Report from Dagstuhl Seminar 13321

Reinforcement Learning

Edited by

Peter Auer[1], Marcus Hutter[2], and Laurent Orseau[3]

1 **Montanuniversität Leoben, AT,** auer@unileoben.ac.at
2 **Australian National University – Canberra, AU,** marcus.hutter@anu.edu.au
3 **AgroParisTech – Paris, FR,** laurent.orseau@agroparistech.fr

Abstract

This Dagstuhl Seminar also stood as the 11th European Workshop on Reinforcement Learning (EWRL11). Reinforcement learning gains more and more attention each year, as can be seen at the various conferences (ECML, ICML, IJCAI, ...). EWRL, and in particular this Dagstuhl Seminar, aimed at gathering people interested in reinforcement learning from all around the globe. This unusual format for EWRL helped viewing the field and discussing topics differently.

Seminar 04.–09. August, 2013 – www.dagstuhl.de/13321
1998 ACM Subject Classification I.2.6 Learning, I.2.8 Problem Solving, Control Methods, and Search, I.2.9 Robotics, I.2.11 Distributed Artificial Intelligence, G.3 Probability and Statistics (Markov processes)
Keywords and phrases Machine Learning, Reinforcement Learning, Markov Decision Processes, Planning
Digital Object Identifier 10.4230/DagRep.3.8.1

1 Executive Summary

Peter Auer
Marcus Hutter
Laurent Orseau

Reinforcement Learning (RL) is becoming a very active field of machine learning, and this Dagstuhl Seminar aimed at helping researchers have a broad view of the current state of this field, exchange cross-topic ideas and present and discuss new trends in RL. It gathered 38 researchers together. Each day was more or less dedicated to one or a few topics, including in particular: The exploration/exploitation dilemma, function approximation and policy search, universal RL, partially observable Markov decision processes (POMDP), inverse RL and multi-objective RL.This year, by contrast to previous EWRL events, several small tutorials and overviews were presented. It appeared that researchers are nowadays interested in bringing RL to more general and more realistic settings, in particular by alleviating the Markovian assumption, for example so as to be applicable to robots and to a broader class of industrial applications.This trend is consistent with the observed growth of interest in policy search and universal RL. It may also explain why the traditional treatment of the exploration/exploitation dilemma received less attention than expected.

Reinforcement Learning, *Dagstuhl Reports*, Vol. 3, Issue 8, pp. 1–26
Editors: Peter Auer, Marcus Hutter, and Laurent Orseau
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Table of Contents

3 Overview of Talks

3.1 Preference-based Evolutionary Direct Policy Search

Robert Busa-Fekete (Universität Marburg, DE)

We introduce a preference-based extension of *evolutionary direct policy search* (EDPS) as proposed by Heidrich-Meisner and Igel [2]. EDPS casts policy learning as a search problem in a parametric policy space, where the function to be optimized is a performance measure like expected total reward, andevolution strategies (ES) such as CMA-ES [1] are used as optimizers. Moreover, since the evaluation of a policy can only be done approximately, namely in terms of a finite number of *rollouts*, the authors make use of *racing algorithms* [3] to control this number in an adaptive manner. These algorithms return a sufficiently reliable ranking over the current set of policies (candidate solutions), which is then used by the ES for updating its parameters and population. A key idea of our approach is to extend EDPS by replacing the *value-based* racing algorithm with a *preference-based* one that operates on a suitable ordinal preference structure and only uses pairwise comparisons between sample rollouts of the policies.

References

1 N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 282–291, 2004.
2 V. Heidrich-Meisner and C. Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *Proceedings of the 26th International Conference on Machine Learning*, pages 401–408, 2009.
3 O. Maron and A.W. Moore. Hoeffding races: accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems*, pages 59–66, 1994.

3.2 Solving Simulator-Defined MDPs for Natural Resource Management

Thomas G. Dietterich (Oregon State University, US)

Joint work of Dietterich, Thomas G.; Taleghan Alkaee, Majid; Crowley, Mark
Main reference T. G. Dietterich, M. Alkaee Taleghan, M. Crowley, "PAC Optimal Planning for Invasive Species Management: Improved Exploration for Reinforcement Learning from Simulator-Defined MDPs," in Proc. of the AAAI Conference on Artificial Intelligence (AAAI'13), AAAI Press. 2013.
URL http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6478
URL http://web.engr.oregonstate.edu/~tgd/publications/dietterich-taleghan-crowley-pac-optimal-planning-for-invasive-species-management-etc-aaai2013.pdf

Natural resource management problems, such as forestry, fisheries, and water resources, can be formulated as Markov decisionprocesses. However, solving them is difficult for two reasons. First, the dynamics of the system are typically available only in the form of a complex and expensive simulator. This means that MDP planning algorithms are needed that minimize the number of calls to the simulator. Second, the systems are spatial. A natural way to formulate the MDP is to divide up the region into cells, where each cell is

modeled with a small number of state variables. Actions typically operate at the level of individual cells, but the spatial dynamics couple the states of spatially-adjacent cells. The resulting state and action spaces of these MDPs are immense. We have been working on two natural resource MDPs. The first involves the spread of tamarisk in river networks. A native of the Middle East, tamarisk has become an invasive plant in the dryland rivers and streams of the western US. Given a tamarisk invasion, a land manager must decide how and where to fight the invasion (e.g., eradicate tamarisk plants? plant native plants? upstream? downstream?). Although approximate or heuristic solutions to this problem would beuseful, our collaborating economists tell us that our policy recommendations will carry more weight if they are provably optimal with high probability. A large problem instance involves 4.7×10^6 states with 2187 actions in each state. On a modern 64-bit machine, the action-value function for this problem can fit into main memory. However, computing the full transition function to sufficient accuracy to support standardvalue iteration requires on the order of 3×10^{20} simulator calls. The second problem concerns the management of wildfire in Eastern Oregon. In this region, prior to European settlement, the native ponderosa pine forests were adapted to frequent, low-intensity fires. These fires allow the ponderosa pine trees (which are well-adapted to survive fire) to grow very tall while preventing the accumulation of fuel at ground level. These trees provide habitat for many animal species and are also very valuable for timber. However, beginning in the early 1900s, all fires were suppressed in this landscape, which has led to the build up of huge amounts of fuel. The result has been large, catastrophic fires that kill even the ponderosa trees and that are exceptionally expensive to control. The goal of fire management is to return the landscape to a state where frequent, low-intensity fires are again the normal behavior. There are two concrete fire management problems: Let Burn (decide which fires to suppress) and Fuel Treatment (decide in which cells to perform fuel reduction treatments). Note that in these problems, the system begins in an unusual, non-equilibrium state, and the goal is to return the system to a desired steady state distribution. Hence, these problems are not problems of reinforcement learning, but rather problems of MDP planning for a specific start state. Many of the assumptions in RL papers, such as ergocity of all policies, are not appropriate for this setting. Note also that it is highly desirable to produce a concrete policy (as opposed to just producing near-optimal behavior viareceding horizon control). A concrete policy can be inspected by stakeholders to identify missing constraints, state variables, and components of the reward function. To solve these problems, we are exploring two lines of research. For tamarisk, we have been building on recent work in PAC-RL algorithms(e.g., MBIE, UCRL, UCRL2, FRTDP, OP) to develop PAC-MDP planning algorithms. We are pursuing two innovations. First, we have developed an exploration heuristic based on an upper bound on the discounted state occupancy probability. Second, we are developing tighter confidence intervals in order to terminate the search earlier. These are based on combining Good-Turing estimates of missing mass (i.e., for unseen outcomes) with sequential confidence intervals for multinomial distributions. These reduce the degree to which we must rely on the union bound, and hence give us tighter convergence. For the LetBurn wildfire problem, we are exploring approximate policy iteration methods. For Fuel Treatment, we are extending Crowley's Equilibrium Policy Gradient methods. These define a local policy function that stochastically chooses the action for cell i based on the actions already chosen for the cells in the surrounding neighborhood. A Gibbs-sampling-style MCMC method repeatedly samples from these local policies until a global equilibrium is reached. This equilibrium defines the global policy. At equilibrium, gradient estimates can be computed and applied to improve the policy.

3.3 ABC and Cover Tree Reinforcement Learning

Christos Dimitrakakis (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Tziortziotis, Nikolaos
Main reference C. Dimitrakakis, N. Tziortziotis, "ABC Reinforcement Learning," arXiv:1303.6977v4 [stat.ML], 2013; to appear in Proc. of ICML'13.
URL http://arxiv.org/abs/1303.6977v4

In this talk, I presented our recent results on methods for Bayesian reinforcement learning using Thompson sampling, but differing significantly on their prior. The first, Approximate Bayesian Computation Reinforcement Learning (ABC-RL), employs an arbitrary prior over a set of simulators and is most suitable in cases where an uncertain simulation model is available. The second,Cover Tree Bayesian Reinforcement Learning (CTB-RL), performs closed-form online Bayesian inference on a cover tree and is suitable for arbitrary reinforcement learning problems, when little is known about the environment and fast inference is essential. ABC-RL introduces a simple, general framework for likelihood-free Bayesian reinforcement learning, through Approximate Bayesian Computation (ABC). The advantage is that we only require a prior distribution ona class of simulators. This is useful when a probabilistic model of the underlying process is too complex to formulate, but where detailed simulation models are available. ABC-RL allows the use of any Bayesian reinforcement learning technique in this case. It can be seen as an extension of simulation methods to both planning and inference. We experimentally demonstrate the potential of this approach in a comparison with LSPI. Finally, we introduce a theorem showing that ABC is sound, in the sense that the KL divergence between the incomputable true posterior and the ABC approximation is bounded by an appropriate choice of statistics. CTBRL proposes an online tree-based Bayesian approach for reinforcement learning. For inference, we employ a generalised context tree model. This defines a distribution on multivariate Gaussian piecewise-linear models, which can be updated in closed form. The tree structure itself is constructed using the cover tree method, which remains efficient in high dimensional spaces. We combine the model with Thompson sampling and approximate dynamic programming to obtain effective exploration policies in unknown environments. The flexibility and computational simplicity of the model render it suitable for many reinforcement learning problems in continuous state spaces. We demonstrate this in an experimental comparison with least squares policy iteration.

References

1 N. Tziortziotis, C. Dimitrakakis and K. BlekasCover Tree Bayesian Reinforcement Learning. arXiv:1305.1809

3.4 Some thoughts on Transfer Learning in Reinforcement Learning: on States and Representation

Lutz Frommberger (Universität Bremen, DE)

Main reference L. Frommberger, "Qualitative Spatial Abstraction in Reinforcement Learning," Cognitive Technologies Series, ISBN 978-3-642-16590-0, Springer, 2010.
URL http://dx.doi.org/10.1007/978-3-642-16590-0

The term "transfer learning" is a fairly sophisticated term for something thatcan be considered a core component of any learning effort of a human or animal: to base the solution to a new problem on experience and learning success of prior learning tasks. This is something that a

learning organism does implicitly from birth on: no task is ever isolated, but embedded in a common surrounding or history. In contrary to this lifelong learning type setting, transfer learning in RL[5] assumes two different MDPs $\mathcal{M}$ and $\mathcal{M}'$ that have something "in common". This commonality is mostlikely given in a task mapping function that maps states and actions from $\mathcal{M}$ to $\mathcal{M}'$ as a basis for reusing learned policies. Task mappings can be given by human supervisors or learned, but mostly there is some instance telling the learning agent what to do to benefit from its experience. In very common words: Here is task $\mathcal{M}$, there is task$\mathcal{M}'$, and this is how you can bridge between them. This is a fairly narrow view on information reuse, and more organic and autonomous variants of knowledge transfer are desirable. Knowledge transfer, may it be in-task (i.e., generalization) or cross-task, exploits similarity between tasks. By task mapping functions, information on similarity is brought into the learning process from outside. This also holds for approaches that do not require an explicit state mapping [2, 4], where relations or agent spaces, e.g., are defined a-priori. What is mostly lacking so far is the agent's ability to recognize similarities on its own and/or seamlessly benefit from prior experiences as an integral part of the new learning effort. An intelligent learning agent should easily notice if certain parts of the current task are identical or similar to an earlier learning task, for example, general movement skills that remain constant over many specialized learning tasks. In prior work, I proposed generalization approaches such as task space tilecoding [1] that allow to reuse knowledge of the actual learning task if certain state variables are identical. This works if structural information is made part of the state space and does not require a mapping function. However, it needs a-priori knowledge of which state variables are critical for action selection in a structural way. Recent approaches foster thehope that such knowledge can be retrieved by the agent itself: e.g.,[3] allows for identification of state variables that have a generally high impact on action selection over one or several tasks. But even if we can identify and exploit certain state variables that encode structural information and have this generalizing impact, these features must at least exist. If they do not exist in the state representations, such approaches fail. For example, if the distance to the next obstacle in front of an agent is this critical value, it does not help if the agent's position and the position of the obstacle are in the state representation, as this implicitly hides the critical information. Thus, again, the question of state space design becomes evident. How can we ensure that relevant information is encoded on the level of features? Or how can we exploit information that is only implicitly given in the state representation? Answering these questions will be necessary to take the next step into autonomous knowledge reuse for RL agents.

References

1 Lutz Frommberger. Task space tile coding: In-task and cross-task generalization in reinforcement learning. In *9th European Workshop on Reinforcement Learning (EWRL9)*, Athens, Greece, September 2011.

2 George D. Konidaris and Andrew G. Barto. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the Twenty Third International Conference on Machine Learning (ICML 2006)*, pp. 489–49, Pittsburgh, PA, June 2006.

3 Matthijs Snel and Shimon Whiteson. Multi-task reinforcement learning: shaping and feature selection. In *Recent Advances in Reinforcement Learning*, pp. 237–248. Springer, 2012.

4 Prasad Tadepalli, Robert Givan, and Kurt Driessens. Relational reinforcement learning: An overview. In *Proc. of the ICML-2004 Workshop on Relational Reinforcement Learning*, pp. 1–9, 2004.

5 Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685,2009.

3.5 Actor-Critic Algorithms for Risk-Sensitive MDPs

Mohammad Ghavamzadeh (INRIA Nord Europe – Lille, FR)

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both discounted and average reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criteria, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

3.6 Statistical Learning Theory in Reinforcement Learning and Approximate Dynamic Programming

Mohammad Ghavamzadeh (INRIA Nord Europe – Lille, FR)

Approximate dynamic programming (ADP) and reinforcement learning (RL) algorithms are used to solve sequential decision-making tasks where the environment (i.e., the dynamics and the rewards) is not completely known and/or the size of thestate and action spaces is too large. In these scenarios, the convergence and performance guarantees of the standard DP algorithms are no longer valid, and adifferent theoretical analysis have to be developed. Statistical learning theory (SLT) has been a fundamental theoretical tool to explain the interaction between the process generating the samples and the hypothesis space used by learning algorithms, and shown when and how-well classification and regression problems can be solved. In recent years, SLT tools have been used to study the performance of batch versions of RL and ADP algorithms with the objective of deriving finite-sample bounds on the performance loss (w.r.t. the optimal policy) of the policy learned by these methods. Such an objective requires to effectively combine SLT tools with the ADP algorithms, and to show how the error is propagated through the iterations of these iterative algorithms.

3.7 Universal Reinforcement Learning

Marcus Hutter (Australian National University, AU)

There is great interest in understanding and constructing generally intelligent systems approaching and ultimately exceeding human intelligence. Universal AI is such a mathematical

theory of machinesuper-intelligence. More precisely, AIXI is an elegant parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. The theory reduces all conceptual AI problems to pure computational questions. After a brief discussion of its philosophical, mathematical, and computational ingredients, I will give a formal definition and measure of intelligence, which is maximized by AIXI. AIXI can be viewed as the most powerful Bayes-optimal sequential decision maker, for which I will present general optimality results. This also motivates some variations such as knowledge-seeking and optimistic agents, and feature reinforcement learning. Finally I present some recent approximations, implementations, and applications of this modern top-down approach to AI.

References

1 M. Hutter. *Universal Artificial Intelligence.* Springer, Berlin, 2005.
2 J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. *A Monte Carlo AIXI approximation. Journal of Artificial Intelligence Research*, 40:95–142, 2011.
3 S. Legg. *Machine Super Intelligence. PhD thesis*, IDSIA, Lugano, Switzerland, 2008.
4 M. Hutter. *One decade of universal artificial intelligence.* In *Theoretical Foundations of Artificial General Intelligence*, pages 67–88. Atlantis Press, 2012.
5 T. Lattimore. *Theory of General Reinforcement Learning. PhD thesis*, Research School of Computer Science, Australian National University, 2014.

3.8 Temporal Abstraction by Sparsifying Proximity Statistics

Rico Jonschkowski (TU Berlin, DE)

Joint work of Jonschkowski, Rico; Toussaint, Marc;

Automatic discovery of temporal abstractions is a key problem in hierarchical reinforcement learning. In previous work, such abstractions were found through the analysis of a set of experienced or demonstrated trajectories [2]. We propose that, from such trajectory data, we may learn temporally marginalized transition probabilities—which we call proximity statistics—that model possible transitions on larger time scales rather than learning 1-step transition probabilities. Viewing the proximity statistics as state values allows the agent to generate greedy policies from them. Making the statistics sparse and combining proximity estimates by *proximity propagation* can substantially accelerate planning compared to value iteration while keeping the size of the statistics manageable. The concept of proximity statistics and its sparsification approach is inspired from recent work in transit-node routing in large road networks [1]. We show that sparsification of these proximity statistics implies an approach to the discovery of temporal abstractions. Options defined by subgoals are shown to be a special case of the sparsification of proximity statistics. We demonstrate the approach and compare various sparsification schemes in a stochastic grid world.

References

1 Holger Bast; Setfan Funke; Domagoj Matijevic; Peter Sanders; Dominik Schultes, *In transit to constant time shortest-path queries in road networks*, Workshop on Algorithm Engineering and Experiments, 46–59, 2007.
2 Martin Stolle; Doina Precup, *Learning options in reinforcement learning*, Proc. of the 5th Int'l Symp. on Abstraction, Reformulation and Approximation, pp. 212–223, 2002.

3.9 State Representation Learning in Robotics

Rico Jonschkowski (TU Berlin, DE)

Joint work of Jonschkowski, Rico; Brock, Oliver
Main reference R. Jonschkowski, O. Brock, "Learning Task-Specific State Representations by Maximizing Slowness and Predictability," in Proc. of the 6th Int'l Workshop on Evolutionary and Reinforcement Learning for Autonomous Robot Systems (ERLARS'13), 2013.
URL http://www.robotics.tu-berlin.de/fileadmin/fg170/Publikationen_pdf/Jonschkowski-13-ERLARS-final.pdf

The success of reinforcement learning in robotic tasks is highly dependent onthe state representation – a mapping from high dimensional sensory observations of the robot to states that can be used for reinforcement learning. Currently, this representation is defined by human engineers – thereby solving an essential part of the robotic learning task. However, this approach does not scale because it restricts the robot to tasks for which representations have been predefined by humans. For robots that are able to learn new tasks, we need state representation learning. We sketch how this problem can be approached by iteratively learning a hierarchy of task-specific state representations following a curriculum. We then focus on a single step in this iterative procedure: learning a state representation that allows the robot to solve asingle task. To find this representation, we optimize two characteristics of good state representations: predictability and slowness. We implement these characteristics in a neural network and show that this approach can find good state representations from visual input in simulated robotic tasks.

3.10 Reinforcement Learning with Heterogeneous Policy Representations

Petar Kormushev (Istituto Italiano di Tecnologia – Genova, IT)

Joint work of Kormushev, Petar; Caldwell, Darwin G.
Main reference P. Kormushev, D. G. Caldwell, "Reinforcement Learning with Heterogeneous Policy Representations," 2013.
URL http://ewrl.files.wordpress.com/2013/06/ewrl11_submission_20.pdf

We propose a novel reinforcement learning approach for direct policy search that can simultaneously: (i) determine the most suitable policy representation for a given task; and (ii) optimize the policy parameters of this representation in order to maximize the reward and thus achieve the task. The approach assumes that there is a heterogeneous set of policy representations available to choose from. A naïve approach to solving this problem would be to take the available policy representations one by one, run a separate RL optimization process (i.e. conduct trials and evaluate the return) for each once, and at the very end pick the representation that achieved the highest reward. Such an approach, while theoretically possible, would not be efficient enough in practice. Instead, our proposed approach is to conduct one single RL optimization process while interleaving simultaneously all available policy representations. This can be achieved by leveraging our previous work in the area of RL based on Particle Filtering (RLPF).

3.11 Theoretical Analysis of Planning with Options

Timothy Mann (Technion – Haifa, IL)

Joint work of Mann, Timothy; Mannor, Shie

We introduce theoretical analysis suggesting how planning can benefit from using options. Experimental results have shown that options often induce faster convergence [3, 4], and previous theoretical analysis has shown that options are well-behaved in dynamic programming[2, 3]. We introduced a generalization of the Fitted Value Iteration (FVI) algorithm [1] that incorporates samples generated by options. Our analysis reveals that when the given set of options contains the primitive actions, our generalized algorithm converges approximately as fast as FVI with only primitive actions. When only temporally extended actions are used for planning convergence can be significantly faster than planning with only primitives, but this method may converge toward a suboptimal policy. We also developed precise conditions where our generalized FVI algorithm converges faster with a combination of primitive and temporally extended actions than with only primitive actions. These conditions turn out to depend critically on whether the iterates produced by FVI underestimate the optimal value function. Our analysis of FVI suggests that options can play an important role in planning by inducing fast convergence.

References

1 Rémi Munos and Csaba Szepesvári.Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.
2 Doina Precup, Richard S Sutton, and Satinder Singh. Theoretical results on reinforcement learning with temporallyabstract options. Machine Learning: ECML-98, Springer, 382–393, 1998.
3 David Silver and Kamil Ciosek. Compositional Planning Using Optimal Option Models.In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012.
4 Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112 (1):181–211, August 1999.

3.12 Learning Skill Templates for Parameterized Tasks

Jan Hendrik Metzen (Universität Bremen, DE)

Joint work of Metzen, Jan Hendrik; Fabisch, Alexander

We consider the problem of learning skills for a parameterized reinforcement learning problem class. That is, we assume that a task is defined by a task parameter vector and, likewise, a skill is considered as a parameterized policy. We propose skill templates, which allow to generalize skills that have been learned using reinforcement learning to similar tasks. In contrast to the recently proposed parameterized skills[1], skill templates also provide a measure of uncertainty for this generalization, which is useful for subsequent adaptation of the skill by means of reinforcement learning. In order to infer a generalized mapping from task parameter space to policy parameter space and an estimate of its uncertainty, we use Gaussian process regression [3]. We represent skills by dynamical movement primitives [2]

and evaluate the approach on a simulated Mitsubishi PA10 arm, where learning a single skill corresponds to throwing a ball to a fixed target position while learning the skill template requires to generalize to new target positions. We show that learning skill templates requires only a small amount of training data and improves learning in the target task considerably.

References

1 B. C. da Silva, G. Konidaris, and A. G. Barto. *Learning Parameterized Skills*, 29th International Conference on Machine Learning, 2012.

2 A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. *Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors*. Neural Computation, 25:2, 328–373, 2013.

3 C. E. Rasmussen, C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

3.13 Online learning in Markov decision processes

Gergely Neu (Budapest University of Technology & Economics, HU)

Joint work of Neu, Gergely; Szepesvari, Csaba; Zimin, Alexander; Gyorgy, Andras; Dick, Travis
Main reference A. Zimin, G. Neu, "Online learning in Markov decision processes by Relative Entropy Policy Search," to appear in Advances in Neural Information Processing 26.

We study the problem of online learning in finite episodic Markov decision processes (MDPs) where the loss function is allowed to change between episodes. The natural performance measure in this learning problem is the regret defined as the difference between the total loss of the best stationary policy and the total loss suffered by the learner. We assume that the learner is given access to a finite action space $\mathcal{A}$ and the state space $\mathcal{X}$ has a layered structure with L layers, so that state transitions are only possible between consecutive layers. We propose several learning algorithms based on applying the well-known Mirror Descent algorithm to the problem described above. For deriving our first method, we observe that Mirror Descent can be regarded as a variant of the recently proposed Relative Entropy Policy Search (REPS) algorithm of [1]. Our corresponding algorithm is called Online REPS or O-REPS. Second, we show how to approximately solve the projection operations required by Mirror Descent without taking advantage of the connections to REPS. Finally, we propose a learning method based on using a modified version of the algorithm of[2] to implement the Continuous Exponential Weights algorithm for the online MDP problem. For these last two techniques, we provide rigorous complexity analyses. More importantly, we show that all of the above algorithms satisfy regret bounds of $O(\sqrt{L\left|\mathcal{X}\right|\left|\mathcal{A}\right|T\log(\left|\mathcal{X}\right|\left|\mathcal{A}\right|/L)})$ in the bandit setting and $O(L\sqrt{T\log(\left|\mathcal{X}\right|\left|\mathcal{A}\right|/L)})$ in thefull information setting (both after T episodes). These guarantees largely improve previously known results under much milder assumptions and cannot be significantly improved under general assumptions.

References

1 Peters, J., Mülling, K., and Altün, Y. (2010). Relative entropy policy search. In *Proc. of the 24th AAAI Conf. on Artificial Intelligence (AAAI 2010)*, pp. 1607–1612.

2 Narayanan, H. and Rakhlin, A. (2010). Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems 23*, pp. 1777–1785.

3.14 Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search

Gerhard Neumann (TU Darmstadt – Darmstadt, DE)

Joint work of Neumann, Gerhard; Daniel, Christian; Kupcsik, Andras; Deisenroth, Marc; Peters, Jan
URL G. Neumann, C. Daniel, A. Kupcsik, M. Deisenroth, J. Peters, "Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search," 2013.
URL http://ewrl.files.wordpress.com/2013/06/ewrl11_submission_1.pdf

The key idea behind information-theoretic policy search is to bound the 'distance' between the new and old trajectory distribution, where the relative entropy is used as 'distance measure'. The relative entropy bound exhibits many beneficial properties, such as a smooth and fast learning process and a closed-form solution for the resulting policy. We summarize our work on information theoretic policy search for motor skill learning where we put particular focus on extending the original algorithm to learn several options for a motor task, select an option for the current situation, adapt the option to the situation and sequence options to solve an overall task. Finally, we illustrate the performance of our algorithm with experiments on real robots.

3.15 Multi-Objective Reinforcement Learning

Ann Nowe (Free University of Brussels, BE)

Joint work of Nowe, Ann; Van Moffaert, Kristof; Drugan, M. Madalina
Main reference K. Van Moffaert, M. M. Drugan, A. Nowé, "Hypervolume-based Multi-Objective Reinforcement Learning," in Proc. of the 7th Int'l Conf. on Evolutionary Multi-Criterion Optimization (EMO'13), LNCS, Vol. 7811, pp. 352–366, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-37140-0_28

We focus on extending reinforcement learning algorithms to multi-objective problems, where the value functions are not assumed to be convex. In these cases, the environment – either single-state or multi-state – provides the agent multiple feedback signals upon performing an action. These signals can be independent, complementary or conflicting. Hence, multi-objective reinforcement learning (MORL) is the process of learning policies that optimize multiple criteria simultaneously. In our talk, we briefly describe our extensions to multi-armed bandits and reinforcement learning algorithms to make them applicable in multi-objective environments. In general, we highlight two main streams in MORL, i.e. either the scalarization or the direct Pareto approach. The simplest method to solve a multi-objective reinforcement learning problem is to use a scalarization function to reduce the dimensionality of the objective space to a single-objective problem. Examples are the linear scalarization function and the non-linear Chebyshev scalarization function. In our talk, we highlight that scalarization functions can be easily applied in general but their expressive power depends heavily on on the fact whether a linear or non-linear transformation to a single dimension is performed [2]. Additionally, they suffer from additional parameters that heavily bias the search process. Without scalarization functions, the problem remains truly multi-objective and$Q$-values have to be learnt for each objective individually and therefore a state-action is mapped to a Q-vector. However, a problem arises in the boots trapping process as multiple actions be can considered equally good in terms of the partial order Pareto dominance

relation. Therefore, we extend the RL boots trapping principle to propagating sets of Pareto dominating Q-vectors in multi-objective environments. In [1], we propose to store the average immediate reward and the Pareto dominating future discounted reward vector separately. Hence, these two entities can converge separately but can also easily be combined with a vector-sum operator when the actual Q-vectors are requested. Subsequently, the separation is also a crucial aspect to determine the actual action sequence to follow a converged policy in the Pareto set.

References

1 K. Van Moffaert, M. M. Drugan, A. Nowé , *Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies*,Conference on Multiple Criteria Decision Making, 2013.
2 M. M. Drugan, A. Nowé , *Designing Multi-Objective Multi-Armed Bandits: An Analysis*, in Proc of International Joint Conference of Neural Networks (IJCNN 2013).

3.16 Bayesian Reinforcement Learning + Exploration

Tor Lattimore (Australian National University – Canberra, AU)

Joint work of Lattimore, Tor; Hutter, Marcus

A reinforcement learning policy π interacts sequentially with an environment μ. In each time-step the policy π takes action $a \in \mathcal{A}$ before receiving observation $o \in \mathcal{O}$ and reward $r \in \mathcal{R}$. The goal of an agent/policy is to maximise some version of the (expected/discounted) cumulative reward. Since we are interested in the reinforcement learning problem we will assume that thetrue environment μ is unknown, but resides in some known set $\mathcal{M}$. The objective is to construct a single policy that performs well in some sense for all/most $\mu \in \mathcal{M}$. This challenge has been tackled for many specific $\mathcal{M}$, including bandits and factored/partially observable/regular MDPs, but comparatively few researchers have considered more general history-based environments. Here we consider arbitrary countable $\mathcal{M}$ and construct a principled Bayesian inspired algorithm that competes with the optimal policy in Cesaro average.

3.17 Knowledge-Seeking Agents

Laurent Orseau (AgroParisTech – Paris, FR)

Joint work of Orseau, Laurent; Lattimore, Tor; Hutter, Marcus
Main reference L. Orseau, T. Lattimore, M. Hutter, "Universal Knowledge-Seeking Agents for Stochastic Environments," in Proc. of the 24th Int'l Conf. on Algorithmic Learning Theory (ALT'13), LNCS, Vol. 8139, pp. 146–160, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-40935-6_12
URL http://www.hutter1.net/publ/ksaprob.pdf

Observing that the optimal Bayesian rational agent AIXI does not explore its environment entirely led us to give a seek a definition of an optimal Bayesian that does so in an optimal way. We recently defined such a knowledge-seeking agent, KL-KSA, designed for countable hypothesis classes of stochastic environments. Although this agent works for arbitrary countable classes and priors, we focus on the especially interesting case where all stochastic

computable environments are considered and the prior is based on Solomonoff's universal prior. Among other properties, we show that KL-KSA learns the true environment in the sense that it learns to predict the consequences of actions it does not take. We show that it does not consider noise to be information and avoids taking actions leading to inescapable traps. We also present a variety of toy experiments demonstrating that KL-KSA behaves according to expectation.

3.18 Toward a more realistic framework for general reinforcement learning

Laurent Orseau (AgroParisTech – Paris, FR)

Joint work of Orseau, Laurent; Ring, Mark
Main reference L. Orseau, M. Ring, "Space-Time Embedded Intelligence," in Proc. of the 5th Int'l Conf. on Artificial General Intelligence (AGI'12), LNCS, Vol. 7716, pp. 209–218, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-35506-6_22

The traditional agent framework, commonly used in reinforcement learning (RL)and elsewhere, is particularly convenient to formally deal with agents interacting with their environment. However, this framework has a number of issues that are usually of minor importance but can become severe when dealing with general RL and artificial general intelligence, where one studies agents that are optimally rational, or can merely have a human-level intelligence. As a simple example, an intelligent robot that is controlled by rewards and punishments through a remote control should by all means try to get hold of this remote control, in order to give itself as many rewards as possible. In a series of paper [1, 2, 4, 3], we studied various consequences of integrating the agent more and more in its environment, leading to a new definition of artificial general intelligence where the agent is fully embedded in the world, to the point where it is even computed by it [3].

References

1 Ring, M., & Orseau, L. (2011). *Delusion, Survival, and Intelligent Agents.* Artificial General Intelligence (AGI) (pp. 11–20). Berlin, Heidelberg: Springer.
2 Orseau, L., & Ring, M. (2011). *Self-Modification and Mortality in Artificial Agents.* Artificial General Intelligence (AGI) (pp. 1–10). Springer.
3 Orseau, L., & Ring, M. (2012). *Space-Time Embedded Intelligence.* Artificial General Intelligence (pp. 209–218). Oxford, UK: Springer Berlin Heidelberg.
4 Orseau, L., & Ring, M. (2012). *Memory Issues of Intelligent Agents.* Artificial General Intelligence (pp. 219–231). Oxford, UK: Springer Berlin Heidelberg.

3.19 Colored MDPs, Restless Bandits, and Continuous State Reinforcement Learning

Ronald Ortner (Montan-Universität Leoben, AT)

Joint work of Ortner, Ronald; Ryabko, Daniil; Auer, Peter; Munos, Rémi
Main reference R. Ortner, D. Ryabko, P. Auer, R. Munos, "Regret Bounds for Restless Markov Bandits," in Proc. of the 23rd Int'l Conf. on Algorithmic Learning Theory (ALT'12), LNCS, Vol . 7568, pp. 214–228, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-34106-9_19

We introduce the notion of colored MDPs that allows to add structural information to ordinary MDPs. Thus, state-action pairs are assigned the same color when they exhibit similar rewards and transition probabilities. This extra information can be exploited by an adaptation of the UCRL algorithm, leading to regret bounds that depend on the number of colors instead of the size of the state-action space. As applications, we are able to derive regret bounds for the restless bandit problem as well as for continuous state reinforcement learning.

3.20 Reinforcement Learning using Kernel-Based Stochastic Factorization

Joëlle Pineau (McGill University – Montreal, CA)

Joint work of Barreto, André M.S.; Precup, D.; Pineau, Joëlle
Main reference A. M. S. Barreto, D. Precup, J. Pineau, "Reinforcement Learning using Kernel-Based Stochastic Factorization," in Proc. of the 25th Annual Conf. on Neural Information Processing Systems (NIPS'11), pp. 720–728, 2011.
URL http://books.nips.cc/papers/files/nips24/NIPS2011_0496.pdf
URL http://www.cs.mcgill.ca/~jpineau/files/barreto-nips11.pdf

Recent years have witnessed the emergence of several reinforcement-learning techniques that make it possible to learn a decision policy from a batch of sample transitions. Among them, kernel-based reinforcement learning (KBRL)stands out for two reasons. First, unlike other approximation schemes, KBRL always converges to a unique solution. Second, KBRL is consistent in the statistical sense, meaning that adding more data improves the quality of the resulting policy and eventually leads to optimal performance. Despite its nice theoretical properties, KBRL has not been widely adopted by the reinforcement learning community. One possible explanation for this is that the size of the KBRL approximator grows with the number of sample transitions, which makes the approach impractical for large problems. In this work, we introduce a novel algorithm to improve the scalability of KBRL. We use a special decomposition of a transition matrix, called stochastic factorization, which allows us to fix the size of the approximator while at the same time incorporating all the information contained in the data. We apply this technique to compress the size of KBRL-derived models to a fixed dimension. This approach is not only advantageous because of the model-size reduction; it also allows a better bias-variance trade-off, by incorporating more samples inthe model estimate. The resulting algorithm, kernel-based stochastic factorization (KBSF), is much faster than KBRL, yet still converges to a unique solution. We derive a theoretical bound on the distance between KBRL's solutionand KBSF's solution. We show that it is also possible to construct the KBSF solution in a fully incremental way, thus freeing the space

complexity of the approach from its dependence on the number of sample transitions. The incremental version of KBSF (iKBSF) is able to process an arbitrary amount of data, which results in a model-based reinforcement learning algorithm that canbe used to solve large continuous MDPs in on-line regimes. We present experiments on a variety of challenging RL domains, including thedouble and triple pole-balancing tasks, the Helicopter domain, the penthatlon event featured in the Reinforcement Learning Competition 2013, and a model of epileptic rat brains in which the goal is to learn a neurostimulation policy to suppress the occurrence of seizures.

3.21 A POMDP Tutorial

Joëlle Pineau (McGill University – Montreal, CA)

URL http://www.cs.mcgill.ca/~jpineau/talks/jpineau-dagstuhl13.pdf

This talk presented key concepts, algorithms, theory and empirical results pertaining to learning and planning in Partially Observable Markov Decision Processes (POMDPs).

3.22 Methods for Bellman Error Basis Function construction

Doina Precup (McGill University – Montreal, CA)

Function approximation is crucial for obtaining good results in large reinforcement learning tasks, but the problem of devising a good function approximator is difficult and often solved in practice by hand-crafting the"right" set of features. In the last decade, a considerable amount of effort has been devoted to methods that can construct value function approximators automatically from data. Among these methods, Bellman error basis functionc onstruction (BEBF) are appealing due to their theoretical guarantees and good empirical performance in difficult tasks. In this talk, we discuss on-going developments of methods for BEBF construction based on random projections (Fard, Grinberg, Pineau and Precup, NIPS 2013) and orthogonal matching pursuit (Farahmand and Precup, NIPS 2012).

3.23 Continual Learning

Mark B. Ring (Anaheim Hills, US)

Joint work of Ring, Mark B.; Schaul, Tom; Schmidhuber, Jürgen
Main reference M. B. Ring, T. Schaul, "The organization of behavior into temporal and spatial neighborhoods," in Proc. of the 2012 IEEE Int'l Conf. on Development and Learning and Epigenetic Robotics (ICDL-EPIROB'12), pp. 1–6, IEEE, 2012.
URL http://dx.doi.org/10.1109/DevLrn.2012.6400883

A continual-learning agent is one that begins with relatively little knowledge and few skills but then incrementally and continually builds up new skills and new knowledge based on

what it has already learned. But what should such an agent do when it inevitably runs out of resources? One possible solution is to prune away less useful skills and knowledge, which is difficult if these are closely connected to each other in a network of complex dependencies. The approach I advocate in this talk is to give the agent at the outset all the computational resources it will ever have, such that continual learning becomes the process of continually reallocating those fixed resources. I will describe how an agent's policy can be broken into many pieces and spread out among many computational units that compete to represent different parts of the agent's policy space. These units can then be arranged across a lower-dimensional manifold according to those similarities, which results in many advantages for the agent. Among these advantages are improved robustness, dimensionality reduction, and an organization that encourages intelligent reallocation of resources when learning new skills.

3.24 Multi-objective Reinforcement Learning

Manuela Ruiz-Montiel (University of Malaga, ES)

Joint work of Ruiz-Montiel, Manuela; Mandow, Lawrence; Pérez-de-la-Cruz, José-Luis
Main reference M. Riuz-Montiel, L. Mandow, J. L. Pérez-de-la-Cruz, "PQ-Learning: Aprendizaje por Refuerzo Multiobjetivo," in Proc. of the XV Conference of the Spanish Association for Artificial Intelligence (CAEPIA'13), pp. 139–148, 2013.
URL http://www.congresocedi.es/images/site/actas/ActasCAEPIA.pdf

In this talk we present PQ-learning, a new Reinforcement Learning (RL)algorithm that determines the rational behaviours of an agent in multi-objective domains. Most RL techniques focus on environments with scalar rewards. However, many real scenarios are best formulated in multi-objective terms: rewards are vectors and each component stands for an objective to maximize. In scalar RL, the environment is formalized as a Markov Decision Problem, defined by a set S of states, a set A of actions, a function $P_{sa}(s')$ (the transition probabilities) and a function $R_{sa}(s')$ (the obtained scalar rewards). The problem is to determine a*policy* $\pi : S \rightarrow A$ that maximizes the *discounted accumulated reward* $R_t = \Sigma_{k=0}^{\infty} \gamma^k r_{t+k+1}$.E.g., Q-learning [1] is an algorithm that learns such policy. It learns the scalar values $Q(s, a) : S \times A \rightarrow \mathbb{R}$, that represent the expected accumulated reward when following a given policy after taking a in s. The selected action a in each state is given by the expression $argmax_a Q(s, a)$. In the multi-objective case the rewards are vectors $\overrightarrow{r} \in \mathbb{R}^n$, so different accumulated rewards cannot be totally ordered; $\overrightarrow{v}$ dominates $\overrightarrow{w}$ when $\exists i : v_i > w_i \wedge \nexists j : v_j < w_j$. Given a set of vectors, those that are not dominated by any other vector are said to lie in the *Pareto front*. We seek the set of policies that yield non-dominated accumulated reward vectors. The literature on multi-objective RL (MORL) is relatively scarce (see Vamplew et al. [2]). Most methods use preferences (lexicographic ordering or scalarization) allowing a total ordering of the value vectors, and approximate the front by running a scalar RL method several times with different preferences. When dealing with non-convex fronts, only a subset of the solutions is approximated. Some multi-objective dynamic programming (MODP) methods calculate all the policies at once, assuming a perfect knowledge of $P_{sa}(s')$ and $R_{sa}(s')$. We deal with the problem of efficiently approximating all the optimal policies at once, without sacrificing solutions nor assuming a perfect knowledge of the model. As far as we know, our algorithm is the first to bring these featurestogether. As we aim to learn a set of policies at once, Q-learning is apromising

starting point, since the policy used to interact with theenvironment is not the same that is learned. At each step, Q-learning shifts the previous estimated Q-value towards its new estimation: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma max_{a'} Q(s',a'))$. In PQ-learning, Q-values are sets of vectors, so the max operator is replaced by $ND(\bigcup_{a'} Q(s',a')))$, where ND calculates the Pareto front. A naive approach to perform the involved set addition is a pairwise summation (imported from MODP methods), but it leads to an uncontrolled growth of the sets and the algorithm becomes impractical, as it sums vectors that correspond to different action sequences. The results of these mixed sums are useless when learning deterministic policies, because two sequences cannot be followed at once. We propose a controlled set addition that only sums those pairs of vectors that correspond to useful action sequences. This is done by associating each vector $\overrightarrow{q}$ with two data structures with information about the vectors that (1) have been updated by $\overrightarrow{q}$ and (2) have contributed to its value. In this talk we describe in detail the application ofPQ-learning to a simple example, and the results that the algorithm yields when applied to two problems of a benchmark [2]. It approximates all the policies in the true Pareto front, as opposed to the naive approach, that produces huge fronts with useless values that dramatically slow down the process.[1]

References

1 C.J. Watkins, *Learning From DelayedRewards*. PhD Thesis, University of Cambridge, 1989.
2 P. Vamplew et al., *Empirical EvaluationMethods For Multiobjective Reinforcement Learning*, in Machine Learning84(1-2) pp. 51-80, 2011.

3.25 Recent Advances in Symbolic Dynamic Programming for Hybrid MDPs and POMDPs

Scott Sanner (NICTA – Canberra, AU)

Joint work of Sanner, Scott; Zamani, Zahra

Many real-world decision-theoretic planning problems are naturallymodeled using mixed discrete and continuous state, action, and observation spaces, yet little work has provided *exact* methodsfor performing exact dynamic programming backups in such problems. This overview talk will survey a number of recent developments in the exact and approximate solution of mixed discrete and continuous (hybrid) MDPs and POMDPs via the technique of symbolic dynamic programming (SDP) as covered in recent work by the authors [1, 2, 3, 4].

References

1 S. Sanner, K. V. Delgado, and L. Nunes de Barros. Symbolic dynamic programming for discrete and continuous state MDPs. In *In Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence (UAI-11)*, Barcelona, Spain, 2011.
2 Z. Zamani, S. Sanner, K. V. Delgado, and L. Nunes de Barros. Robust optimization for hybrid mdps with state-dependent noise. In *Proc. of the 23rd International Joint Conf. on Artificial Intelligence (IJCAI-13)*, Beijing, China, 2013.

[1] This work is partially funded by: grant TIN2009-14179 (SpanishGovernment, Plan Nacional de I+D+i) and Universidad de Málaga, Campusde Excelencia Internacional Andalucía Tech. Manuela Ruiz-Montiel is funded by the Spanish Ministry of Education through the National F.P.U. Program.

3 Z. Zamani, S. Sanner, and C. Fang. Symbolic dynamic programming for continuous state and action mdps. In *In Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI-12)*, Toronto, Canada, 2012.
4 Z. Zamani, S. Sanner, P. Poupart, and K. Kersting. Symbolic dynamic programming for continuous state and observation pomdps. In *In Proc. of the 26th Annual Conf. on Advances in Neural Information Processing Systems (NIPS-12)*, Lake Tahoe, Nevada, 2012.

3.26 Deterministic Policy Gradients

David Silver (University College – London, GB)

Joint work of David Silver

In this talk we consider deterministic policy gradient algorithms for reinforcement learning with continuous actions. The deterministic policy gradient has a particularly appealing form: it is the expected gradient of the action-value function. This simple form means that the deterministic policy gradient can be estimated much more efficiently than the usual stochastic policy gradient. To ensure adequate exploration, we introduce an off-policy actor-critic algorithm that learns a deterministic target policy from an exploratory behaviour policy. We demonstrate that deterministic policy gradient algorithms can significantly outperform their stochastic counterparts in high-dimensional action spaces.

3.27 Sequentially Interacting Markov Chain Monte Carlo BasedPolicy Iteration

Orhan Sönmez (Boğaziçi University – Istanbul, TR)

Joint work of Sönmez, Orhan; Cemgil, A. Taylan

In this ongoing research, we introduce a policy iteration method where policies are evaluated using sequentially interacting Markov chain Monte Carlo (SIMCMC) [1] for planning in discrete time continuous state space Markov decision processes (MDPs).In order to do so, we utilize the expectation-maximization algorithm derived for solving MDPs [2] and employ a SIMCMC sampling scheme in its intractable expectation step. Fortunately, the maximization step has a closed form solution due to Markov properties. Meanwhile, we approximate the policy as a function over the continuous state space using Gaussian processes [3]. Hence, in the maximization step, we simply select the state-action pairs of the trajectories sampled by SIMCMC as the support of the Gaussian process approximator. We are aware that SIMCMC methods are not the best choice with respect to sample efficiency compared to sequential Monte Carlo samplers (SMCS)[4]. However, they are more appropriate for online settings due to their estimation at any time property which SMCSs lack. As a future work, we are investigating different approaches to develop an online reinforcement learning algorithm based on SIMCMC policy evaluation. As a model based approach, the dynamics of the model would be approximated with Gaussian processes [5].

References

1 Brockwell, A., Del Moral, P., Doucet A., Sequentially Interacting Markov Chain Monte Carlo Methods. emphThe Annals of Statistics, 38(6):3870–3411, December 2010.
2 Toussaint, M., Storkey, A., Probabilistic Inference for Solving Discrete and Continuous State Markov Decision Processes. In *Proc. of the 23rd Int'l Conf. on MachineLearning*, pp. 945–952, 2006.
3 Deisenroth, M., Rasmussen C.E., Peters, J., Gaussian Process Dynamic Programming. *Neurocomputing*, 72(7-9):1508–1524,March 2009.
4 Del Moral, P., Doucet A., Sequential Monte Carlo Samplers.*Journal of the Royal Statistical Society – Series B: StatisticalMethodology*, 68(3):411–436, June 2006.
5 Deisenroth, M., Rasmussen C.E.PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *Proc. of the 28th Int'l Conf. on Machine Learning*, pp. 465–472, 2006.

3.28 Exploration versus Exploitation in Reinforcement Learning

Peter Sunehag (Australian National University, AU)

Main reference T. Lattimore, M. Hutter, P. Sunehag, "The Sample-Complexity of General Reinforcement Learning," arXiv:1308.4828v1 [cs.LG]; and in Proc. of the 30th Int'l Conf. on Machine Learning (ICML'13), JMLR W&CP 28(3):28–36, 2013.
URL http://arxiv.org/abs/1308.4828v1
URL http://jmlr.org/proceedings/papers/v28/lattimore13.pdf

My presentation was a tutorial overview of the exploration vs exploitation dilemma in reinforcement learning. I began in the multi-armed bandit setting and went through Markov Decision Processes to the general reinforcement learning setting that has only recently been studied. The talk discussed the various strategies for dealing with the dilemma like optimismin a frequentist setting or posterior sampling in the Bayesian setting, as well as the performance measures like sample complexity in discounted reinforcement learning or regret bounds for undiscounted the setting. Itwas concluded that sample complexity bounds can be proven in much more general settings than regret bounds. Regret bounds need some sort of recoverability guarantees while unfortunately sample complexity says less about how much reward the agent will achieve. The speaker's recommendation is to try to achieve optimal sample complexity but only within the class of rational agents described by an axiomatic system developed from classical rational choice decision theory.

3.29 The Quest for the Ultimate TD(λ)

Richard S. Sutton (University of Alberta – Edmonton, CA)

Joint work of Maei, H. R.; Sutton, R. S.
Main reference H. R. Maei, R. S. Sutton, "GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces," in Proc. of the 3rd Conf. on Artificial General Intelligence (AGI'10), Advances in Intelligent Systems Research Series, 6 pp., 2010.
URL http://dx.doi.org/10.2991/agi.2010.22
URL http://webdocs.cs.ualberta.ca/ sutton/papers/maei-sutton-10.pdf

TD(λ) is a computationally simple model-free algorithm for learning to predict long term consequences. It has been used to learn value functions, to form multi-scale models of

the world, and to compile planning into policies for immediate action. It is a natural core algorithm for artificial intelligence based on reinforcement learning. Before realizing its full potential, however, TD(λ) needs to be generalized in several ways: to off-policy learning, as has already been partially done, to maximally general parameterization, as has also been partially done, and to off-policy eligibility traces,which was previously thought impossible but now perhaps we can see how this too can be done. In all these ways we see a glimmer of a perfected and complete algorithm—something inspired by TD(λ), with all its positive computational and algorithmic features, and yet more general, flexible, and powerful. Seeking this perfected algorithm is the quest for the ultimate TD (λ); this talk is a status report on the goals for it and the prospects for achieving them.

References

1 H. R. Maei, R. S. Sutton. GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. *In Proceedings of the Third Conference on Artificial General Intelligence*, Lugano, Switzerland, 2010.

3.30 Relations between Reinforcement Learning, Visual Input, Perception and Action

Martijn van Otterlo (Radboud University Nijmegen, NL)

Typical reinforcement learning (RL) algorithms learn from *traces* of state, action, state, action, ... sequences, in order to optimize action selection for each state (wrt. a reward criterium). The field of RL has come up with many algorithms based on abstraction and generalization[7]. In my view general RL amounts to feedback-based, *interactive experimentation* with particular abstraction levels forstates, actions and tasks. However, despite all previous efforts direct couplings of RL with complex visual input (e.g. raw images) are still rare. In roughly the recent decade, RL has been combined with so-called *relational* knowledge representation for states and languages[5, 6]. Also, many forms of *decision-theoretic planning*, using abstract or relational version of Bellman equations, can employ powerful knowledge representation schemes[4]. An interesting development is that also in the computervision community, people wish to employ similar relational generalization overvisual input, due to advances in (probabilistic) logical learning (e.g. see our recent work on the interpretation of houses from images [1] and robotics [3, 2]). My talk is about a possibilities for relational integration of both relational action and vision. The real potential of relational representations is that states can *share* information with actions (e.g. parameters, or more specifically *objects*). Possibilities exist to define novel languages for *interactive experimentation* with relational abstraction levels in the context of both complex visual input and complex behavioral output. This includes new types of interactions – for example dealing with scarce human feedback, new types of experimentation – for example incorporating visual feedback and physical manipulation, and new types of abstraction levels – such as probabilistic programming languages. I will present first steps towards amore tight integration of relational vision and relational action for interactive [2] learning settings. In addition I present several new problem domains.

[2] See also the IJCAI Workshop on Machine Learning forInteractive Systems (http://mlis-workshop.org/2013/)

References

1 L. Antanas, M. van Otterlo, J. Oramas, T. Tuytelaars, and L. De Raedt. There are plenty of places like home: Using relational representations in hierarchy for distance-based image understanding. *Neurocomputing*, 2013; *in press* (http://dx.doi.org/10.1016/j.neucom.2012.10.037).

2 B. Moldovan, P. Moreno, and M. van Otterlo. On the use of probabilistic relational affordance models for sequential manipulation tasks in robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

3 B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4373–4378, 2012.

4 M. van Otterlo. Intensional dynamic programming: A Rosetta stone for structured dynamic programming. *Journal of Algorithms*, 64:169–191, 2009.

5 M. van Otterlo. *The Logic of Adaptive Behavior*. IOS Press, Amsterdam, The Netherlands, 2009.

6 M. van Otterlo. Solving relational and first-order logical Markov decision processes: A survey. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, chapter 8. Springer, 2012.

7 M. Wiering and M. van Otterlo. *Reinforcement Learning: State-of-the-Art*. Springer, 2012.

3.31 Universal RL: applications and approximations

Joel Veness (University of Alberta – Edmonton, CA)

Joint work of Veness, Joel; Hutter, Marcus
Main reference J. Veness, "Universal RL: applications and approximations," 2013.
URL http://ewrl.files.wordpress.com/2012/10/ewrl11_submission_27.pdf

While the main ideas underlying Universal RL have existed for over a decadenow (see [1] for historical context), practical applications are only juststarting to emerge. In particular, the direct approximation introduced by Venesset al. [2, 3] was shown empirically to compare favorably to a number of other model-based RL techniques on small, partially observable environments with initially unknown, stochastic dynamics. Since then, a variety of additional techniques have been introduced that allow for the construction of far more sophisticated approximations. We present and review some of the mainideas that have the potential to lead to larger scale applications.

References

1 M. Hutter. *One decade of universal artificial intelligence*. In Theoretical Foundations of Artificial General Intelligence, pages 67–88. Atlantis Press, 2012.

2 J. Veness, K. S. Ng, M. Hutter, and D. Silver. Reinforcement Learning via AIXI Approximation. In AAAI, 2010.

3 J. Veness, K. S. Ng, M. Hutter, W. T. B. Uther, and D. Silver. A Monte-Carlo AIXI Approximation. Journal of Artificial Intelligence Research (JAIR), 40:95–142, 2011.

3.32 Learning and Reasoning with POMDPs in Robots

Jeremy Wyatt (University of Birmingham – Birmingham, GB)

Sequential decision making under state uncertainty is a well understood if intractable problem.In this talk I show various ways that approximate methods for belief state planning and learningcan be developed to solve practical robot problems including those that scale to high dimensional continuous state and action spaces, problems with incomplete information, and problems requiring real-time decision making.

4 Schedule

- Monday
  Morning
  - Self-introduction of participants
  - **Peter Sunehag:** Tutorial on exploration/exploitation

  Afternoon
  - **Tom Dietterich:** Solving Simulator-Defined MDPs for Natural Resource Management
  - **Csaba Szepesvari, Gergely Neu:** Online learning in Markov decision processes
  - **Ronald Ortner:** Continuous RL and restless bandits
  - **Martijn van Otterlo:** Relations Between Reinforcement Learning, Visual Input, Perception and Action
- Tuesday
  Morning
  - **M. Ghavamzadeh and A. Lazaric:** Tutorial on Statistical Learning Theory in RL and Approximate Dynamic Programming
  - **Gerard Neumann:** Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search
  - **Doina Precup:** Methods for Bellman Error Basis Function construction
  - **Joëlle Pineau:** Reinforcement Learning using Kernel-Based Stochastic Factorization

  Afternoon
  - **Petar Kormushev:** Reinforcement Learning with Heterogeneous Policy Representations
  - **Mark B. Ring:** Continual learning
  - **Rich Sutton:** The quest for the ultimate TD algorithm
  - **Jan H Metzen:** Learning Skill Templates for Parameterized Tasks
  - **Robert Busa-Fekete:** Preference-based Evolutionary Direct Policy Search
  - **Orhan Sönmez:** Sequentially Interacting Markov Chain Monte Carlo Based Policy Iteration
- Wednesday
  Morning
  - **Marcus Hutter:** Tutorial on Universal RL
  - **Joel Veness:** Universal RL – Applications and Approximations
  - **Laurent Orseau:** Optimal Universal Explorative Agents
  - **Tor Lattimore:** Bayesian Reinforcement Learning + Exploration
  - **Laurent Orseau:** More realistic assumptions for RL

Afternoon

- **Scott Sanner:** Tutorial on Symbolic Dynamic Programming
- **Rico Jonschkowski:** Representation Learning for Reinforcement Learning in Robotics
- **Timothy Mann:** Theoretical Analysis of Planning with Options
- **Mohammad Ghavamzadeh:** SPSA based Actor-Critic Algorithm for Risk Sensitive Control
- **David Silver:** Deterministic Policy Gradients

- Thursday

Morning

- **Joëlle Pineau:** Tutorial on POMDPs
- **Jeremy Wyatt:** Learning and Reasoning with POMDPs in Robots
- **Scott Sanner:** Recent Advances in Symbolic Dynamic Programming for Hybrid MDPs and POMDPs
- **Lutz Frommberger:** Some thoughts on Transfer Learning in RL – On States and Representation
- **Will Uther:** Tree-based MDP – PSR
- **Nils Siebel:** Neuro-Evolution

Afternoon Hiking

- Friday

Morning

- **Christos Dimitrakakis:** ABC and cover-tree reinforcement learning
- **Manuela Ruiz-Montiel:** Multi-objective Reinforcement Learning
- **Ann Nowe:** Multi-Objective Reinforcement Learning
- **Rico Jonschkowski:** Temporal Abstraction in Reinforcement Learning with Proximity Statistics
- **Christos Dimitrakakis:** RL competition

Participants

- Peter Auer
Montan-Universität Leoben, AT
- Manuel Blum
Albert-Ludwigs-Universität Freiburg, DE
- Robert Busa-Fekete
Universität Marburg, DE
- Yann Chevaleyre
University of Paris North, FR
- Marc Deisenroth
TU Darmstadt, DE
- Thomas G. Dietterich
Oregon State University, US
- Christos Dimitrakakis
EPFL – Lausanne, CH
- Lutz Frommberger
Universität Bremen, DE
- Jens Garstka
FernUniversität in Hagen, DE
- Mohammad Ghavamzadeh
INRIA Nord Europe – Lille, FR
- Marcus Hutter
Australian National Univ., AU
- Rico Jonschkowski
TU Berlin, DE
- Petar Kormushev
Italian Institute of Technology – Genova, IT
- Tor Lattimore
Australian National Univ., AU
- Alessandro Lazaric
INRIA Nord Europe – Lille, FR
- Timothy Mann
Technion – Haifa, IL
- Jan Hendrik Metzen
Universität Bremen, DE
- Gergely Neu
Budapest University of Technology & Economics, HU
- Gerhard Neumann
TU Darmstadt, DE
- Ann Nowé
Free University of Brussels, BE
- Laurent Orseau
AgroParisTech – Paris, FR
- Ronald Ortner
Montan-Universität Leoben, AT
- Joëlle Pineau
McGill Univ. – Montreal, CA
- Doina Precup
McGill Univ. – Montreal, CA
- Mark B. Ring
Anaheim Hills, US
- Manuela Ruiz-Montiel
University of Malaga, ES
- Scott Sanner
NICTA – Canberra, AU
- Nils T. Siebel
Hochschule für Technik und Wirtschaft – Berlin, DE
- David Silver
University College London, GB
- Orhan Sönmez
Bogaziçi Univ. – Istanbul, TR
- Peter Sunehag
Australian National Univ., AU
- Richard S. Sutton
University of Alberta, CA
- Csaba Szepesvári
University of Alberta, CA
- William Uther
Google – Sydney, AU
- Martijn van Otterlo
Radboud Univ. Nijmegen, NL
- Joel Veness
University of Alberta, CA
- Jeremy L. Wyatt
University of Birmingham, GB