispatching mechanism. These extensions further serve as compelling benchmarks for tailored instances. We demonstrate through extensive numerical experiments that DRL can solve single maintainer instances up to optimality, regardless of the chosen initial solution. Experiments with hospital networks containing up to 35 assets show that the proposed DRL algorithm is scalable. Lastly, the trained policies are shown to be robust against network modifications such as removing an asset or an engineer or yield a suitable initial solution for the DRL approach.

Fleming-Viot particle systems to accelerate optimal policy learning in the presence of costly rare events (2:30 PM)

Presenter: MASTROPIETRO, Daniel (INP Toulouse, CNRS-IRIT)

In this talk we present Fleming-Viot particle systems to increase the efficiency in discovering rare events that have an impact in the learning speed of optimal policies. The approach is used to learn the critic of Actor-Critic policy gradient methods that learn optimal parameters of parameterized policies, giving rise to what we call the FVAC method. We have successfully applied FVAC to two different contexts where it has shown an advantage over a benchmark Monte-Carlo or TD Actor-Critic method: (i) network systems, where the objective is to learn an optimal acceptance policy of incoming jobs with large rejection costs; and (ii) a classical RL environment, where the objective is to find the shortest path to the exit in a labyrinth.

Parallel session: Reinforcement learning in MDPs with large state spaces - A001 (1:30 PM - 3:00 PM)

Organizers and chairs: R. Srikant and Yashaswini Murthy

Joint learning and scheduling in queueing systems (1:30 PM)

Presenter: YING, Lei (University of Michigan, Ann Arbor)

This talk presents our recent results on joint learning and scheduling in queueing systems.

Neural Inventory Control in Networks via Hindsight Differentiable Policy Optimization (2:00 PM)

Presenter: ALVO, Matias (Columbia Business School (Decision, Risk and Operations division))

Inventory management offers unique opportunities for reliably evaluating and applying deep reinforcement learning (DRL). We introduce Hindsight Differentiable Policy Optimization (HDPO), facilitating direct optimization of a policy's hindsight performance using stochastic gradient descent. HDPO leverages two key elements: (i) an ability to backtest any policy's performance on a sample of historical "noise" traces, and (ii) the differentiability of the total cost incurred on any subsample with respect to policy parameters. We assess this approach in four problem classes where we can benchmark performance against the true optimum. Our algorithms consistently achieve near-optimal performance across all these classes, even when dealing with up to 60-dimensional raw state vectors. Moreover, we propose a natural neural network architecture to address problems with weak (or aggregate) coupling constraints between locations in an inventory network. This architecture utilizes weight duplication for "sibling" locations and state summarization. We demonstrate empirically that this design significantly enhances sample efficiency and provide justification through an asymptotic performance guarantee. Lastly, we assess our approach in a setting that incorporates real sales data from a retailer, demonstrating its substantial superiority over predict-then-optimize strategies.

A Doubly Robust Approach to Sparse Reinforcement Learning (2:30 PM)

Presenter: ZEEVI, Assaf (columbia university)

We propose a new regret minimization algorithm for episodic sparse linear Markov decision process (SMDP) where the state-transition distribution is a linear function of observed features. The only previously known algorithm for SMDP requires the knowledge of the sparsity parameter and oracle access to a reference policy. We overcome these limitations by combining the doubly robust method that allows one to use feature vectors of \emph{all} actions with a novel analysis technique that enables the algorithm to use data from all periods in all episodes. This algorithm is shown to achieve best possible regret (up to log factors)

Poster session - Hall of building C (3:00 PM - 4:00 PM)

Dynamic Scheduling and Trajectory Planning for Urban Intersections with Heterogeneous Autonomous Traffic (3:00 PM)

Presenter: JOSHI, Purva (TU/e)

The anticipated launch of fully autonomous vehicles presents an opportunity to develop and implement novel traffic management systems, such as for urban intersections. Platoon-forming algorithms, in which vehicles are grouped together with short inter-vehicular distances just before arriving at an intersection at high speed, seem promising from a capacity-improving standpoint. In this work, we present a performance evaluation framework which not only captures the intersection access dynamics via a queueing model (or, more specifically, a polling model) with multiple customer types, but also explicitly accounts for the vehicle trajectory planning via a joint optimisation procedure. We further focus on deriving computationally fast and interpretable closed-form expressions for safe and efficient vehicle trajectories during the process of platoon formation, and show that these closed-form trajectories are equivalent to those obtained via the joint optimisation procedure. Additionally, we conduct a numerical study to obtain approximations for the capacity of an intersection under the platoon-forming framework.

Reinforcement learning and regret bounds for admission control (3:00 PM)

Presenter: WEBER, Lucas (Inria)

The expected regret of any reinforcement learning algorithm is lower bounded by $\Omega\left(\sqrt{DXAT}\right)$ for undiscounted returns, where $D$ is the diameter of the Markov decision process, $X$ the size of the state space, $A$ the size of the action space and $T$ the number of time steps. However, this lower bound is general. A smaller regret can be obtained by taking into account some specific knowledge of the problem structure. In this article, we consider an admission control problem to an $M/M/c/S$ queue with $m$ job classes and class-dependent rewards and holding costs. Queuing systems often have a diameter that is at least exponential in the buffer size $S$, making the previous lower bound prohibitive for any practical use. We propose an algorithm inspired by UCRL2, and use the structure of the problem to upper bound the expected total regret by $O(S\log T + \sqrt{mT \log T})$ in the finite server case. In the infinite server case, we prove that the dependence of the regret on $S$ disappears.

Time-Constrained Robust MDPs (3:00 PM)

Presenter: ZOUITINE, Adil (SUPAERO)

Robust reinforcement learning is essential for deploying reinforcement learning algorithms in real-world scenarios where environmental uncertainty predominates. Traditional robust reinforcement learning often depends on rectangularity assumptions, where adverse probability measures of outcome states are assumed to be independent across different states and actions. This assumption, rarely fulfilled in practice, leads to overly conservative policies. To address this problem, we introduce a new time-constrained robust MDP (TC-RMDP) formulation that considers multifactorial, correlated, and time-dependent disturbances, thus more accurately reflecting real-world dynamics. This formulation goes beyond the conventional rectangularity paradigm, offering new perspectives and expanding the analytical framework for robust RL. We propose three distinct algorithms, each using varying levels of environmental information, and evaluate them extensively on continuous control benchmarks. Our results demonstrate that these algorithms yield an efficient tradeoff between performance and robustness, outperforming traditional deep robust RL methods in time-constrained environments while preserving robustness in classical benchmarks. This study revisits the prevailing assumptions in robust RL and opens new avenues for developing more practical and realistic RL applications.

Learning payoffs while routing in skill-based queues (3:00 PM)

Presenter: VAN KEMPEN, Sanne (TU/e)

We consider skill-based routing in queueing systems with heterogeneous customers and servers, where the quality of service is measured by customer-server dependent random rewards and the reward structure is a priori unknown to the system operator. We analyze routing policies that simultaneously learn the system pa- rameters and optimize the reward accumulation, while satisfying queueing stability constraints. To this end, we introduce a model that integrates queueing dynamics and decision making. We use learning techniques from the multi-armed bandit (MAB) framework to propose a definition of regret against a suitable oracle reward and formulate an instance-dependent asymptotic regret lower bound. Since our lower bound is of the same order as results in the classical MAB setting, an asymptotically optimal learning algorithm must exploit the structure of the queueing system to learn as efficiently as in the classical setting, where decisions are not constrained by state space dynamics. We discuss approaches to overcome this by leveraging the analysis of the transient behavior of the queueing system.

Boosting Rare Event Simulation in Markov Processes (3:00 PM)

Presenter: GARCIA, Ernesto (LAAS)

Under constraints on the total simulation time available for a Markov process, we look for regimes where parallel independent simulations can effectively sample unlikely regions of the state space.

Non-preemptive scheduling with non-observable environment (3:00 PM)

Presenter: HIRA, Thomas (IRIT)

We investigate a non-preemptive scheduling problem within a class of non-observable environments, framed as a restless multi-armed bandit (RMAB) problem characterized by a Markovian dynamics and partial observability. Each arms of this RMAB is modeled as independent Gilbert-Elliot channels with different parameters and the current state of each arms is not observable by the decision-maker so we relied on a belief state for our analysis. The goal is to derive optimal policies that maximize long-run average rewards given the constraints of limited information and non-preemptive service. Our approach involves computing the generative function of the remaining service time based on the current belief state, then using the expected remaining service time to define an index over the different arms. We demonstrate that this index is optimal in the positively auto-correlated case. Additionally, we compare our results to models with observability or preemptive scheduling to quantify the performance loss.

Parallel session: Multi-agent systems - A002 (4:00 PM - 5:30 PM)

Session chair: Konstantin Avrachenkov

Backlogged Bandits for Network Utility Maximization (4:00 PM)

Presenter: STEIGER, Juaren (Queen's University)

We consider network utility maximization for job admission, routing, and scheduling in a queueing network with unknown job utilities as a type of multi-armed bandit problem. This "Backlogged Bandit" problem is a bandit learning problem with delayed feedback due to the end-to-end delay of a job waiting in the queue of each node in its path through the network. While recent work has explored techniques for learning under delayed feedback in this problem, such as the parallel instance technique, we find that the celebrated drift-plus-penalty technique classically used in the optimization of queueing networks already adequately controls the feedback delay in some problem instances. To that end, we focus our attention on developing theoretical techniques to study this style of algorithm in the Backlogged Bandits framework. In this talk, we present our recent work that explores the special case of routing in a bipartite (single-hop) queueing network, and discuss the challenges and our progress toward the general multi-hop case.

Interpersonal trust: An asymptotic analysis of a stochastic coordination game with multi-agent learning (4:30 PM)

Presenter: MEYLAHN, Benedikt (Korteweg-de Vries Institute for Mathematics, University of Amsterdam)

We study the interpersonal trust of a population of agents, asking whether chance may decide if a population ends up in a high trust or low trust state. We model this by a discrete time, random matching stochastic coordination game. Agents are endowed with an exponential smoothing learning rule about the behaviour of their neighbours. We find that, with probability one in the long run the whole population either always cooperates or always defects. By simulation we study the impact of the distributions of the payoffs in the game and of the exponential smoothing learning (memory of the agents). We find, that as the agent memory increases or as the size of the population increases, the actual dynamics start to resemble the expectation of the process. We conclude that it is indeed possible that different populations may converge upon high or low trust between its citizens simply by chance, though the game parameters (context of the society) may be quite telling.

Multiagent Q-learning with `satisficing' criteria (5:00 PM)

Presenter: BORKAR, Vivek (Indian Institute of Technology Bombay)

We consider multiagent Q-learning with each agent having her own reward function, but all agents influencing the transition mechanism. By relaxing the exact optimality to a requirement of `satisficing', modelled as driving the average costs to prescribed acceptable regions, we propose a scheme that provably achieves this.

Parallel session: Reinforcement learning and queueing I - A001 (4:00 PM - 5:30 PM)

Session chair: Nicolas Gast

Stability and performance of multi-class queueing systems with unknown service rates: A combined scheduling-learning approach (4:00 PM)

Presenter: ANTON, Elene (Université de Pau et des Pays de l'Adour (UPPA))

We consider a system with N different service modes handling several traffic classes, where a scheduling agent decides which service option to use at each time slot. Each service mode provides service simultaneously to all the traffic classes, at different random rates (possibly zero). Moreover, each job experiences a random slowdown when in service, which is independent among jobs and service options. The scheduling agent can observe the global queue state, but does not have any advance knowledge of the instantaneous rates, service rate distributions or slowdown rate distributions associated with each of the service modes. We propose a threshold-based algorithm where the threshold value is compared to the sum of the square of the queue lengths at that time slot. If the threshold value is exceeded, then the scheduling agent switches the service mode, and it keeps using the same service mode otherwise. We show that the proposed scheduling algorithm achieves maximum stability and also analyse the mean response time of the various traffic classes. In a second step, we extend the scheduling agent in order to learn which service mode to use at each queue state.

Dynamic Scheduling of a Multiclass Queue in the Halfin-Whitt Regime: A Computational Approach for High-Dimensional Problems (4:30 PM)

Presenter: KASIKARALAR, Ebru (University of Chicago Booth School of Business)

We consider a multi-class queueing model of a telephone call center, in which a system manager dynamically allocates available servers to customer calls. Calls can terminate through either service completion or customer abandonment, and the manager strives to minimize the expected total of holding costs plus abandonment costs over a finite horizon. Focusing on the Halfin-Whitt heavy traffic regime, we derive an approximating diffusion control problem, and building on earlier work by Han et al. (2018), develop a simulation-based computational method for the solution of such problems, one that relies heavily on deep neural network technology. Using this computational method, we propose a policy for the original (pre-limit) call center scheduling problem. Finally, the performance of this policy is assessed using test problems based on publicly available call center data. For the test problems considered so far, our policy does as well as the best benchmark we could find. Moreover, our method is computationally feasible at least up to dimension 100, that is, for call centers with 100 or more distinct customer classes.

Designing M/G/1 Scheduling Policies from Job Size Samples (5:00 PM)

Presenter: RAMAKRISHNA, Shefali (Cornell University)

The Gittins policy is known to minimize mean response time in the M/G/1 with unknown job sizes, provided the job size distribution is known. In practice, however, one does not have direct access to the job size distribution. Motivated by this, we consider the problem of designing a scheduling policy based on a finite number of samples from the job size distribution. This is an open problem. A natural idea is a policy one might call *empirical Gittins*, which constructs a policy using the empirical distribution of the samples. But it is unclear how empirical Gittins compares to the true Gittins policy (constructed from the true job size distribution). This talk will present initial results on M/G/1 scheduling with only sample access to the job size distribution. We show that empirical Gittins is a good proxy for true Gittins in the finite support case. Specifically, we show that empirical Gittins is a $(1+\epsilon)$-approximation for mean response time with probability $1-\delta$, provided at least $O\bigl(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\bigr)$ samples. Our results are based on applications of the recently derived "WINE" queueing identity. We will also present initial progress towards the general case, which combine WINE with the "SOAP" analysis of M/G/1 scheduling policies.

Cocktail (6:00 PM - 7:30 PM)

Wednesday, June 19, 2024

Keynote: Adam Wierman (Caltech) - Amphi B00 (9:30 AM - 10:30 AM)

Learning Augmented Algorithms for MDPs (9:30 AM)

Abstract: Making use of modern black-box AI tools such as deep reinforcement learning is potentially transformational for sustainable systems such as data centers, electric vehicles, and the electricity grid. However, such machine-learned algorithms typically do not have formal guarantees on their worst-case performance, stability, or safety. So, while their performance may improve upon traditional approaches in “typical” cases, they may perform arbitrarily worse in scenarios where the training examples are not representative due to, e.g., distribution shift. Thus, a challenging open question emerges: Is it possible to provide guarantees that allow black-box AI tools to be used in safety-critical applications? In this talk, I will provide an overview of an emerging area studying learning-augmented algorithms that seeks to answer this question in the affirmative. I will survey recent results in the area, focusing on online optimization and MDPs, and then describe applications of these results to the design of sustainable data centers and electric vehicle charging.

Coffee break - B006-B007 (10:30 AM - 11:00 AM)

Keynote: Jim Dai (Cornell University) - Amphi B00 (11:00 AM - 12:00 PM)

Inpatient Overflow Management with Proximal Policy Optimization (11:00 AM)

Abstract: Overflow patients to non-primary wards can effectively alleviate congestion in hospitals, while undesired overflow also leads to issues like mismatched service quality. Therefore, we need to trade-off between con- gestion and undesired overflow. This overflow management problem is modeled as a discrete-time Markov Decision Process with large state and action space. To overcome the curse-of-dimensionality, we decompose the action at each time into a sequence of atomic actions and use an actor-critic algorithm, Proximal Policy Optimization (PPO), to update policy. Moreover, we tailor the design of neural network which represents policy to account for the daily periodic pattern of the system flows. Under hospital settings of different scales, the PPO policies consistently outperform some commonly used state-of-art policies significantly.

Lunch - C101-C103 (12:00 PM - 1:30 PM)

Parallel session: Online learning - A002 (1:30 PM - 3:00 PM)

Session chair: Tejas Bodas

Model-Free Robust ϕ-Divergence Reinforcement Learning Using Both Offline and Online Data (1:30 PM)

Presenter: PANAGANTI, Kishan (California Institute of Technology)

The robust ϕ-regularized Markov Decision Process (RRMDP) framework focuses on designing control policies that are robust against parameter uncertainties due to mismatches between the simulator (nominal) model and real-world settings. This work makes \emph{two} important contributions. First, we propose a \textit{model-free} algorithm called \textit{Robust ϕ-regularized fitted Q-iteration} (RPQ) for learning an ϵ-optimal robust policy that uses only the historical data collected by rolling out a behavior policy (with \textit{robust exploratory} requirement) on the nominal model. To the best of our knowledge, we provide the \textit{first} unified analysis for a class of ϕ-divergences achieving robust optimal policies in high-dimensional systems with general function approximation. Second, we introduce the \textit{hybrid robust ϕ-regularized reinforcement learning} framework to learn an optimal robust policy using both historical data and online sampling. Towards this framework, we propose a model-free algorithm called \textit{Hybrid robust Total-variation-regularized Q-iteration} (HyTQ). Finally, we provide theoretical guarantees on the performance of the learned policies of our algorithms on systems with arbitrary large state space using function approximation.

Score-Aware Policy-Gradient Methods and Performance Guarantees using Local Lyapunov Conditions (2:00 PM)

Presenter: SENEN–CERDA, Albert (IRIT, LAAS–CNRS, and Université de Toulouse)

In this talk, we introduce a policy-gradient method for model-based Reinfocement Learning (RL) that exploits a type of stationary distribution commonly obtained from Markov Decision Processes (MDPs) in stochastic networks, queueing systems and statistical mechanics. Specifically, when the stationary distribution of the MDP belongs to an exponential family that is parametrized by policy parameters, we can improve existing policy gradient methods for average-reward RL. Our key identification is a family of gradient estimators, called Score-Aware Gradient Estimators (SAGEs), that in the aforementioned setting, enable policy gradient estimation without relying on value-function approximation. This contrasts with other common policy-gradient algorithms, such as actor-critic methods. We first show that policy-gradient with SAGE locally converges, including in cases when the objective function is nonconvex, presents multiple maximizers, and the state space of the MDP is not finite. Under appropriate assumptions such as starting sufficiently close to a maximizer, the policy under stochastic gradient ascent with SAGE has an overwhelming probability of converging to the associated optimal policy. Other key assumptions are that a local Lyapunov function exists, and a nondegeneracy property of the Hessian of the objective function holds locally around a maximizer. Furthermore, we conduct a numerical comparison between a SAGE-based policy-gradient method and an actor-critic method. We specifically focus on several examples inspired from stochastic networks, queueing systems, and models derived from statistical phyiscs, where parameterizable exponential families are commonplace. Our results demonstrate that a SAGE-based method finds close-to-optimal policies faster than an actor-critic method.

Tabular and Deep Reinforcement learning for Gittins Index (2:30 PM)

Presenter: BODAS, Tejas (IIIT Hyderabad)

In the realm of multi-arm bandit problems, the Gittins index policy is known to be optimal in maximizing the expected total discounted reward obtained from pulling the Markovian arms. In most realistic scenarios however, the Markovian state transition probabilities are unknown and therefore the Gittins indices cannot be computed. One can then resort to reinforcement learning (RL) algorithms that explore the state space to learn these indices while exploiting to maximize the reward collected. In this work, we propose tabular (QGI) and Deep RL (DGN) algorithms for learning the Gittins index that are based on the retirement formulation for the multi-arm bandit problem. When compared with existing RL algorithms that learn the Gittins index, our algorithms have a lower run time, require less storage space (small Q-table size in QGI and smaller replay buffer in DGN), and illustrate better empirical convergence to the Gittins index. This makes our algorithm well suited for problems with large state spaces and is a viable alternative to existing methods. As a key application, we demonstrate the use of our algorithms in minimizing the mean flowtime in a job scheduling problem when jobs are available in batches and have an unknown service time distribution.

Parallel session: Some applications of reinforcement learning to networks - A001 (1:30 PM - 3:00 PM)

Organizer and chair: Vivek Borkar

Learning LP-indices in Average-Reward Restless Multi-Armed Bandits (1:30 PM)

Presenter: AVRACHENKOV, Konstantin (INRIA Sophia Antipolis)

Restless Multi-Armed Bandits (RMABs) are extensively used in scheduling, resource allocation, marketing and clinical trials, just to name a few application areas. RMABs are Markov Decision Processes with two actions (active and passive modes) for each arm and with a constraint on the number of active arms per time slot. Since in general RMABs are PSPACE-complete, several heuristics such as Whittle index and LP index have been proposed. In this talk, I present a reinforcement learning scheme for LP indices with almost sure convergence guarantee in the tabular setting and an empirically efficient Deep Q-learning variant. Several examples, including scheduling in queueing systems, will be presented. This is a joint work V.S. Borkar and P. Shah from IIT Bombay.

Optimal congestion signaling under customer heterogeneity with private types (2:00 PM)

Presenter: JHUNJHUNWALA, Prakirt (Columbia Business School)

In an unobservable queue, where customers lack the complete wait time information, a throughput-maximizing server aims to exploit the information asymmetry by strategically signaling coarse congestion information to incentivize customers' arrival into the system. The customers make a calculated decision about joining the queue by creating a belief of their utility given the congestion signal provided by the server. Using the Bayesian persuasion framework to model the customers' response, we map the problem of designing an optimal signaling mechanism to finding an optimal policy in a Constrained MDP problem. Afterward, we exploit the Constrained MDP formulation to derive the structure of the optimal policy. When customers are heterogeneous, we discover a counter-intuitive phenomenon where the optimal signaling mechanism attains a laminar structure, as opposed to a monotone structure commonly seen in MDP settings. We also show that the laminar structure of the optimal policy is also prevalent in a large class of admission control problems.

Energy Management in Smart Grids Using Finite Horizon SOR Q-learning (2:30 PM)

Presenter: BHATNAGAR, Shalabh (Indian Institute of Science)

The smart grid is comprised of different microgrids and is supported by different technologies. The microgrid has to make a lot of decisions, and this has to be automated for efficiency. We have Markov Decision Processes as a frame- work for sequential decision-making under uncertainty and reinforcement learning techniques to learn the optimal decisions. In this work, we formulate this problem in the setting of finite horizon Markov Decision Processes and propose two algorithms. One is a typical finite horizon algorithm similar to [1], and the other is an improvement of this algorithm using the technique of successive over relaxation. This is then applied to the smart grid problem.

Thursday, June 20, 2024

Keynote: Vincent François-Lavet (VU, Amsterdam) - Amphi B00 (9:30 AM - 10:30 AM)

Generalization in deep reinforcement learning and the role of representation learning for sequential decision-making tasks (9:30 AM)

Abstract: The concept of generalization in the context of deep reinforcement learning will be explained along with the different elements that can be used to improve generalization. The key role of representation learning for improved generalization and interpretability will then be shown. Finally, the role of representation learning for exploration and transfer learning will also be discussed.

Coffee break - B006-B007 (10:30 AM - 11:00 AM)

Keynote: Bruno Gaujal (INRIA) - Amphi B00 (11:00 AM - 12:00 PM)

The sliding regret (11:00 AM)

Optimistic reinforcement learning algorithms in Markov decision processes essentially rely on two ingredients to guarantee regret efficiency. The first one is the choice of well-tuned confidence bounds and the second is the design of a pertinent rule to end episodes. While many efforts have been dedicated to improve the tightness of confidence bounds, the management of episodes has remained essentially unaltered since the introduction of the doubling trick (DT) in UCRL2 (Auer et Al.2009). In this talk, I will present two solutions to move beyond (DT). The first one is the performance test (PT) that ends an episode as soon as the performance of the current policy becomes obviously sub-optimal. The second one is the vanishing multiplicative (VM) rule that is as simple as DT to implement and replace the doubling criterion by a weaker one. Both solutions keep the regret of the algorithm unaltered and induce a drastic reduction of the local regret taken from the start of exploration episodes (start of episodes where a sub-optimal policy is used). More specifically, classical algorithms such as UCRL2, KL-UCRL or UCRL2B, patched with our new rules get an immediate benefit. Their regret upper bound remain the same (up to a small negligible additive term) while their asymptotic local regret at exploration times decreases from Omega(T) to O(log T). I will also comment on numerical experiments confirming our asymptotic findings. The regret of algorithms under (VM) or (PT) becomes slightly better and significantly smoother, while the local regret at exploration times becomes sub-linear, even over finite times.

Lunch - C101-C103 (12:00 PM - 1:30 PM)

Parallel session: Algorithmic collusion: Foundations for understanding the emergence of anticompetitive behaviour - A001 (1:30 PM - 3:00 PM)

Organizer and chair: Janusz Meylahn

Reinforcement learning in a prisoner's dilemma (1:30 PM)

Presenter: DOLGOPOLOV, Artur (Bielefeld University)

I characterize the outcomes of a class of model-free reinforcement learning algorithms, such as stateless Q-learning, in a prisoner's dilemma. The behavior is studied in the limit as players stop experimenting after sufficiently exploring their options. A closed form relationship between the learning rate and game payoffs reveals whether the players will learn to cooperate or defect. The findings have implications for algorithmic collusion and also apply to asymmetric learners with different experimentation rules.

Less than meets the eye: simultaneous experiments as a source of algorithmic seeming collusion (2:00 PM)

Presenter: LAMBIN, Xavier (ESSEC Business School)

This article challenges the idea of algorithmic collusion as proposed in Calvano et al. (2020) and subsequent literature. Identifying a critical mistake, we dispute the notion that supracompetitive prices result from collusive mechanisms where high prices are sustained by reward and punishment strategies. Instead, our analysis suggests that both phenomena originate from simultaneous experimentation and learning inertia inherent in reinforcement learning, without a causal link between them. Such seeming collusion can emerge rapidly in memoryless environments and with myopic agents, cautioning against misinterpreting the phenomena as collusion. Our findings advocate for simpler approaches to address algorithmic supra-competitive pricing issues. **below is an extended abstract:** Algorithmic decision-making has become ubiquitous in our lives, and its impact is increasing at an unprecedented rate. From our social media feeds to the stock market, from self-driving cars to medical diagnoses, algorithms are increasingly being used to automate decision-making processes. In an influential paper Calvano et al. [2020b] (henceforth CCDP), show that basic and independent reinforcement algorithms, when trained simultaneously, consistently achieve supra-competitive outcomes. Furthermore, the responses of the algorithms to out-of-equilibrium stimuli resemble reward-punishment schemes that may be used to sustain collusion. The authors conclude that algorithms genuinely collude and provide policy recommen dations and guidance to antitrust authorities. In particular, Calvano et al. [2020a] present tests, based on responses to stimuli, that regulators can employ to verify whether algorithms are autonomously engaging in collusion.This research has been followed up by many other studies that have used and expanded the notion of algorithmic collusion, such as Hettich [2021], Dolgopolov [2021], Banchio and Skrzypacz [2022], Klein [2021], Werner [2022], Qiu et al. [2022], Xu et al. [2023], to name just a few. Should these findings be confirmed, they could have significant implications for antitrust regulations, necessitating urgent action. It is therefore no surprise that the issue of Artificial Intelligence (AI) collusion has garnered attention from regulatory agencies, with it being a top priority on the agendas of many organizations (see e.g. OECD [2017], Autoridade Da Concorrència [2019], ACB [2019], Ezrachi and Stucke [2018], McSweeny and O'Dea [2017], Competition Bureau [2018] and Petit [2017]). The results of CCDP and subsequent works are, however, increasingly debated. Critics raise possible methodological or design issues (Meylahn et al. [2022], Abada et al. [2022], Eschenbaum et al. [2022]), or question the interpretation of the results (Epivent and Lambin [2024], Abada and Lambin [2023]). Asker et al. [2022, 2023] emphasize the critical role of algorithmic learning protocols on supracompetitive limit prices. Calvano et al. [2023] claim that the “spurious” collusion results of Asker et al. [2023] are driven by the specific exploration mode they implement (synchronous learning), together with optimistic initialization of Q-matrices. Overall, the literature provides no formal explanation for the observational facts described in CCDP. This is mostly due to the fact it is notoriously difficult to draw theoretical results in multi-agent Q-learning processes. This paper employs a simplified exploration procedure to elucidate the dynamics at play. It demonstrates that apparent collusion arises due to the specific learning process inherent in simultaneously-trained reinforcement learning algorithms: by construction, the initial valuations of actions are based on experiments that are performed while the other agents are also experimenting. These valuations may differ significantly from the profits observed in "play" mode when all agents play only their preferred (or "greedy") actions. Still, the learning procedure is such that the erroneous valuations persist over time. When the rate of exploration decreases jointly, we show that agents may fail to identify profitable independent deviations and converge to prices (much) greater than Nash. Our theory is similar in spirit to that of Banchio and Mantegazza (2022), though we address much more general demand systems than the prisoner dilemma, with a specific application to the economic environment of CCDP. Our results are not restricted to cycles that possibly include cooperative actions, but also rationalize the convergence to singleton or fixed-point (supra-competitive) strategies, which represent 64 % of the simulations in CCDP. We use a mean-field assumption, eliminating the need for continuous time approximation of the learning process. Compared to Asker et al. (2022), we provide a complete characterization of the initial and final ``greedy actions'' in the Q-learning context, when endowed with a simple exploration procedure. We offer a comprehensive description of the underlying mechanism, including the characterization of the convergence point of the algorithm of CCDP. A notable contribution to the literature is our explanation for the apparent reward and punishment schemes identified in CCDP and subsequent works, clarifying that these schemes are not the cause of the observed high prices. Our theory is confronted with the results of CCDP, which we replicate faithfully. The first important step is to show that the memoryless version of CCDP also yields supra-competitive prices, which refutes the claim that the high prices are due to ``genuine'' collusion with high prices caused by reward and punishment schemes. In a second step, the results from our theoretical models are shown to explain the main observations of this literature: simultaneous learning causes the convergence to high prices. Finally, the apparent reward and punishment schemes are also shown to be spurious. From these observations, we note that the misinterpretation of high prices and apparent punishment schemes as evidence of genuine collusion has led to misguided policy recommendations. We propose to correct the interpretation and to implement more straightforward policy interventions against supracompetitive prices.

Quantifying the likelihood of collusion by provably convergent reinforcement learning (2:30 PM)

Presenter: MEYLAHN, Janusz (University of Twente)

Recent advances in decentralized multiagent reinforcement learning (MARL) have led to the development of algorithms that are provably convergent in a variety of Markov game subclasses. One of these is the Decentralized Q-learning (DQ) algorithm by Arslan and Yüksel (2017) which is provably convergent in weakly acyclic games. In this talk, I will present a new characterization of weak acyclicity and use it to show that the prisoner's dilemma with a memory of one period is weakly acyclic. This new characterization naturally leads to an identification of the basins of attraction of all possible strategy equilibria of the DQ algorithm. Since only a subset of strategy equilibria leads to robust collusion, we can use this to quantify the likelihood of observing algorithmic collusion. In addition, I will discuss the effect that fluctuations in the learning process and the addition of a third intermediate action to the prisoner's dilemma have on the likelihood of collusion.

Parallel session: Reinforcement learning for energy networks - A002 (1:30 PM - 3:00 PM)

Organizers and chairs: Sandjai Bhulai and Alessandro Zocca

Decentralized MARL and Preference Learning: Balancing Grid Efficiency and Privacy in Demand Response (1:30 PM)

Presenter: CREMER, Jochen (Delft University of Technology)

This talk introduces a decentralized Multi-Agent Reinforcement Learning (MARL) with inverse reinforcement learning for Electricity Demand Response (DR) programs in the residential sector, aiming at grid stability and efficiency while prioritizing user privacy. The approach is for incentive-based DR that addresses the grid's capacity limits and congestion challenges and aligns with the residential consumers' preferences and comfort without reducing their privacy. The approach uses a Disjunctively Constrained Knapsack Problem optimization within a MARL setup to model energy management systems within a household. This optimization is complemented by an inverse reinforcement learning model that generates personalized appliance operation schedules based on historical consumption data, thereby learning user preferences and comfort levels. Studies on the Pecan-street data using 25 households were undertaken. The results demonstrate reductions in the Peak-to-Average ratio (PAR) of energy consumption by 15%.

Multi-Agent Reinforcement Learning for Power Grid Topology Optimization (2:00 PM)

Presenter: VAN DER SAR, Erica (Vrije Universiteit Amsterdam)

Recent challenges in operating power networks arise from increasing energy demands and unpredictable renewable sources like wind and solar. While reinforcement learning (RL) shows promise in managing these networks, through topological actions like bus and line switching, efficiently handling large action spaces as networks grow is crucial. In this talk we present a hierarchical multi-agent reinforcement learning (MARL) framework tailored for these expansive action spaces, leveraging the power grid’s inherent hierarchical nature. Experimental results indicate the MARL framework’s competitive performance with single-agent RL methods. We also compare different RL algorithms for lower-level agents alongside different policies for higher-order agents.

Scalable Grid Topology Reconfiguration using Consensus-Based Multi Agent Reinforcement Learning (2:30 PM)

Presenter: DE MOL, Barbera (University of Groningen/TenneT TSO (NED))

Power network control is a crucial aspect of modern society, as it allows electricity to be a reliable resource for daily living, industry, and transportation. Controlling electricity is a highly complex task that represents a sequential decision-making problem with large state and action spaces. The state space represents a combinatorial explosion of all possible ways the network can be reconfigured through topological remedial actions. As the power network is a real-world infrastructure that is constantly changing and needs to be acted on real-time, solutions to operate this network need to be scalable while still achieving reliable results. Additionally, similar issues occur at different grid scales, and it is beneficial to deploy a similar solution to each of these scales. In order to improve the scalability of solutions in networks with a combinatorial state space while maintaining similar performance, a multi-agent reinforcement learning approach is proposed. Here, individual agents propose actions that are evaluated by a central controller. The introduction of multiple (autonomous) agents that interact within a shared environment reduces the central computational load and improves scalability.

Coffee break - B006-B007 (3:00 PM - 3:30 PM)

Parallel session: Learning and optimization - A002 (3:30 PM - 5:00 PM)

Organizer and chair: R. Srikant and Yashaswini Murthy

Matching Impatient and Heterogeneous Demand and Supply while Learning (3:30 PM)

Presenter: WARD, Amy (The University of Chicago Booth)

We study a two-sided network where heterogeneous demand (customers) and heterogeneous supply (workers) arrive randomly over time to get matched. Customers and workers arrive with a randomly sampled patience time (also known as reneging time in the literature), and are lost if forced to wait longer than that time to be matched. The system dynamics depend on the matching policy, which determines when to match a particular customer class with a particular worker class. The issue in developing a matching policy is that model primitives are unknown. Our objective is to develop a policy that has small regret, where regret is the difference between the cumulative value of matches, minus costs incurred when demand and supply wait, between a proposed policy, that does not have knowledge of model primitives, and a benchmark policy, that does have full knowledge of model primitives. Our benchmark policy is an asympotically optimal policy (on fluid scale, as demand and supply rates grow large). A key challenge is that the benchmark policy depends on the patience time distributions, and may be different for different distributions, even when the mean is the same.

Pseudo-Bayesian Optimization (4:00 PM)

Presenter: CHEN, Haoxian (Columbia University)

Bayesian Optimization aims to optimize expensive black-box functions using minimal function evaluations. Its key idea is to strategically model the unknown function structure via a surrogate model and, importantly, quantify the associated uncertainty that allows a sequential search of query points to balance exploitation-exploration. While Gaussian process (GP) has been a flexible and favored surrogate model, its scalability issues have spurred recent alternatives whose convergence properties are nonetheless more opaque. Motivated by these dilemmas, we propose an axiomatic framework, which we call Pseudo-Bayesian Optimization, that elicits the minimal requirements to guarantee black-box optimization convergence beyond GP-based methods. The design freedom in our framework subsequently allows us to construct algorithms that are both scalable and empirically superior. In particular, we show how using simple local regression, together with an uncertainty quantifier that adapts the "randomized prior" idea in reinforcement learning, not only guarantees convergence but also consistently outperforms state-of-the-art benchmarks in examples ranging from high-dimensional synthetic experiments to realistic hyperparameter tuning and robotic applications.

Artificial Replay: How to get the most out of your data (4:30 PM)

Presenter: BANERJEE, Siddhartha (Cornell University)

How best to incorporate historical data for initializing control policies is an important open question for using RL in practice: more data should help get better performance, but naively initializing policies using historical samples can suffer from spurious data and imbalanced data coverage, leading to computational and storage issues. To get around this, we will propose a simple meta-algorithm called Artificial Replay for incorporating historical data in control policies. We will first illustrate this for multi-armed bandits, showing how our approach uses a fraction of the historical data compared to a full warm-start, while achieving identical regret guarantees. Next we will extend this to a much more general class of problems we call Markov Decision Processes with Exogenous Inputs (or Exo-MDPS), where the uncertainty affecting the system can be represented as being exogenous to the system state. Here, we will show how our algorithms achieve data-efficiency by leveraging a key insight: using samples of the exogenous input we can infer counterfactual consequences, that then accelerate policy improvements. We will discuss how we can show formal regret guarantees for such systems using the compensated coupling, and also demonstrate its use in virtual machine allocation on real datasets from a large public cloud provider, where our approach outperforms domain-specific heuristics, as well as alternative state-of-the-art reinforcement learning algorithms.

Parallel session: Reinforcement learning and queueing II - A001 (3:30 PM - 5:00 PM)

Session chair: Yuan Zhong

Instability and stability of parameter agnostic policies in parallel server systems (3:30 PM)

Presenter: ZHONG, Yuan (University of Chicago)

We analyze the stability properties of parameter agnostic service policies for parallel server systems. Parameter agnostic policies rely only on the current systems state to make service decisions, thus eliminating the need for knowledge about system parameters, making them potentially appealing to deploy in practice. We focus on a broad and natural class of parameter agnostic policies, which are characterized by increasing switching curves, and we consider their performances in the X-model, a simplest class of parallel server systems for which the stability question is unknown. Our main result is large negative: Essentially, for any switching curve policy, there exists instances of stabilizable parameters under which the given policy leads to instability. The proof involves a novel coupling with a state-dependent birth-death process, which may be of independent interest. In addition, we study various classes of parameter agnostic policies and characterize their regions of instability.

Learning payoffs while routing in skill-based queues (4:00 PM)

Presenter: VAN KEMPEN, Sanne (Eindhoven University of Technology)

We consider skill based routing in queueing networks with heterogeneous customers and servers, where the quality of service is measured by customer-server dependent random rewards and the reward structure is a priori unknown to the system operator. We analyze routing policies that simultaneously learn the system parameters and optimize the reward accumulation, while satisfying queueing stability constraints. To this end, we introduce a model that integrates queueing dynamics and decision making. We use learning techniques from the multi-armed bandit (MAB) framework to propose a definition of regret against a suitable oracle reward and formulate an instance-dependent asymptotic regret lower bound. Since our lower bound is of the same order as results in the classical MAB setting, an asymptotically optimal learning algorithm must exploit the structure of the queueing system to learn as efficiently as in the classical setting, where decisions are not constrained by state space dynamics. We discuss approaches to overcome this by leveraging the analysis of the transient behavior of the queueing system.

Online Load Balancing and Auto Scaling: Stochastic approximation algorithms for LP-based policy (4:30 PM)

Presenters: LOPES, Lucas (UFRJ), REIFFERS, Alexandre (IMT Atlantique)

This talk introduces a novel online algorithm for tuning load balancers coupled with auto-scalers, considering bursty traffic arriving at finite queues with large buffer sizes and with large action space, and when the parameters of the queuing system are not known. When the policy of the queue is known, the problem can be modeled as a weakly coupled Markov Decision Process (MDP). LP-based policies can be used to solve weakly coupled Markov Decision Processes (MDP) and are known to be efficient heuristics. However, to compute such a heuristic, it is necessary to know the transition matrix of the underlying Markov chain. Moreover, a second challenge is that the number of variables and constraints may be too high to efficiently use the LP-based policy in practice. In this talk, we will present a new online algorithm to learn efficiently the LP-based policy. Our algorithm is based on a two-time scale algorithm, and MDP aggregation techniques. We can derive finite bounds between the errors made by our algorithms with the optimal policy.

Friday, June 21, 2024

Keynote: Emmanuel Rachelson (ISAE-SUPAERO) - Amphi B00 (9:30 AM - 10:30 AM)

Lipschitz Lifelong Reinforcement Learning: transferring value functions across MDPs (9:30 AM)

Abstract: How close are the optimal value functions of two Markov decision processes that share the same state and action spaces but have different dynamics and rewards? In this talk, we will consider the problem of knowledge transfer when an agent is facing a series of reinforcement learning (RL) tasks. We will introduce a novel metric between Markov decision processes (MDPs) and establish that close MDPs have close optimal value functions. These theoretical results lead us to a value-transfer method for Lifelong RL, which we use to build a PAC-MDP algorithm with improved convergence rate. Beyond value transfer, this talk will open up on challenges and opportunities deriving from such an analysis.

Coffee break - B006-B007 (10:30 AM - 11:00 AM)

Parallel session: Reinforcement learning for real-life applications - A001 (11:00 AM - 12:30 PM)

Non-Stationary Gradient Descent for Optimal Auto-Scaling in Serverless Platforms (11:00 AM)

Presenter: GAUJAL, Bruno (Inria)

To efficiently manage serverless computing platforms, a key aspect is the auto-scaling of services, i.e., the set of computational resources allocated to a service adapts over time as a function of the traffic demand. The objective is to find a compromise between user-perceived performance and energy consumption. In this paper, we consider the "scale-per-request" auto-scaling pattern and investigate how many function instances (or servers) should be spawned each time an unfortunate job arrives, i.e., a job that finds all servers busy upon its arrival. We address this problem by following a stochastic optimization approach: taking advantage of the ability to observe the system \emph{state} over time, we develop a stochastic gradient descent scheme of the Kiefer--Wolfowitz type. At each iteration, the proposed scheme computes an estimate of the number of servers to spawn each time an unfortunate job arrives to minimize some cost function. Under natural assumptions, we show that the sequence of estimates produced by our scheme is asymptotically optimal almost surely. In addition, we prove that its convergence rate is $O(n^{-2/3})$ where n is the number of iterations. From a mathematical point of view, the stochastic optimization framework induced by auto-scaling exhibits non-standard aspects that we approach from a general point of view. We consider the setting where a controller can only get samples of the transient -- rather than stationary -- behavior of the underlying stochastic system. To handle this difficulty, we develop arguments that exploit properties of the mixing time of the underlying Markov chain. By means of numerical simulations, we validate the proposed approach and quantify its gain with respect to common existing scale-up rules.

A Deep-Learning Approach to High-Dimensional Impulse Control with Applications to Inventory Management (11:30 AM)

Presenter: VAN EEKELEN, Wouter Johannes (University of Chicago, Booth School of Business)

We consider impulse control problems where the system controller can intervene in the state process by means of jumps in the underlying state space. So far, it has not been possible to efficiently solve these problems numerically in high dimensions due to the dreaded “curse of dimensionality.” To tackle this challenge, we introduce a novel deep-learning framework. Grounded in the theory of backward stochastic differential equations (BSDEs), this framework relies crucially on probabilistic identities that exploit a deep connection between impulse control and stochastic target problems. We demonstrate the efficacy of our approach for a class of joint replenishment problems with Brownian demands, in which procurement fixed costs can be saved by replenishing a group of different types of items at a time.

Using Reinforcement and Optimization for grid operation - A solution to the L2PRN 23 competition (12:00 PM)

Presenter: LAIR, Nicolas (Artelys)

Context In the context of the emerging risks faced by the electrical grid, a number of initiatives have been launched by major players to devise innovative ways of operating the power grid based on optimization and machine learning [1-3]. Among them,

RTE, French TSO, is animating the L2RPN competition (Learning to Run a Power Network) [2] to encourage the development of solutions based on Reinforcement Learning approach. In this competition, agents learns to operate a synthetic grid network in real time during one-week scenarios and to deal with forced outages. A successful agent is able to optimize cost operation while avoiding blackouts and favoring low carbon emission energy source. The competition is based on the GridAlive software ecosystem to model the grid and the interaction of the agent [4]. ## Multi-agent framework combining Reinforcement Learning, Optimization and Expert Heuristics For our participation in the 2023 edition, we developed a solution that ranked first on the private leaderboard of the competition. It is based on a multi-agent framework that allows a cooperation between specialized agents. The solution mainly relies on a topology agent that is learned based on a curriculum learning approach [5, 6]. It is build in an iterative way, from greedy agents whose roles are to identify most relevant actions on the grid, to model-based agent trained by reinforcement learning to take into account the dynamic of scenarios. The topology agent is backed by a resdispatch agent when no satisfactory topological configuration could be find and by expert agents to deal with individual case. ## References [1] ARPA-E. Grid Optimization Competition | Challenge 3. url: https://gocompetition.energy.gov/challenges/challenge-3/. [2] Antoine Marot et al. “Learning to run a power network challenge for training topology controllers”. In: Electric Power Systems Research 189 (Dec. 2020), p. 106635. issn: 0378-7796. doi: 10.1016/J.EPSR.2020.106635. [3] Jan Viebahn et al. “Potential and challenges of AI-powered decision support for short-term system operations”. In: (2022). [4] RTE. Grid Alive. url: https://github.com/rte-france/gridAlive. [5] Yoshua Bengio et al. “Curriculum learning”. In: ICML 09: Proceedings of the 26th Annual International Conference on Machine Learning. Vol. 382. New York, New York, USA: ACM Press, 2009, pp. 1–8. isbn: 9781605585161. doi: 10.1145/1553374.1553380. [6] Malte Lehna et al. “Managing power grids through topology actions: A comparative study between advanced rule-based and reinforcement learning agents”. In: Energy and AI 14 (Oct. 2023), p. 100276. issn: 2666-5468. doi: 10.1016/J.EGYAI.2023.100276.

Parallel session: Reinforcement learning for wireless scheduling - A002 (11:00 AM - 12:30 PM)

Organizer and chair: I-Hong Hou

Achieving Regular and Fair Learning in Combinatorial Multi-Armed Bandit (11:00 AM)

Presenter: LI, Bin (Pennsylvania State University)

Combinatorial multi-armed bandit refers to the model that aims to maximize cumulative rewards in the presence of uncertainty. Motivated by two important wireless network applications, in addition to maximizing cumulative rewards, it is important to ensure fairness among arms (i.e., the minimum average reward required by each arm) and reward regularity (i.e., how often each arm receives the reward). In this paper, we develop a parameterized regular and fair learning algorithm to achieve these three objectives. In particular, the proposed algorithm linearly combines virtual queue-lengths (tracking the fairness violations), Time-Since-Last-Reward (TSLR) metrics, and Upper Confidence Bound (UCB) estimates in its weight measure. Here, TSLR is similar to age-of-information and measures the elapsed number of rounds since the last time an arm received a reward, capturing the reward regularity performance, and UCB estimates are utilized to balance the tradeoff between exploration and exploitation in online learning. Through capturing a key relationship between virtual queue-lengths and TSLR metrics and utilizing several non-trivial Lyapunov functions, we analytically characterize zero cumulative fairness violation, reward regularity, and cumulative regret performance under our proposed algorithm. These findings are corroborated by our extensive simulations.

Structured Reinforcement Learning for Delay-Optimal Data Transmission in Dense mmWave Networks (11:30 AM)

Presenter: LI, Jian (Stony Brook University)

We study the data packet transmission problem (mmDPT) in dense cell-free millimeter wave (mmWave) networks, i.e., users sending data packet requests to access points (APs) via uplinks and APs transmitting requested data packets to users via downlinks. Our objective is to minimize the average delay in the system due to APs' limited service capacity and unreliable wireless channels between APs and users. This problem can be formulated as a restless multi-armed bandits problem with fairness constraint (RMAB-F). Since finding the optimal policy for RMAB-F is intractable, existing learning algorithms are computationally expensive and not suitable for practical dynamic dense mmWave networks. In this work, we propose a structured reinforcement learning (RL) solution for mmDPT by exploiting the inherent structure encoded in RMAB-F. To achieve this, we first design a low-complexity and provably asymptotically optimal index policy for RMAB-F. Then, we leverage this structure information to develop a structured RL algorithm called mmDPT-TS, which provably achieves an $\tilde{\cO}(\sqrt{T})$ Bayesian regret. More importantly, mmDPT-TS is computation-efficient and thus amenable to practical implementation, as it fully exploits the structure of index policy for making decisions. Extensive emulation based on data collected in realistic mmWave networks demonstrate significant gains of mmDPT-TS over existing approaches.

Multi-Agent Reinforcement Learning for Collaborative Decision-Making in Network Optimization (12:00 PM)

Presenter: LAN, Tian (George Washington University)

Abstract: Reinforcement Learning has demonstrated tremendous success in many challenging tasks with superhuman performance. Nevertheless, many of the decision-making problems in network optimization/scheduling naturally involve the participation of multiple decision-making agents (e.g., a network of routers/switches and a group of decentralized controllers) and thus need to be modeled as Multi-Agent Reinforcement Learning (MARL) problems. As the number of agents grows, we start to encounter “the curse of many agents” -- the exponential growth of MARL problem space significantly hinders the development of collaborative exploration strategies as well as the learning of joint decision-making policies. In this talk, we will present our recent research on multi-agent learning algorithms and multi-agent option/skill discovery, as well as their applications to traffic engineering, scheduling, and 5G slice management.

Lunch - C101-C103 (12:30 PM - 2:00 PM)

Keynote: Sean Meyn (University of Florida) - Amphi B00 (2:00 PM - 3:00 PM)

The Projected Bellman Equation in Reinforcement Learning (2:00 PM)

Abstract: A topic of discussion throughout the 2020 Simons program on reinforcement learning: is the Q-learning algorithm convergent outside of the tabular setting? It is now known that stability can be assured using a matrix gain algorithm, but this requires assumptions, which begs the next question: does a solution to the projected Bellman equation exist? This is the most minimal requirement for convergence of any algorithm. The question was resolved in very recent work. A solution does exist, subject to two assumptions: the function class is linear, and (far more crucial) the input used for training is a form of epsilon-greedy policy with sufficiently small epsilon. Moreover, under these conditions it is shown that the Q-learning algorithm is stable, in terms of bounded parameter estimates. Convergence remains one of many open topics for research. In short, sufficient optimism is not only valuable for algorithmic efficiency, but is a means to algorithmic stability.