European Workshop
On
Reinforcement Learning

EWRL

The 13th European Workshop on Reinforcement Learning (EWRL 2016)

Dates: December 3-4 2016
Location: Pompeu Fabra University, Barcelona, Spain. Campus Ciutadella. Ramon Turró building (building number 13). Carrer Ramon Turró, 1

PROGRAM

Saturday 3rd

8:30 – 8:40 Opening remarks

8:40 – 9:20 Invited talk: **Bruno Scherrer** – INRIA

Periodic Markov Decision Process
After introducing the standard infinite-horizon discounted optimal control problem formalized by Markov Decision Processes (MDPs), I shall describe a simple extension to periodic problems, that is where the reward and transition functions are periodic functions of time. I will describe how to naturally adapt the standard dynamic programming (DP) algorithms to this extension. I will discuss the properties of these algorithms, in particular when they are run approximately. I will mention the following somewhat surprising result: the bigger the period of the problem, the less sensitive to approximation the DP algorithms. Among other things, this will bring some new light on a series of very similar results suggesting to consider looking for non-stationary periodic solutions even in the (stationary) standard case.

9:20 – 9:40 Contributed talk: **Ian Osband**
Why is Posterior Sampling Better for RL?

9:40 – 10:00 Contributed talk: **Marc Abeille**
Linear Thompson Sampling Revisited

10:00 – 10:30 Coffee break

10:30 – 11:10 Invited talk: **Hector Geffner** – Universitat Pompeu Fabra

Exploiting structure for planning and eventually for learning
Research in planning has uncovered effective techniques for dealing with problems involving large numbers of state variables and huge state spaces. This includes the automatic derivation of heuristic functions, structure-based transformations, and width-based search. In this talk, I'll review these techniques and discuss their potential relevance to learning.

11:10 – 11:50 Invited talk: **John Langford** – Microsoft Research

The Contextual Reinforcement Learning Research Program
The theory of Reinforcement Learning for Markov Decision Processes is a well-developed failure as evidenced by common practice where people instead use algorithms with glaring weaknesses like epsilon-greedy Q-learning. Why? And how do we go forward? I will discuss a new approach based upon _contextual_ analysis where you compete with a set of policies, similar to existing supervised learning algorithms. This is both an overview and discussion of new results.

11:50 – 12:10 Contributed talk: **Tom Zahavy**
A Deep Hierarchical Approach to Lifelong Learning in Minecraft

12:10 – 13:30 Poster session 1

- **Robust Kalman Temporal Difference.** *Shirli Di-Castro Shashua and Shie Mannor*
- **A Lower Bound for Multi-Armed Bandits with Expert Advice.** *Yevgeny Seldin and Gábor Lugosi*
- **Magical Policy Search: Data Efficient Reinforcement Learning with Guarantees of Global Optimality.** *Philip Thomas and Emma Brunskill*
- **Approximations of the Restless Bandit Problem.** *Steffen Grunewalder and Azadeh Khaleghi*
- **Bayesian Optimal Policies for Asynchronous Bandits with Known Trends.** *Mohammed Amine Alaoui, Tanguy Urvoy and Fabrice Clérot*
- **Exploration Potential.** *Jan Leike*
- **Corrupt Bandits.** *Pratik Gajane, Tanguy Urvoy and Emilie Kaufmann*
- **Iterative Hierarchical Optimization for Misspecified Problems.** *Daniel J. Mankowitz, Timothy Mann and Shie Mannor.*
- **A Deep Hierarchical Approach to Lifelong Learning in Minecraft.** *Chen Tessler, Shahar Givony, Daniel J. Mankowitz, Tom Zahavy and Shie Mannor*
- **Situational Awareness by Risk-Conscious Skills.** *Daniel J. Mankowitz, Aviv Tamar and Shie Mannor.*

13:30 – 15:00 Lunch break (on your own)

15:00 – 15:40 Invited talk: **Remi Munos** – Google DeepMind and INRIA

Safe and efficient off-policy reinforcement learning

Our goal is to design a RL algorithm with two desired properties: (1) Off-policy learning: useful for exploration, or when we use memory replay, or observe log-data, (2) Use multi-steps returns, in order to propagate rewards faster and avoid accumulation of approximation/estimation errors when doing one-step Bellman updates with function approximation. Both properties are useful in a deep RL setting. We introduce and analyse an algorithm, Retrace, which use multi-steps returns and can safely and efficiently use any off-policy data. As corollary we prove the convergence of Watkin's Q(λ) to Q (open problem since 1989). We report experiments on the Atari domain.

15:40 – 16:00 Contributed talk: **Michal Moshkovitz**
Principled Option Learning in Markov Decision Processes

16:00 – 16:20 Contributed talk: **Ronan Fruit**
Exploration–Exploitation in MDPs with Options

16:20 – 16:50 Coffee break

16:50 – 17:30 Invited talk: **Doina Precup** – McGill University

How to construct good temporal abstractions

The ability to represent, learn and plan with temporally extended actions is an important ingredient of intelligent systems. While planning with temporally extended actions in reinforcement learning, using the options framework, is well understood, creating such abstractions autonomously from data has remained challenging. I will briefly review the option-critic architecture (AAAI'17), which is capable of learning both the internal policies and the termination conditions of options, as well as the policy over options, without the need to provide any additional rewards or subgoals, by using a policy gradient-style approach. However, this framework still leaves open the question of what constitutes a good set of options. In order to move towards an answer, I will discuss new insights into the relationship of the Bellman operator in the options framework to matrix splitting, an approach traditionally used to speed up convergence of iterative solvers for large linear systems of equations. Based on standard comparison theorems for matrix splittings, we can analyze the asymptotic rate of convergence varies as a function of the inherent timescales of the options. This new perspective highlights a trade-off between asymptotic performance and the cost of computation associated with building a good set of options. Joint work with Pierre-Luc Bacon and Jean Harb.

17:30 – 18:10 Invited talk: **Ronald Ortner** – Montanuniversität Leoben

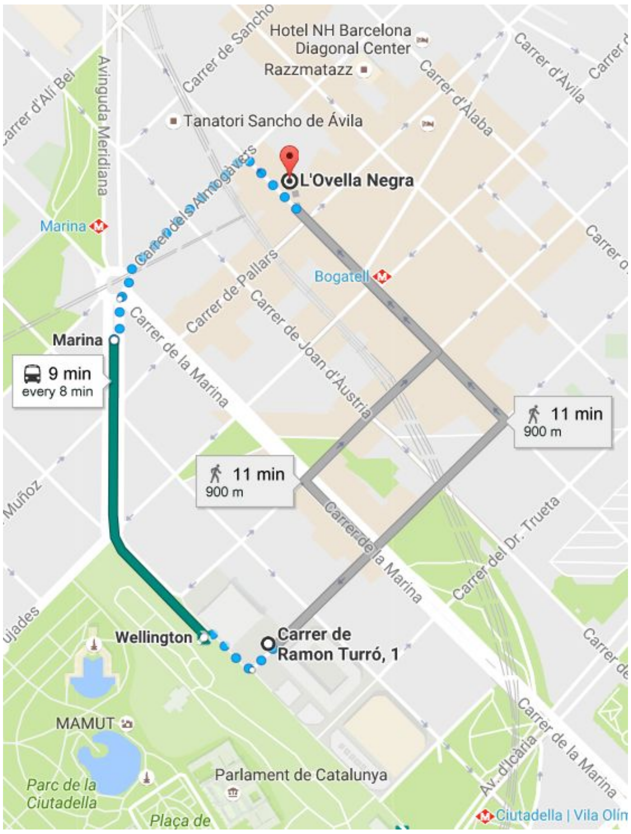
Some Open Problems for Average Reward MDPs
This talk will consider theoretical guarantees in form of regret bounds for reinforcement learning in MDPs with average reward criterion. In this setting, some open problems for discrete as well as for continuous domains will be presented and discussed.

18:10 – 19:00 Panel discussion 1:
Informed exploration in Reinforcement Learning

19:00 – 21:00 Dinner break (on your own)

21:00 – Event: Bad Axes Concert
Megataverna Poble Nou. Ovella Negra
Location: Carrer de Zamora, 78, 08018 Barcelona

The concert will be downstairs. To avoid drunk teenagers, please turn right as soon as you enter



Sunday 4th

9:00 – 9:20 Contributed talk: **Shie Mannor**
Situational Awareness by Risk-Conscious Skills.

9:20 – 9:40 Contributed talk: **Christoph Dann**
Memory Lens: How Much Memory Does an Agent Use?

9:40 – 10:00 Contributed talk: **Bilal Piot**
Batch policy iteration algorithms for continuous domains.

10:00 – 10:30 Coffee break

10:30 – 11:10 Invited talk: **Mohammad Ghavamzadeh** – Adobe Research and INRIA

Learning Safe Policies in Sequential Decision-Making Problems

In online advertisement as well as many other fields such as health informatics and computational finance, we often have to deal with the situation in which we are given a batch of data generated by the current strategy(ies) of the company (hospital, investor), and we are asked to generate a good or an optimal strategy. Although there are many techniques to find a good policy given a batch of data, there are not much results to guarantee that the obtained policy will perform well in the real system without deploying it. On the other hand, deploying a policy might be risky, and thus, requires convincing the product (hospital, investment) manager that it is not going to harm the business. This is why it is extremely important to devise algorithms that generate policies with performance guarantees. In this talk, we discuss four different approaches to this fundamental problem, we call them model-based, model-free, online, and risk-sensitive. In the model-based approach, we first use the batch of data and build a simulator that mimics the behavior of the dynamical system under studies (online advertisement, hospital¹s ER, financial market), and then use this simulator to generate data and learn a policy. The main challenge here is to have guarantees on the performance of the learned policy, given the error in the simulator. This line of research is closely related to the area of robust learning and control. In the model-free approach, we learn a policy directly from the batch of data (without building a simulator), and the main question is whether the learned policy is guaranteed to perform at least as well as a baseline strategy. This line of research is related to off-policy evaluation and control. In the online approach, the goal is to control the exploration of the algorithm in a way that never during its execution the loss of using it instead of the baseline strategy is more than a given margin. In the risk-sensitive approach, the goal is to learn a policy that manages risk by minimizing some measure of variability in the performance in addition to maximizing a standard *criterion. We present algorithms based on these approaches and demonstrate their usefulness in real-world applications such as personalized ad recommendation, energy arbitrage, traffic signal control, and American option pricing.*

11:10 – 11:50 Invited talk: **Alessandro Lazaric** – INRIA

Spectral Methods for Reinforcement Learning

In many learning problems, the actual structure of the environment is not fully revealed to the learner (e.g., a hidden Markov state in partially observable MDPs). On the other hand, correctly reconstructing the hidden structure of the problem allows to significantly speed up the learning process and improve the overall performance. In this talk, we will show how recent spectral tensor decomposition methods for latent variable models can be successfully integrated in online learning algorithms. By combining the finite-time guarantees on the accuracy of the estimated model with the analysis of standard exploration-exploitation algorithms, we show that regret guarantees can be easily derived. In particular, we will review the application of tensor decomposition methods to the multi-armed bandit and the POMDP settings.

11:50 – 12:10 Contributed talk: **Martha White**
Accelerated Gradient Temporal Difference Learning

12:10 – 13:30 Poster session 2

- **Decoding multitask DQN in the world of Minecraft**. Lydia Liu, Urun Dogan and Katja Hofmann
- **Spatio-Temporal Abstractions in Reinforcement Learning Through Neural Encoding.** Nir Baram, Tom Zahavy and Shie Mannor
- **Non-Deterministic Policy Improvement Stabilizes Approximated Reinforcement Learning.** Wendelin Böhmer, Rong Guo and Klaus Obermayer
- **Automatic Representation for Life-Time Value Recommender Systems.** Assaf Hallak, Elad Yom-Tov and Yishay Mansour
- **Using Policy Gradients to Account for Changes in Behavior Policies under Off-policy Control**. Lucas Lehnert and Doina Precup
- **Deep Reinforcement Learning Solutions for Energy Microgrids Management.** Vincent Francois-Lavet, David Taralla, Damien Ernst and Raphael Fonteneau
- **Toward a data efficient neural actor-critic.** Matthieu Zimmer, Yann Boniface and Alain Dutech

13:30 – 15:00 Lunch break (on your own)

15:00 – 15:40 Invited talk: **Emma Brunskill** – Carnegie Mellon University

Helping Unlock the Potential of RL

Reinforcement learning still has a huge unrealized potential to better the world. Part of the reason for this is that many high stakes applications, like education or healthcare, are often (1) wary of employing algorithms without a good understanding of their potential performance, and also (2) may benefit substantially from the continued interaction of a human-in-the-loop. To tackle these issues, I will discuss some of our recent efforts in developing better statistical estimators for off policy evaluation, and RL agents that can actively, with the help of a human, change their underlying domain specification, towards unlocking the benefit of RL for a much wider array of disciplines.

15:40 – 16:00 Contributed talk: **Amir-massoud Farahmand**

Value-Aware Loss Function for Model Learning in Reinforcement Learning

16:00 – 16:20 Contributed talk: **Assaf Hallak**

Consistent On-Line Off-Policy Evaluation

16:20 – 16:50 Coffee break

16:50 – 17:30 Invited talk: **Sergey Levine** – University of Washington and Google

Deep Robotic Learning

Reinforcement learning has long held the promise to enable fully autonomous learning of complex motion skills for robotic manipulation, locomotion, and other behaviors. However, in order to achieve practical and tractable learning for high-dimensional robotic systems, most methods have needed to make certain compromises, such as using low-dimensional hand-designed policy classes, hand-engineered abstractions, or other simplifications. Deep neural networks can represent extremely complex functions, and could allow for the entire pipeline from perception to action to be learned even for complex robotic skills. However, severe algorithmic and practical challenges remain. In this talk, I will discuss algorithms for deep reinforcement learning that are efficient and practical for real-world robotic manipulation skills, discuss how we can overcome the enormous challenges posed by sample complexity, how we can provide robots with a sufficient diversity of data for learning generalizable motion skills, and how we can overcome limited supervision and partial access to reward functions. Though the focus of the talk will be on robotic learning in particular, many of the algorithms we develop for efficient robotic reinforcement learning are broadly applicable to a range of systems that must make intelligent decisions with limited data in the real world.

17:30 – 18:10 Invited talk: **Gerhard Neumann** – University of Lincoln

Information-theoretic methods for learning versatile, reusable skills

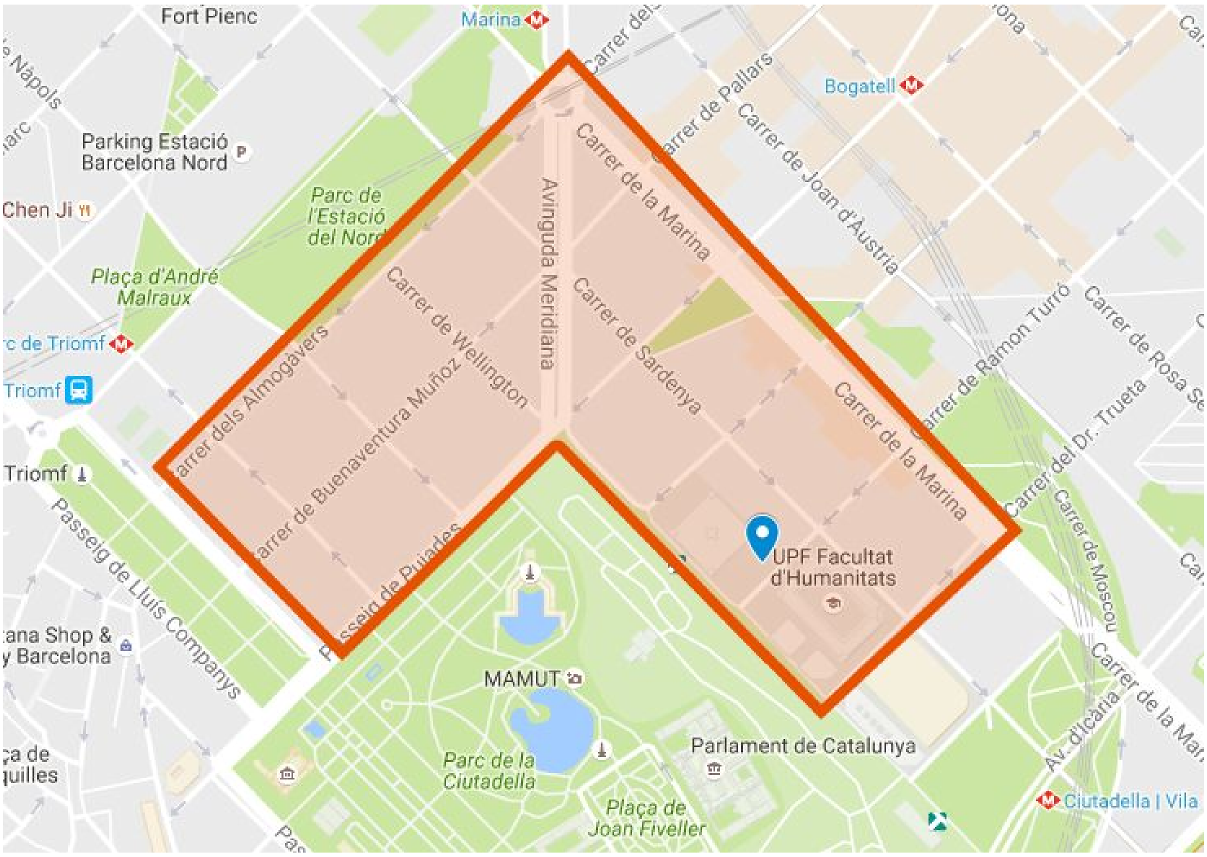
In this talk, I will present our work on information theoretic policy search methods for learning complex motor skills. Throughout my presentation, I will use a simulated robotic table tennis application as working example. We developed new algorithms that are based on information theoretic constraints and compatible value function approximation that can be efficiently used to learn motor skills represented as movement primitives. We extended our framework to learn how to generalize skills, learn reactive skills can react to perturbations, select skills and learn when to switch between the skills, which we all evaluated on our simulated table tennis plattform. I will also present shortly our latest results how to apply similar methods to learning in robot swarms.

18:10 – 19:00 Panel discussion 2

From mountain car to Atari games: what did we learn?

Food & drinks in the neighborhood

There are several good restaurants nearby the venue. We suggest that, during lunchtime, you stay within the area enclosed by Carrer de la Marina, Carrer dels Almogàvers, and carrer de Ramon Turró. (Outside of this area, you may either run into a boring neighborhood with no restaurants, too close to the beach which increases the risk of being ripped off for mediocre food, or just end up so far that you can't make it back to the talks.)



Below is a very short list of arbitrarily selected restaurants that we can recommend.

Spanish:
Taberna Gallega, Carrer de Wellington, 72
Bar Palets, Av. d'Icària, 153
Santa Fe, Buenaventura Muñoz (Wellington)
Italian:
La Forchetta, Avinguda Meridiana, 2
Gastropub:
Café Menssana, Carrer de Sardenya, 48
Babol Burger, Carretera Antiga de Mataró, 22
Sandwiches:
Café de la Pompeu, Carrer de Ramon Trias Fargas, 25-27
Food court (sort of small):
El Centre de la Vila, Carrer de Salvador Espriu, 61

Program Committee

Christos Dimitrakakis
Marc Bellemare
Christian Daniel
Marc Deisenroth
Amir-massoud Farahmand
Victor Gabillon
Matthieu Geist
Mohammad Ghavamzadeh
Mohammad Gheshlaghi Azar
Nan Jiang
Anders Jonsson
Akshay Krishnamurthy
Tor Lattimore
Alessandro Lazaric
Ashique Rupam Mahmood
Timothy Mann
Jérémie Mary
Rémi Munos
Laurent Orseau
Ronald Ortner
Ian Osband
Bilal Piot
Doina Precup
Marcello Restelli
Scott Sanner
Georgios Theocharous
Michal Valko

Organizing Committee

Gergely Neu
Vicenç Gómez
Csaba Szepesvari

Google DeepMind

Artificial Intelligence
www.elsevier.com/locate/artint

upf. Universitat Pompeu Fabra Barcelona

Departament de Tecnologies de la Informació i les Comunicacions

EXCELENCIA MARÍA DE MAEZTU