

# 8<sup>th</sup> European Workshop on Reinforcement Learning

June 30 – July 3, 2008  
Lille, France



## Program

<b>June 30, Monday - 10:00</b> . . . . .	<b>1</b>
Regularized Fitted Q-iteration: Application To Bounded Resource Planning <i>Amir Massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvari and Shie Mannor</i>	
Regularized Policy Iteration <i>Amir Massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvari and Shie Mannor</i>	
<b>June 30, Monday - 11:05</b> . . . . .	<b>2</b>
Sample-Based Learning And Search With Permanent And Transient Memories <i>David Silver, Rich Sutton and Martin Mueller</i>	
Parameter Tuning By The Cross-Entropy Method <i>Guillaume Chaslot, Mark Winands, Istvan Szita and Jaap Van Den Herik</i>	
Including Expert Knowledge In Bandit-Based Monte-Carlo Planning, With Application To Computer-Go <i>Louis Chatriot, Sylvain Gelly, Hoock Jean-Baptiste, Julien Perez, Arpad Rimmel and Olivier Teytaud</i>	
<b>June 30, Monday - 14:15</b> . . . . .	<b>3</b>
United We Stand: Population Based Methods For Solving Unknown POMDPs <i>Noel Welsh and Jeremy Wyatt</i>	
Reinforcement Learning With History Lists <i>Stephan Timmer and Martin Riedmiller</i>	
A Near Optimal Policy For Channel Allocation In Cognitive Radio <i>Sarah Filippi, Olivier Cappé, Fabrice Clérot and Eric Moulines</i>	
<b>June 30, Monday - 15:45</b> . . . . .	<b>4</b>
Evaluation Of Batch-Mode Reinforcement Learning Methods For Solving DEC-MDPs With Changing Action Sets <i>Thomas Gabel and Martin Riedmiller</i>	
Solving Analytic Multi-Agent Stochastic Processes <i>Luke Dickens, Krysia Broda and Alessandra Russo</i>	
<b>June 30, Monday - 16:50</b> . . . . .	<b>-</b>
Tutorial: Multi-Automata Learning <i>Ann Nowé, Katja Verbeeck and Peter Vrancx</i>	
<b>July 1, Tuesday - 09:10</b> . . . . .	<b>5</b>
Multigrid Reinforcement Learning With Reward Shaping <i>Marek Grzes and Daniel Kudenko</i>	
Reinforcement Learning With The Use Of Costly Features <i>Robby Goetschalckx, Scott Sanner and Kurt Driessens</i>	
Tile Coding Based On Hyperplane Tiles <i>Daniele Loiacono and Pier Luca Lanzi Pier Luca Lanzi</i>	
Using Decision Trees As The Answer Networks In Temporal Difference-Networks <i>Laura-Andreea Antanas, Kurt Driessens, Jan Ramon and Tom Croonenborghs</i>	
<b>July 1, Tuesday - 11:05</b> . . . . .	<b>6</b>
The Many Faces Of Optimism: A Unifying Approach <i>Istvan Szita and Andras Lorincz</i>	
On Upper-Confidence Bound Policies For Non-Stationary Bandit Problems <i>Aurélien Garivier and Eric Moulines</i>	
Reinforcement Learning By Direct Optimal Value Estimation And Regret Minimization <i>Manuel Loth and Philippe Preux</i>	
<b>July 1, Tuesday - 14:15</b> . . . . .	<b>-</b>
Invited Talk <i>Richard S. Sutton</i>	

<b>July 1, Tuesday - 15:45</b> . . . . .	<b>7</b>
Adaptive Treatment Of Epilepsy Via Batch-mode Reinforcement Learning <i>Arthur Guez, Robert Vincent, Massimo Avoli and Joelle Pineau</i>	
Reinforcement Learning With Markov Logic Networks <i>Weiwei Wang, Xingguo Chen and Yang Gao</i>	
Use Of Reinforcement Learning In Two Real Applications <i>Jose D. Martin-Guerrero, Emilio Soria, Marcelino Martínez, Antonio José Serrano, Rafael Magdalena and Juan Gómez-Sanchis</i>	
<b>July 1, Tuesday - 17:15</b> . . . . .	<b>8</b>
Knows What It Knows: A Framework For Self-Aware Learning <i>Lihong Li, Michael Littman and Thomas Walsh</i>	
Reinforcement Learning In The Presence Of Rare Events <i>Jordan Frank, Shie Mannor and Doina Precup</i>	
A Metric Analogue To MDP Homomorphisms <i>Jonathan Taylor, Doina Precup and Prakash Panangaden</i>	
<b>July 2, Wednesday - 09:00</b> . . . . .	<b>-</b>
Invited Talk <i>Dimitri Bertsekas</i>	
<b>July 2, Wednesday - 10:00</b> . . . . .	<b>9</b>
New Error Bounds For Approximations From Projected Linear Equations <i>Huizhen Yu and Dimitri Bertsekas</i>	
Model-based Reinforcement Learning With State Aggregation <i>Cosmin Paduraru, Robert Kaplow, Doina Precup and Joelle Pineau</i>	
<b>July 2, Wednesday - 11:05</b> . . . . .	<b>10</b>
Basis Expansion In Natural Actor Critic Methods <i>Sertan Girgin and Philippe Preux</i>	
Variable Metric Reinforcement Learning Methods Applied To The Noisy Mountain Car Problem <i>Verena Heidrich-Meisner and Christian Igel</i>	
Policy Learning – A Unified Perspective With Applications In Robotics <i>Jan Peters, Jens Kober and Duy Nguyen-Tuong</i>	
<b>July 2, Wednesday - 14:15</b> . . . . .	<b>11</b>
Exploiting Additive Structure In Factored MDPs For Reinforcement Learning <i>Thomas Degris, Olivier Sigaud and Pierre-Henri Wuillemin</i>	
Hierarchical Reinforcement Learning In Factored MDPs <i>Olga Kozlova, Olivier Sigaud and Christophe Meyer</i>	
Bayesian Reward Filtering <i>Matthieu Geist, Olivier Pietquin and Gabriel Fricout</i>	
<b>July 2, Wednesday - 15:45</b> . . . . .	<b>12</b>
Transfer Of Samples In Batch Reinforcement Learning <i>Alessandro Lazaric, Marcello Restelli and Andrea Bonarini</i>	
Privacy-Preserving Reinforcement Learning <i>Jun Sakuma, Shigenobu Kobayashi and Rebecca Wright</i>	
Multi-Agent Model-Based Reinforcement Learning Experiments In The Pursuit Evasion Game <i>Bruno Bouzy and Marc Metivier</i>	
<b>July 2, Wednesday - 17:15</b> . . . . .	<b>13</b>
A Family Of Reinforcement Learning Algorithms <i>Marco Wiering and Hado Van Hasselt</i>	
Empirical Bernstein Stopping <i>Volodymir Mnih and Csaba Szepesvari</i>	
Algorithms And Bounds For Sampling-based Approximate Policy Iteration <i>Christos Dimitrakakis and Michail Lagoudakis</i>	
Rollout Sampling Approximate Policy Iteration <i>Christos Dimitrakakis and Michail Lagoudakis</i>	

<b>July 3, Thursday - 09:00</b> . . . . .	<b>-</b>
Invited Talk <i>Jan Peters</i>	
<b>July 3, Thursday - 10:00</b> . . . . .	<b>14</b>
Efficient Reinforcement Learning In Parameterized Models: Discrete Parameter Case. <i>Kirill Dyagilev, Shie Mannor and Nahum Shimkin</i>	
Robustness Analysis Of SARSA( $\lambda$ ): Different Models Of Reward And Initialisation <i>Marek Grzes and Daniel Kudenko</i>	
<b>July 3, Thursday - 11:05</b> . . . . .	<b>15</b>
Lazy Planning Under Uncertainty By Optimizing Decisions On An Ensemble Of Incomplete Disturbance Trees <i>Boris Defourny, Damien Ernst and Louis Wehenkel</i>	
Optimistic Planning Of Deterministic Systems <i>Jean-Francois Hren and Remi Munos</i>	
Policy Optimization By Implicit Probabilistic Simulation <i>Carl Rasmussen and Marc Deisenroth</i>	
<b>July 3, Thursday - 14:15</b> . . . . .	<b>16</b>
Reinforcement Learning Of Perceptual Coupling For Motor Primitives <i>Jens Kober and Jan Peters</i>	
Applications Of Reinforcement Learning To Structured Prediction <i>Francis Maes, Ludovic Denoyer and Patrick Gallinari</i>	
Policy Iteration For Learning An Exercise Policy For American Options <i>Yuxi Li and Dale Schuurmans</i>	
<b>July 3, Thursday - 15:45</b> . . . . .	<b>17</b>
Adaptive Aggregation For Reinforcement Learning With Efficient Exploration: Deterministic Domains <i>Andrey Bernstein and Nahum Shimkin</i>	
Approximate Policy Iteration For Generalized Semi-Markov Decision Processes: An Improved Algorithm <i>Emmanuel Rachelson, Patrick Fabiani and Frédérick Garcia</i>	
Markov Decision Processes With Arbitrary Reward Processes <i>Jia Yuan Yu, Shie Mannor and Nahum Shimkin</i>	
<b>July 3, Thursday - 17:15</b> . . . . .	<b>18</b>
Relational TD Reinforcement Learning <i>Christophe Rodrigues, Pierre Gérard and Celine Rouveirol</i>	
Reinforcement Learning With Markov Logic Networks <i>Weiwei Wang, Xingguo Chen and Yang Gao</i>	
Classifier-Based Policy Representation <i>Michail Lagoudakis and Ioannis Rexakis</i>	

# Session 1

**June 30, Monday - 10:00**

## Regularized Fitted Q-iteration: Application To Bounded Resource Planning

*Amir Massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvari and Shie Mannor*

We consider bounded resource planning in a Markovian decision problem, i.e., the problem of finding a good policy given access to a generative model of the environment and a limit on the computational resources. We propose to use fitted Q-iteration algorithm with penalized least-squares regression as the regression subroutine to address the problem of selecting an appropriate function approximator in each iteration. The algorithm is presented in detail for the case when the function space is a reproducing-kernel Hilbert space underlying a user chosen kernel-function. We derive bounds on the quality of solutions found and argue how data-dependent penalties can lead to almost optimal performance. A simple example is used to illustrate the benefits of using a penalized procedure.

## Regularized Policy Iteration

*Amir Massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvari and Shie Mannor*

In this paper we consider approximate policy iteration based reinforcement learning algorithms. In order to implement a flexible function approximation method we propose the use of a non-parametric methods with regularization, providing a convenient way to control the complexity of the function approximator through changing a single parameter. This idea is explored in the context of policy iteration for the purpose of evaluating policies. Two specific ways of implementing regularization are proposed: One for LSTD, one for the recent (modified) Bellman residual minimization (BRM) method. We derive efficient implementations when regularized solutions are sought for over reproducing kernel Hilbert spaces. For the BRM method we prove generalization bounds.

# Session 2

June 30, Monday - 11:05

## Sample-Based Learning And Search With Permanent And Transient Memories

*David Silver, Rich Sutton and Martin Mueller*

We present a reinforcement learning architecture, Dyna-2, that encompasses both sample-based learning and sample-based search, and that generalises across states during both learning and search. We apply Dyna-2 to high performance Computer Go. In this domain the most successful planning methods are based on sample-based search algorithms, such as UCT, in which states are treated individually, and the most successful learning methods are based on temporal-difference learning algorithms, such as Sarsa, in which linear function approximation is used. In both cases, an estimate of the value function is formed, but in the first case it is transient, computed and then discarded after each move, whereas in the second case it is more permanent, slowly accumulating over many moves and games. The idea of Dyna-2 is for the transient planning memory and the permanent learning memory to remain separate, but for both to be based on linear function approximation and both to be updated by Sarsa. To apply Dyna-2 to 9x9 Computer Go, we use a million binary features in the function approximator, based on templates matching small fragments of the board. Using only the transient memory, Dyna-2 performed at least as well as UCT. Using both memories combined, it significantly outperformed UCT. Our program based on Dyna-2 achieved a higher rating on the Computer Go Online Server than any handcrafted or traditional search based program.

## Parameter Tuning By The Cross-Entropy Method

*Guillaume Chaslot, Mark Winands, Istvan Szita and Jaap Van Den Herik*

Recently, Monte-Carlo Tree Search (MCTS) has become a popular approach for intelligent play in games. Amongst others, it is successfully used in most state-of-the-art Go programs. To improve the playing strength of these Go programs any further, many parameters dealing with MCTS should be fine-tuned. In this paper, we propose to apply the Cross-Entropy Method (CEM) for this task. The method is comparable to Estimation-of-Distribution Algorithms (EDAs), a new area of evolutionary computation. We tested CEM by tuning various types of parameters in our Go program Mango. The experiments were performed in matches against the open-source program GNU Go. They revealed that a program with the CEM-tuned parameters played better than without. Moreover, Mango plus CEM outperformed the regular Mango for various time settings and board sizes. From the results we may conclude that parameter tuning by CEM genuinely improved the playing strength of Mango, for various time settings. This result may be generalized to other game engines using MCTS.

## Including Expert Knowledge In Bandit-Based Monte-Carlo Planning, With Application To Computer-Go

*Louis Chatriot, Sylvain Gelly, Hoock Jean-Baptiste, Julien Perez, Arpad Rimmel and Olivier Teytaud*

We present our work on the introduction of expert knowledge in a Bandit-Based Monte-Carlo Planning algorithm applied to the game of Go. The contributions include (i) opening books (ii) bias in the tree part (iii) improvement of the Monte-Carlo playouts and point out general elements around Bandit-Based Monte-Carlo Planning, namely the risk of diversity loss in random playouts, the efficiency of counter-examples for Monte-Carlo design or for the introduction of bias. The resulting program has recently won a non-blitz game against a professional player in 9x9 Go.

# Session 3

June 30, Monday - 14:15

## United We Stand: Population Based Methods For Solving Unknown POMDPs

*Noel Welsh and Jeremy Wyatt*

Solving large unknown POMDPs is an open research problem. Policy search is one solution method that is attractive as it scales in the size of the policy, which is typically much simpler than the environment. We present a global search algorithm capable of finding good policies for POMDPs that are substantially larger than previously reported results. Our algorithm is general; we show it can be used with, and improves the performance of, existing local search techniques such as gradient ascent. Sharing information between the members of the population is the key to our algorithm and we show it results in better performance than equivalent parallel searches that do not share information. Unlike previous work our algorithm does not require the size of the policy be known in advance.

## Reinforcement Learning With History Lists

*Stephan Timmer and Martin Riedmiller*

To represent an optimal policy for a partially observable Markov decision process (POMDP), it is necessary to use some form of memory. Perfect memory is provided by the belief space, the space of probability distributions over states. Unfortunately, computing policies defined on the belief space requires a model and is expensive in terms of computation time. In this article, we will present a model-free algorithm for solving deterministic POMDPs by using memory based on history lists. In contrast to belief states, history lists do not allow to compute optimal policies, but are far more practical and make the learning process much more efficient. We show that by using abstract state spaces, our method also applies for MDPs with continuous state spaces.

## A Near Optimal Policy For Channel Allocation In Cognitive Radio

*Sarah Filippi, Olivier Cappé, Fabrice Clérot and Eric Moulines*

Channel allocation problems can be modeled by a Partially Observable Markov Decision Processes (POMDP). We develop an approach to compute a near optimal policy performing better than the policy introduced by Zhao and al (2007) for this problem. Our method consists in solving a continuous state Markov Decision Process (MDP) using an internal state which is specific for this problem and smaller than the standard belief state. A technique to find a near optimal policy for this MDP involves performing value backups at specific internal points.

# Session 4

**June 30, Monday - 15:45**

## Evaluation Of Batch-Mode Reinforcement Learning Methods For Solving DEC-MDPs With Changing Action Sets

*Thomas Gabel and Martin Riedmiller*

DEC-MDPs with changing action sets and partially ordered transition dependencies have recently been suggested as a sub-class of general DEC-MDPs that features provably lower complexity. In this paper, we investigate the usability of a coordinated batch-mode reinforcement learning algorithm for this class of distributed problems. Our agents acquire their local policies independent of the other agents by repeated interaction with the DEC-MDP and concurrent evolution of their policies, where the learning approach employed builds upon a specialized variant of a neural fitted Q iteration algorithm, enhanced for use in multi-agent settings. We applied our learning approach to various scheduling benchmark problems and obtained encouraging results that show that problems of current standards of difficulty can very well approximately, and in some cases optimally be solved.

## Solving Analytic Multi-Agent Stochastic Processes

*Luke Dickens, Kryisia Broda and Alessandra Russo*

Stochastic Processes exist everywhere we look and modern modelling and learning techniques allow us to automate control of such processes. However, multi-agent models in this context tend to be quite complicated and geared towards particular solution methods. This paper examines the Finite Analytic Stochastic Process (FASP), a flexible and natural alternative to similar modelling tools currently available, which is founded on the important concepts of state encapsulation and transparent mechanics. We show that small systems written in the FASP style can be solved analytically, for arbitrary numbers of agents and with multiple (possibly conflicting) measure/utility signals - analogous to general-sum extensive games from the game theory literature. We discuss the benefits and limitations of this approach and how the results might be interpreted in the multi-agent context.

# Session 5

July 1, Tuesday - 09:10

## Multigrid Reinforcement Learning With Reward Shaping

*Marek Grzes and Daniel Kudenko*

Potential-based reward shaping has been shown to be a powerful method to improve the convergence rate of reinforcement learning agents. It is a flexible technique to incorporate background knowledge into temporal-difference learning in a principled way. However, the question remains how to compute the potential which is used to shape the reward that is given to the learning agent. In this paper we propose a way to solve this problem in reinforcement learning with state space discretisation. In particular, we show that the potential function can be learned online in parallel with the actual reinforcement learning process. If the Q-function is learned for states determined by a given grid, a V-function for states with lower resolution can be learned in parallel and used to approximate the potential for ground learning. The novel algorithm is presented and experimentally evaluated.

## Reinforcement Learning With The Use Of Costly Features

*Robby Goetschalckx, Scott Sanner and Kurt Driessens*

In many practical reinforcement learning problems, the state space is too large to permit an exact representation of the value function, much less the time required to compute it. In such cases, a common solution approach is to compute an approximation of the value function in terms of state features. However, relatively little attention has been paid to the cost of computing these state features. For example, search-based features may be useful for value prediction, but their computational cost must be traded off with their impact on value accuracy. To this end, we introduce a new cost-sensitive sparse linear regression paradigm for value function approximation in reinforcement learning where the learner is able to select only those costly features that are sufficiently informative to justify their computation. We illustrate the learning behavior of our approach using a simple experimental domain that allows us to explore the effects of a range of costs on the cost-performance trade-off.

## Tile Coding Based On Hyperplane Tiles

*Daniele Loiacono and Pier Luca Lanzi Pier Luca Lanzi*

In problems involving large and continuous state-action spaces, the success of reinforcement learning heavily relies on function approximation techniques. Tile coding is a well known function approximator that has been successfully applied to many reinforcement learning problems. In this paper we introduce hyperplane tile coding, in which the usual tiles are replaced with hyperplane tiles to improve the generalization capabilities over the state-action space. In particular, the experimental analysis reported in this work suggest that with hyperplane tile coding broad generalizations result only in a soft degradation of the performance, whereas with usual tile coding they might dramatically decrease the performance.

## Using Decision Trees As The Answer Networks In Temporal Difference-Networks

*Laura-Andreea Antanas, Kurt Driessens, Jan Ramon and Tom Croonenborghs*

Temporal difference networks (or TD-Nets) offer a framework for predictive state representations. TD-Nets break up into two parts: the question network and the answer network. The question network defines which questions about future observations are of importance, while the answer network provides a way to update the answers to those questions as the environment changes. Currently, TD-Nets use logistic regression functions to represent the answer networks. We propose the use of probability trees in their stead. Trees offer a different but powerful way of generalisation and using them may be beneficial in a number of applications. Moreover, we believe this aids in a better understanding of the strengths and weaknesses of TD-Nets and represents an important first step towards the application of temporal difference networks in environments with more extensive, i.e. complex and numerous, observations than those currently employed. We compare the learning behavior of TD-Nets using logistic regression and probability trees using an array of experiments in two simple grid worlds and a ring world.

# Session 6

July 1, Tuesday - 11:05

## The Many Faces Of Optimism: A Unifying Approach

*Istvan Szita and Andras Lorincz*

The exploration-exploitation dilemma has been an intriguing and unsolved problem within the framework of reinforcement learning. “Optimism in the face of uncertainty” and model building play central roles in advanced exploration methods. Here, we integrate several concepts and obtain a fast and simple algorithm. We show that the proposed algorithm finds a near-optimal policy in polynomial time, and give experimental evidence that it is robust and efficient compared to its ascendants.

## On Upper-Confidence Bound Policies For Non-Stationary Bandit Problems

*Aurélien Garivier and Eric Moulines*

In this paper, we consider stochastic nonstationary multi-armed bandit problems (MABP). MABP are considered as a paradigm of the trade-off between exploring the environment to find profitable actions and exploiting what is already known. In the stationary case, the distributions of the rewards do not change in time, Upper-Confidence Bound (UCB) policies have been shown to be rate optimal.

A challenging variant of the MABP is the non-stationary bandit problem where the gambler must decide which arm to play while facing the possibility of a changing environment. In this paper, we consider the situation where the distributions of rewards remain constant over epochs and changes at unknown time instants. We analyze two algorithms: the discounted UCB and the sliding-window UCB. We establish for these two algorithms an upper-bound for the expected regret by upper-bounding the expectation of the number of times a suboptimal arm is played. For that purpose, we derive a Hoeffding type inequality for self normalized deviations with a random number of summands. We establish a general lower-bound for the regret in presence of abrupt changes in the arms reward distributions. We show that the discounted UCB and the sliding-windows UCB both match the lower-bound up to a logarithmic factor.

## Reinforcement Learning By Direct Optimal Value Estimation And Regret Minimization

*Manuel Loth and Philippe Preux*

This short paper briefly introduces an online Reinforcement Learning algorithm of which the concept is to replace the idea of policy improvement — explicit in policy iteration algorithms and implicit in value iteration ones — by the idea of knowledge improvement about the value function of the optimal policy, along a learning policy driven by regret minimization. This is achieved by maintaining for each state a probability distribution over its optimal value, relative to the information gathered so far. This approach can have a high sample efficiency, from both the way updates are performed and the fact that the exploration/exploitation trade-off is well handled by an UCB policy.

# Session 7

July 1, Tuesday - 15:45

## Adaptive Treatment Of Epilepsy Via Batch-mode Reinforcement Learning

*Arthur Guez, Robert Vincent, Massimo Avoli and Joelle Pineau*

This paper highlights the crucial role that modern machine learning techniques can play in the optimization of treatment strategies for patients with chronic disorders. In particular, we focus on the task of optimizing a deep-brain stimulation strategy for the treatment of epilepsy. The challenge is to choose which stimulation action to apply, as a function of the observed EEG signal, so as to minimize the frequency and duration of seizures. We apply recent techniques from the reinforcement learning literature—namely fitted Q-iteration and extremely randomized trees—to learn an optimal stimulation policy using labeled training data from animal brain tissues. Our results show that these methods are an effective means of reducing the incidence of seizures, while also minimizing the amount of stimulation applied. If these results carry over to the human model of epilepsy, the impact for patients will be substantial.

## Reinforcement Learning With Markov Logic Networks

*Weiwei Wang, Xingguo Chen and Yang Gao*

In this paper, we propose a method to combine reinforcement learning and markov logic network which can easily introduce priori knowledge with the weights of first-order formulas, compactly represent state and learn weight efficiently. Most methods in RL are tabular methods, and thus they lack the ability to handle high-dimension problems. Even with function approximation, we often take no account of the inherent relations or connections of the features, otherwise we need to introduce additional features to represent such connections. Markov logic networks(MLN) combines first-order logic and graphical model and it has the ability to compactly represent a wide variety of knowledge. Introducing MLN to RL will bring us a new method to deal with many difficult problems in RL which need some relational representation of state, such as blocks world. Our new reinforcement learning algorithm with Markov logic networks(RMLN) brings a solution to this kind of problems. In our framework, MLN does inference for the action queries and selects a best action, RL uses the successive state, current state and the reward to update the weights of formulas in MLN. With RMLN, priori knowledge can be easily introduced to the learning system and the learning process will become more efficient. Experiment on blocks world illustrates the promise of RMLN.

## Use Of Reinforcement Learning In Two Real Applications

*Jose D. Martin-Guerrero, Emilio Soria, Marcelino Martínez, Antonio José Serrano, Rafael Magdalena and Juan Gómez-Sanchis*

In this paper, we present two successful applications of Reinforcement Learning (RL) in real life. First, the optimization of anemia management in patients undergoing Chronic Renal Failure is presented. The aim is to individualize the treatment (Erythropoietin dosages) in order to stabilize patients within a targeted range of Hemoglobin (Hb). Results show that the use of RL increases the ratio of patients within the desired range of Hb. Thus, patients' quality of life is increased, and additionally, Health Care System reduces its expenses in anemia management. Second, RL is applied to modify a marketing campaign in order to maximize long-term profits. RL obtains an individualized policy depending on customer characteristics that increases long-term profits at the end of the campaign. Results in both problems show the robustness of the obtained policies and suggest their use in other real-life problems..

# Session 8

July 1, Tuesday - 17:15

## Knows What It Knows: A Framework For Self-Aware Learning

*Lihong Li, Michael Littman and Thomas Walsh*

We introduce a learning framework that combines elements of the well-known PAC and mistake-bound models. The KWIK (knows what it knows) framework was designed particularly for its utility in learning settings where active exploration can impact the training examples the learner is exposed to, as is true in reinforcement-learning and active-learning problems. We catalog several KWIK-learnable classes and list some open problems in this area.

## Reinforcement Learning In The Presence Of Rare Events

*Jordan Frank, Shie Mannor and Doina Precup*

We consider the task of reinforcement learning in an environment in which rare significant events occur independently of the actions selected by the controlling agent. If these events are sampled according to their natural probability of occurring, convergence of standard reinforcement learning algorithms is likely to be very slow, and the learning algorithms may exhibit high variance. In this work, we assume that we have access to a simulator, in which the rare event probabilities can be artificially altered. Then, importance sampling can be used to learn with this simulation data. We introduce algorithms for policy evaluation, using both tabular and function approximation representations of the value function. We prove that in both cases, the reinforcement learning algorithms converge. In the tabular case, we also analyze the bias and variance of our approach compared to TD-learning. We evaluate empirically the performance of the algorithm on random Markov Decision Processes, as well as on a large network planning task.

## A Metric Analogue To MDP Homomorphisms

*Jonathan Taylor, Doina Precup and Prakash Panangaden*

We define a metric for measuring behavior similarity between states in a Markov decision process (MDP), in which a state-dependent mapping between actions can be defined. We show that the kernel of our metric corresponds exactly to the classes of states defined by MDP homomorphisms (Ravindran & Barto, 2003). We prove that the difference in the optimal value function of different states can be upper bounded by the value of this metric, and that the bound is tighter than that provided by bisimulation metrics (Ferns et al, 2004, 2005). Our results hold both for discrete and for continuous actions. We provide an algorithm for constructing approximate homomorphisms, by using this metric to identify states that can be grouped together, as well as actions that can be matched. Previous research on this topic is based mainly on heuristics. We illustrate our algorithm on small examples, and show that it has significantly better empirical performance than bisimulation metrics, providing more compact representations of the state space.

# Session 9

**July 2, Wednesday - 10:00**

## New Error Bounds For Approximations From Projected Linear Equations

*Huizhen Yu and Dimitri Bertsekas*

We consider linear fixed point equations and their approximations by projection on a low dimensional subspace. We derive new bounds on the approximation error of the solutions, which are expressed in terms of low dimensional matrices and can be computed by simulation. When the fixed point mapping is a contraction, as is typically the case in Markovian decision processes (MDP), one of our bounds is always sharper than the standard worst case bounds, and another one is often sharper. Our bounds also apply to the non-contraction case, including policy evaluation in MDP with nonstandard projections that enhance exploration. There are no error bounds currently available for this case to our knowledge.

## Model-based Reinforcement Learning With State Aggregation

*Cosmin Paduraru, Robert Kaplow, Doina Precup and Joelle Pineau*

We address the problem of model-based reinforcement learning in infinite state spaces. One of the simplest and most popular approaches is state aggregation: discretize the state space, build a transition model over the resulting aggregate states, then use this model to compute a policy. In this paper, we provide theoretical results that bound the performance of model-based reinforcement learning with state aggregation as a function of the number of samples used to learn the model and the quality of the discretization. To the best of our knowledge, these are the first sample complexity results for model-based reinforcement learning in continuous state spaces. We also investigate how our bounds compare with the empirical performance of the analyzed method.

# Session 10

July 2, Wednesday - 11:05

## Basis Expansion In Natural Actor Critic Methods

*Sertan Girgin and Philippe Preux*

In reinforcement learning, the aim of the agent is to find a policy that maximizes its expected return. Policy gradient methods try to accomplish this goal by directly approximating the policy using a parametric function approximator; the expected return of the current policy is estimated and its parameters are updated by steepest ascent in the direction of the gradient of the expected return with respect to the policy parameters. In general, the policy is defined in terms of a set of basis functions that capture important features of the problem. Since the quality of the resulting policies directly depend on the set of basis functions, and defining them gets harder as the complexity of the problem increases, it is important to be able to find them automatically. In this paper, we propose a new approach which uses cascade-correlation learning architecture for automatically constructing a set of basis functions within the context of Natural Actor-Critic (NAC) algorithms. Such basis functions allow more complex policies be represented, and consequently improve the performance of the resulting policies. We also present the effectiveness of the method empirically.

## Variable Metric Reinforcement Learning Methods Applied To The Noisy Mountain Car Problem

*Verena Heidrich-Meisner and Christian Igel*

Two variable metric reinforcement learning methods, the natural actor-critic algorithm and the covariance matrix adaptation evolution strategy, are compared on a conceptual level and analyzed experimentally on the mountain car benchmark task with and without noise.

## Policy Learning – A Unified Perspective With Applications In Robotics

*Jan Peters, Jens Kober and Duy Nguyen-Tuong*

Policy Learning approaches are among the best suited methods for high-dimensional, continuous control systems such as anthropomorphic robot arms and humanoid robots. In this paper, we show two contributions: firstly, we show a unified perspective which allows us to derive several policy learning algorithms from a common point of view, i.e, policy gradient algorithms, natural-gradient algorithms and EM-like policy learning. Secondly, we present several applications to both robot motor primitive learning as well as to robot control in task space. Results both from simulation and several different real robots are shown.

# Session 11

July 2, Wednesday - 14:15

## Exploiting Additive Structure In Factored MDPs For Reinforcement Learning

*Thomas Degris, Olivier Sigaud and Pierre-Henri Wuillemin*

SDYNA is a framework able to address large, discrete and stochastic reinforcement learning problems. It incrementally learns a FMDP representing the problem to solve while using fmdp planning techniques to build an efficient policy. SPITI, an instantiation of SDYNA, uses a planning method based on dynamic programming which cannot exploit the additive structure of a FMDP. In this paper, we present two new instantiations of SDYNA, namely ULP and UNATLP, using a linear programming based planning method that can exploit the additive structure of a FMDP and address problems out of reach of SPITI.

## Hierarchical Reinforcement Learning In Factored MDPs

*Olga Kozlova, Olivier Sigaud and Christophe Meyer*

In this paper, we present the *texdyna* algorithm designed to solve large Markov Decision Problems with unknown structure by integrating hierarchical abstraction techniques of Semi-Markov Decision Processes and factorization techniques of Factored Markov Decision Processes. We validate our approach on the taxi problem.

## Bayesian Reward Filtering

*Matthieu Geist, Olivier Pietquin and Gabriel Fricout*

A wide variety of function approximation schemes have been applied to reinforcement learning. However, Bayesian filtering approaches, which have been shown efficient in other fields such as neural network training, have been little studied. We propose a general Bayesian filtering framework for reinforcement learning, as well as a specific implementation based on sigma point Kalman filtering and kernel machines. This allows us to derive an efficient off-policy model-free approximate temporal differences algorithm which will be demonstrated on two simple benchmarks.

# Session 12

July 2, Wednesday - 15:45

## Transfer Of Samples In Batch Reinforcement Learning

*Alessandro Lazaric, Marcello Restelli and Andrea Bonarini*

The main objective of transfer in reinforcement learning is to reduce the complexity of learning the solution of a target task by effectively reusing the knowledge retained from solving a set of source tasks. In this paper, we introduce a novel algorithm that transfers samples (i.e., tuples  $(s, a, s', r)$ ) from source to target tasks. Under the assumption that tasks have similar transition models and reward functions, we propose a method to select samples from the source tasks that are mostly similar to the target task, and, then, to use them as input for batch reinforcement-learning algorithms. As a result, the number of samples an agent needs to collect from the target task to learn its solution is reduced. We empirically show that, following the proposed approach, the transfer of samples is effective in reducing the learning complexity, even when some source tasks are significantly different from the target task.

## Privacy-Preserving Reinforcement Learning

*Jun Sakuma, Shigenobu Kobayashi and Rebecca Wright*

We consider a problem of distributed reinforcement learning (DRL) from private perceptions. In our setting, agents' perceptions, such as states, rewards, and actions, are not only distributed but also are desired to be kept private. This can occur when agents' perceptions include private or confidential information. Conventional DRL algorithms could be applied to such problems, but do not necessarily guarantee privacy preservation. Additionally, DRL which learns only from local perceptions often sacrifice optimality. In this work, we design solutions that achieve optimal policies without requiring the agents to share their private information by means of well-known cryptographic tools, secure function evaluation.

## Multi-Agent Model-Based Reinforcement Learning Experiments In The Pursuit Evasion Game

*Bruno Bouzy and Marc Metivier*

This paper describes multi-agent learning experiments performed on tactical sequences of the pursuit evasion game on very small grids. Its aim is to underline the performance difference between a centralized approach, and a distributed approach when using Rmax, a model-based reinforcement learning algorithm. The prey's goal is to go out of the grid, and the predators' goal is to kill the prey. The prey may learn or not. The predators learn in two ways: in the centralized approach they are part of one single learning agent, and, in the distributed approach, each predator is a learning agent in itself. Every agents learn to accomplish its goal by using Rmax. Our results compare the centralized approach with the distributed approach. Future works mainly include scaling up to larger boards using model-free algorithms, and exploring partial observability of agents.

# Session 13

July 2, Wednesday - 17:15

## A Family Of Reinforcement Learning Algorithms

*Marco Wiering and Hado Van Hasselt*

This paper describes several new online model-free reinforcement learning (RL) algorithms. The aim is to compare these algorithms experimentally with existing algorithms, namely: Q-learning, Sarsa, Actor-Critic, QV-learning, and ACLA. We designed 4 new reinforcement algorithms, namely: QV2, QVMAX, QVMAX2, and Sarsa+. We show experiments on five maze problems of varying complexity, the first problem is simple, but the other four maze tasks are of a dynamic or partially observable nature. Furthermore, we show experimental results on the cart pole balancing problem.

## Empirical Bernstein Stopping

*Volodymir Mnih and Csaba Szepesvari*

Sampling is a popular way of scaling up machine learning algorithms to large datasets. The question often is how many samples are needed. Adaptive stopping algorithms monitor the performance in an online fashion and make it possible to stop early, sparing valuable computation time. We concentrate on the setting where probabilistic guarantees are desired and demonstrate how recently-introduced empirical Bernstein bounds can be used to design stopping rules that are efficient. We provide upper bounds on the sample complexity of the new rules as well as empirical results on model selection and boosting in the filtering setting. The results bear relevance to RL in that picking the winner can be considered as a prototypical action selection problem.

## Algorithms And Bounds For Sampling-based Approximate Policy Iteration

*Christos Dimitrakakis and Michail Lagoudakis*

Several approximate policy iteration schemes without value functions, which focus on policy representation using classifiers and address policy learning as a supervised learning problem, have been proposed recently. Finding good policies with such methods requires not only an appropriate classifier, but also reliable examples for the best actions, covering all of the state space. One major question is how to find a good covering efficiently. However, up to this time, little work has been done to reduce the sample complexity of such methods, especially in continuous state spaces. This paper focuses on the simplest possible classification strategy / policy representation for such spaces (a discretised grid) and performs a sample-complexity comparison between previously the simplest (and commonly) sample allocation strategy, which allocates samples equally at each state under consideration, and an almost as simple method, which is shown to require significantly fewer samples.

## Rollout Sampling Approximate Policy Iteration

*Christos Dimitrakakis and Michail Lagoudakis*

Several researchers have recently investigated the connection between reinforcement learning and classification. We are motivated by proposals of approximate policy iteration schemes without value functions which focus on policy representation using classifiers and address policy learning as a supervised learning problem. This paper proposes variants of an improved policy iteration scheme which addresses the core sampling problem in evaluating a policy through simulation as a multi-armed bandit machine. The resulting algorithm offers comparable performance to the previous algorithm achieved, however, with significantly less computational effort. An order of magnitude improvement is demonstrated experimentally in two standard reinforcement learning domains: inverted pendulum and mountain-car.

# Session 14

**July 3, Thursday - 10:00**

## Efficient Reinforcement Learning In Parameterized Models: Discrete Parameter Case.

*Kirill Dyagilev, Shie Mannor and Nahum Shimkin*

We consider reinforcement learning in the parameterized setup, where the model is known to belong to a parameterized family of Markov Decision Processes (MDPs). We further impose here the assumption that set of possible parameters is finite, and consider the discounted return. We propose an on-line algorithm for learning in such parameterized models, dubbed the Parameter Elimination (PEL) algorithm, and analyze its performance in terms of the total mistake bound criterion (also known as the sample complexity of exploration). The algorithm relies on Wald's Sequential Probability Ratio Test to eliminate unlikely parameters, and uses an optimistic policy for effective exploration. We establish that, with high probability, the total mistake bound for the algorithm is linear (up to a logarithmic term) in the size of the parameter space, independently of the cardinality of the state and action spaces.

## Robustness Analysis Of SARSA( $\lambda$ ): Different Models Of Reward And Initialisation

*Marek Grzes and Daniel Kudenko*

In the paper the robustness of SARSA( $\lambda$ ), the reinforcement learning algorithm with eligibility traces, is confronted with different models of reward and initialisation of the Q-table. Most of the empirical analyses of eligibility traces in the literature have focused mainly on the step-penalty reward. We analyse two general types of rewards (final goal and step-penalty rewards) and show that learning with long traces, i.e., with high values of  $\lambda$  can lead to suboptimal solutions in some situations. Problems are identified and discussed. Specifically, obtained results show that SARSA( $\lambda$ ) is sensitive to different models of reward and initialisation. In some cases the asymptotic performance can be significantly reduced.

# Session 15

July 3, Thursday - 11:05

## Lazy Planning Under Uncertainty By Optimizing Decisions On An Ensemble Of Incomplete Disturbance Trees

*Boris Defourny, Damien Ernst and Louis Wehenkel*

This paper addresses the problem of solving discrete-time optimal sequential decision making problems having a disturbance space  $W$  composed of a finite number of elements. In this context, the problem of finding from an initial state  $x(0)$  an optimal decision strategy can be stated as an optimization problem which aims at finding an optimal combination of decisions attached to the nodes of a disturbance tree modeling all possible sequences of disturbances  $w(0), w(1), \dots, w(T-1)$  in  $W^T$  over the optimization horizon  $T$ . A significant drawback of this approach is that the resulting optimization problem has a search space which is the Cartesian product of  $O(|W|^{T-1})$  decision spaces  $U$ , which makes the approach computationally impractical as soon as the optimization horizon grows, even if  $W$  has just a handful of elements. To circumvent this difficulty, we propose to exploit an ensemble of randomly generated incomplete disturbance trees of controlled complexity, to solve their induced optimization problems in parallel, and to combine their predictions at time  $t = 0$  to obtain a (near-)optimal first-stage decision. Because this approach postpones the determination of the decisions for subsequent stages until additional information about the realization of the uncertain process becomes available, we call it lazy. Simulations carried out on a robot corridor navigation problem show that even for small incomplete trees, this approach can lead to near-optimal decisions.

## Optimistic Planning Of Deterministic Systems

*Jean-Francois Hren and Remi Munos*

If one possesses a model of a controlled deterministic system, then from any state, one may consider the set of all possible reachable states starting from that state and using any sequence of actions. This forms a tree whose size is exponential in the planning time horizon. Here we ask the question: given a finite number of computational resources (e.g. CPU time), which may not be known ahead of time, what is the best way to explore this tree, such that once all resources have been used, the algorithm would be able to propose an action (or a sequence of actions) whose performance is as close as possible to optimality? The performance with respect to optimality is assessed in terms of the regret (with respect to the sum of discounted future rewards) resulting from choosing the action returned by the algorithm instead of an optimal action. In this paper we investigate an optimistic exploration of the tree, where the most promising states are explored first, and compare this approach to a naive uniform exploration. Bounds on the regret are derived both for uniform and optimistic exploration strategies. Numerical simulations illustrate the benefit of optimistic planning.

## Policy Optimization By Implicit Probabilistic Simulation

*Carl Rasmussen and Marc Deisenroth*

We consider learning to control a discrete-time dynamical system with continuous states and actions. The dynamics of the system are assumed to be unknown, and both the dynamics and the policy need to be learned. We utilize full Gaussian process models for both tasks. We demonstrate the viability of the approach on an illustrative toy application.

# Session 16

July 3, Thursday - 14:15

## Reinforcement Learning Of Perceptual Coupling For Motor Primitives

*Jens Kober and Jan Peters*

Reinforcement learning is a natural choice for the learning of complex motor tasks by reward-related self-improvement. As the space of movements is high-dimensional and continuous, a policy parametrization is needed which can be used in this context. Traditional motor primitive approaches deal largely with open-loop policies which can only deal with small perturbations. In this paper, we present a new type of motor primitive policies which serve as closed-loop policies together with an appropriate learning algorithm. Our new motor primitives are an augmented version of the dynamic systems motor primitives that incorporates perceptual coupling to external variables. We show that these motor primitives can perform complex tasks such as a Ball-in-a-Cup or Kendama task even with large variances in the initial conditions where a human would hardly be able to learn this task. We initialize the open-loop policies by imitation learning and the perceptual coupling with a handcrafted solution. We first improve the open-loop policies and subsequently the perceptual coupling using a novel reinforcement learning method which is particularly well-suited for motor primitives.

## Applications Of Reinforcement Learning To Structured Prediction

*Francis Maes, Ludovic Denoyer and Patrick Gallinari*

Supervised learning is about learning functions given a set of input and corresponding output examples. A recent trend in this field is to consider structured outputs such as sequences, trees or graphs. When predicting such structured data, learning models have to select solutions within very large discrete spaces. The combinatorial nature of this problem has recently led to learning models integrating a search component. In this paper, we show that Structured Prediction (SP) can be seen as a sequential decision problem. We introduce SP-MDP: a Markov Decision Process based formulation of Structured Prediction. Learning the optimal policy in SP-MDP is shown to be equivalent as solving the SP problem. This allows us to apply classical Reinforcement Learning (RL) algorithms to SP. We present experiments on two tasks. The first, sequence labeling, has been extensively studied and allows us to compare the RL approach with traditional SP methods. The second, tree transformation, is a challenging SP task with numerous large-scale real-world applications. We show successful results with general RL algorithms on this task on which traditional SP models fail.

## Policy Iteration For Learning An Exercise Policy For American Options

*Yuxi Li and Dale Schuurmans*

Options are important financial instruments, whose prices are usually determined by computational methods. Computational finance is a compelling application area for reinforcement learning research, where hard sequential decision making problems abound and have great practical significance. In this paper, we investigate reinforcement learning methods, in particular, least squares policy iteration (LSPI), for the problem of learning an exercise policy for American options. We also investigate TVR, another policy iteration method. We compare LSPI, TVR with LSM, the standard least squares Monte Carlo method from the finance community. We evaluate their performance on both real and synthetic data. The results show that the exercise policies discovered by LSPI and TVR gain larger payoffs than those discovered by LSM, on both real and synthetic data. Furthermore, for LSPI, TVR and LSM, policies learned from real data generally gain larger payoffs than policies learned from simulated samples. Our work shows that solution methods developed in reinforcement learning can advance the state of the art in an important and challenging application area, and demonstrates furthermore that computational finance remains an under-explored area for deployment of reinforcement learning methods.

# Session 17

July 3, Thursday - 15:45

## Adaptive Aggregation For Reinforcement Learning With Efficient Exploration: Deterministic Domains

*Andrey Bernstein and Nahum Shimkin*

We propose a model-based learning algorithm, the Adaptive Aggregation Algorithm (AAA), that aims to solve the online, continuous state space reinforcement learning problem in a deterministic domain. The proposed algorithm uses an adaptive state aggregation approach, going from coarse to fine grids over the state space, which enables to use finer resolution in the "important" areas of the state space, and coarser resolution elsewhere. We consider an on-line learning approach, in which we discover these important areas on-line, using an uncertainty intervals exploration technique. Polynomial learning rates in terms of mistake bound (in a PAC framework) are presented for this algorithm, under appropriate continuity assumptions.

## Approximate Policy Iteration For Generalized Semi-Markov Decision Processes: An Improved Algorithm

*Emmanuel Rachelson, Patrick Fabiani and Frédéric Garcia*

In the context of time-dependent problems of planning under uncertainty, most of the problem's complexity comes from the concurrent interaction of simultaneous processes. Generalized Semi-Markov Decision Processes represent an efficient formalism to capture both concurrency of events and actions and uncertainty. We introduce GSMDP with observable time and hybrid state space and present a new algorithm based on Approximate Policy Iteration to generate efficient policies. This algorithm relies on simulation-based exploration and makes use of SVM regression. We experimentally illustrate the strengths and weaknesses of this algorithm and propose an improved version based on the weaknesses highlighted by the experiments.

## Markov Decision Processes With Arbitrary Reward Processes

*Jia Yuan Yu, Shie Mannor and Nahum Shimkin*

We consider a control problem where the decision maker interacts with a standard Markov decision process with the exception that the reward functions vary arbitrarily over time. We extend the notion of Hannan consistency to this setting, showing that, in hindsight, the agent can perform almost as well as every deterministic policy. We present efficient online algorithms in the spirit of reinforcement learning that ensure that the agent's performance loss, or regret, vanishes over time, provided that the environment is oblivious to the agent's actions. However, counterexamples indicate that the regret does not vanish if the environment is not oblivious.

# Session 18

July 3, Thursday - 17:15

## Relational TD Reinforcement Learning

*Christophe Rodrigues, Pierre Gérard and Celine Rouveirol*

Relational Reinforcement Learning (RRL) addresses the use of relational representations of states and actions in RL rather than the usual attribute-values. Most works in this field aims at improving relational function approximation, or at adapting advanced techniques to the relational framework. However, little has been done so far to investigate basic Temporal Difference in RRL. In this paper, we propose adaptations of Sarsa and regular Q-learning to the relational case, by using an incremental relational function approximator RIB. In the experimental study, we highlight how changing the RL algorithms impacts generalisation in relational regression.

## Reinforcement Learning With Markov Logic Networks

*Weiwei Wang, Xingguo Chen and Yang Gao*

In this paper, we propose a method to combine reinforcement learning and markov logic network which can easily introduce priori knowledge with the weights of first-order formulas, compactly represent state and learn weight efficiently. Most methods in RL are tabular methods, and thus they lack the ability to handle high-dimension problems. Even with function approximation, we often take no account of the inherent relations or connections of the features, otherwise we need to introduce additional features to represent such connections. Markov logic networks(MLN) combines first-order logic and graphical model and it has the ability to compactly represent a wide variety of knowledge. Introducing MLN to RL will bring us a new method to deal with many difficult problems in RL which need some relational representation of state, such as blocks world. Our new reinforcement learning algorithm with Markov logic networks(RMLN) brings a solution to this kind of problems. In our framework, MLN does inference for the action queries and selects a best action, RL uses the successive state, current state and the reward to update the weights of formulas in MLN. With RMLN, priori knowledge can be easily introduced to the learning system and the learning process will become more efficient. Experiment on blocks world illustrates the promise of RMLN.

## Classifier-Based Policy Representation

*Michail Lagoudakis and Ioannis Rexakis*

Motivated by recent proposals that view a reinforcement learning problem as a collection of classification problems, we investigate various aspect of policy representations using classifiers. In particular, we derive optimal policies for two standard reinforcement learning domains (inverted pendulum and mountain car) in both deterministic and stochastic versions and we examine their internal structure. We then proceed in an evaluation of the representational ability of a variety of classifiers for these policies. Apart from the original formulation where the entire policy is represented by a single multi-class classifier, we also examine and suggest a better formulation whereby the policy is represented by a collection of binary classifiers, one for each action. We strongly believe that our results offers significant insight in making the “reinforcement learning via classification” technology successfully applicable to real-world learning problems.

## Author Index

- Antanas, Laura-Andreea, 5  
 Avoli, Massimo, 7
- Bernstein, Andrey, 17  
 Bertsekas, Dimitri, 9  
 Bonarini, Andrea, 12  
 Bouzy, Bruno, 12  
 Broda, Krysia, 4
- Cappé, Olivier, 3  
 Chaslot, Guillaume, 2  
 Chatriot, Louis, 2  
 Chen, Xingguo, 7, 18  
 Clérot, Fabrice, 3  
 Croonenborghs, Tom, 5
- Defourny, Boris, 15  
 Degris, Thomas, 11  
 Deisenroth, Marc, 15  
 Denoyer, Ludovic, 16  
 Dickens, Luke, 4  
 Dimitrakakis, Christos, 13  
 Driessens, Kurt, 5  
 Dyagilev, Kirill, 14
- Ernst, Damien, 15
- Fabiani, Patrick, 17  
 Farahmand, Amir Massoud, 1  
 Filippi, Sarah, 3  
 Frank, Jordan, 8  
 Fricout, Gabriel, 11
- Gérard, Pierre, 18  
 Gómez-Sanchis, Juan, 7  
 Gabel, Thomas, 4  
 Gallinari, Patrick, 16  
 Gao, Yang, 7, 18  
 Garcia, Frédéric, 17  
 Garivier, Aurélien, 6  
 Geist, Matthieu, 11  
 Gelly, Sylvain, 2  
 Ghavamzadeh, Mohammad, 1  
 Girgin, Sertan, 10  
 Goetschalckx, Robby, 5  
 Grzes, Marek, 5, 14  
 Guez, Arthur, 7
- Hasselt, Hado Van, 13  
 Heidrich-Meisner, Verena, 10  
 Herik, Jaap Van Den, 2  
 Hren, Jean-Francois, 15
- Igel, Christian, 10
- Jean-Baptiste, Hoock, 2
- Kaplow, Robert, 9  
 Kobayashi, Shigenobu, 12  
 Kober, Jens, 10, 16  
 Kozlova, Olga, 11  
 Kudenko, Daniel, 5, 14
- Lagoudakis, Michail, 13, 18  
 Lanzi, Pier Luca Lanzi Pier Luca, 5  
 Lazaric, Alessandro, 12  
 Li, Lihong, 8  
 Li, Yuxi, 16  
 Littman, Michael, 8  
 Loiacono, Daniele, 5  
 Lorincz, Andras, 6  
 Loth, Manuel, 6
- Maes, Francis, 16  
 Magdalena, Rafael, 7  
 Mannor, Shie, 1, 8, 14, 17  
 Martínez, Marcelino, 7  
 Martin-Guerrero, Jose D., 7  
 Metivier, Marc, 12  
 Meyer, Christophe, 11  
 Mnih, Volodymir, 13  
 Moulines, Eric, 3, 6  
 Mueller, Martin, 2  
 Munos, Remi, 15
- Nguyen-Tuong, Duy, 10
- Paduraru, Cosmin, 9  
 Panangaden, Prakash, 8  
 Perez, Julien, 2  
 Peters, Jan, 10, 16  
 Pietquin, Olivier, 11  
 Pineau, Joelle, 7, 9  
 Precup, Doina, 8, 9  
 Preux, Philippe, 6, 10
- Rachelson, Emmanuel, 17  
 Ramon, Jan, 5  
 Rasmussen, Carl, 15  
 Restelli, Marcello, 12  
 Rexakis, Ioannis, 18  
 Riedmiller, Martin, 3, 4  
 Rimmel, Arpad, 2  
 Rodrigues, Christophe, 18  
 Rouveirol, Celine, 18  
 Russo, Alessandra, 4
- Sakuma, Jun, 12  
 Sanner, Scott, 5  
 Schuurmans, Dale, 16  
 Serrano, Antonio José, 7  
 Shimkin, Nahum, 14, 17  
 Sigaud, Olivier, 11  
 Silver, David, 2

Soria, Emilio, 7  
Sutton, Rich, 2  
Szepesvari, Csaba, 1, 13  
Szita, Istvan, 2, 6

Taylor, Jonathan, 8  
Teytaud, Olivier, 2  
Timmer, Stephan, 3

Vincent, Robert, 7

Walsh, Thomas, 8  
Wang, Weiwei, 7, 18  
Wehenkel, Louis, 15  
Welsh, Noel, 3  
Wiering, Marco, 13  
Winands, Mark, 2  
Wright, Rebecca, 12  
Wuillemin, Pierre-Henri, 11  
Wyatt, Jeremy, 3

Yu, Huizhen, 9  
Yu, Jia Yuan, 17