Reinforcement Learning

Reinforcement learning ranks among the biggest challenges for machine learning. Just controlling a known dynamical system is hard on its own – interacting with an *unknown* system poses even harder decision problems, such as the infamous exploration-exploitation tradeoff. Most research in this area is still confined to theoretical analysis and simplistic experiments, but the promise of autonomous machines justifies the effort. Over the past years, members of the department contributed to reinforcement learning in theory and experiment.

Non-Parametric Dynamic Programming Kroemer et al. [9] showed that a non-parametric kernel density representation of system dynamics unifies several popular policy evaluation methods: their Galerkin method joins Least-Squares Temporal Difference learning, Kernelized Temporal Difference learning, and a type of discrete-state Dynamic Programming, as well as a novel method of improved performance.

EM-like Reinforcement Learning Policy search, a successful approach to reinforcement learning, directly maximizes the expected return of a policy – in contrast to value function approximation, which derives policies from a learnt value function. However, few of its variants scale to many dimensions, as they are based on gradient descent over many trials. To improve efficiency, Kober & Peters [2] reduced the problem to reward-weighted imitation, treating rewards received after actions as improper probabilities indicating the actions' success. Their idea resembles Expectation Maximization, giving good actions a higher probability to be reused. This framework also unifies previous algorithms, and allows the derivation of novel ones, such as episodic reward-weighted regression and PoWER.

Relative Entropy Policy Search Policy improvements in policy search often invalidate previously collected information, causing premature convergence and implausible solutions. These problems may be addressed by constraining the information loss. Relative Entropy Policy Search (REPS) [10] bounds the information loss while maximizing expected return. REPS differs significantly from previous policy gradient approaches. It yields an exact update shown to work well on reinforcement learning benchmarks. REPS can be generalized hierarchically [6] using a gating network to choose among several option policies. This hierarchical REPS learns versatile solutions while increasing learning speed and the quality of the learnt policy.

Bayesian reinforcement learning Probability theory gives a uniquely coherent answer to the exploration-exploitation dilemma: From the Bayesian perspective, reinforcement learning is about including possible future observations in considerations about optimal behaviour. Since probabilistic models can predict future data, this process can be rigorously formalized. It amounts to modelling knowledge as an additional dynamic variable to be controlled. In general, the combinatorial number of possible futures is intractable; however, Hennig [8] showed that the Gaussian process (GP) framework, in which predictions involve linear algebra calculations, allows approximating optimal exploration-exploitation with classic numerical methods for the solution of stochastic differential equations.

Reinforcement learning with Gaussian Processes Deisenroth et al. [1] used GPs for approximate dynamic programming in reinforcement learning, as probabilistic function approximators for the value function, and as models of the system dynamics. Using the predictive uncertainty for guidance, active learning methods could explore the state space efficiently.

Rasmussen and Deisenroth [7] proposed a particularly efficient use of GPs for optimal control over continuous states for non-bifurcating systems with low sampling rate. In their work, GPs capture information gained, as well as remaining uncertainty due to noise and lack of experience. The system's behavior is predicted by propagating state and action distributions through time, tractability is achieved approximating distributions by moment matched Gaussians. This "virtual simulation" is used to optimize the control policy. Their algorithms learn from even limited interactions with the environment due to the power of using probabilistic forward models for such indirect experience rehearsal.

Apprenticeship learning via inverse reinforcement learning Unguided exploration can be hazardous for

systems like robots. This is addressed by imitation learning from example actions provided by an expert, where the autonomous agent learns a policy generalizing the demonstrations to new states. This *behavioral cloning* may fail when the dynamics of expert and learner differ. Indeed, even simple repetition of the expert's actions does not always yield the same results. An alternative is to infer the expert's reward function from the expert's behavior, then use it to learn in the new system. This avoids exhaustive exploration by searching for policies close to the expert's. Previous work required a model of the expert's dynamics, but Boularias et al. [4] presented a model-free inverse reinforcement learning algorithm, using importance sampling to adapt expert examples to the learner's dynamics. Tested on several benchmarks, the algorithm proved more efficient than the state of the art.

Generalization in both forward and inverse reinforcement learning depends on the projection of states onto *features* to describe reward and value function. Features, especially visual ones, are often subject to noise, for example in robot grasping and manipulation tasks. To solve this problem, Boularias et al. [5] combined control and structured output prediction over Markov Random Fields to represent the action distribution. Their method is robust to noise in a grasping task, and can also be used in other applications requiring control from vision.

Data-dependent Analysis of Reinforcement Learning Many analyses of reinforcement learning focus on worst-case scenarios, although reality is often not adversarial. Seldin et al. [3] used PAC-Bayesian inequalities for martingales in a data-dependent analysis of the exploration-exploitation trade-off [11]. They studied stochastic multi-armed bandits with side information (also known as contextual bandits), a general framework where at each round of the game the agent is presented side information (e.g., symptoms of a patient in a medical application) and has to find the best action (e.g., the best drug to prescribe given the symptoms). This model class is also used for personalized advertising on the Internet. Their analysis includes the actual usage of side information by the algorithm, rather than the total amount of side information provided. This allows offering a lot of side information and letting the algorithm decide what is relevant, improving the run time of the algorithm exponentially over the state of the art.

Publications

Journal Articles

[1] MP Deisenroth, CE Rasmussen, and J Peters. Gaussian process dynamic programming. *Neurocomputing*, 72(7-9):1508–1524, 3 2009. 1

[2] J Kober and J Peters. Policy search for motor primitives in robotics. *Machine Learning*, 84(1-2):171–203, 7 2011. 1

[3] Y Seldin, F Laviolette, N Cesa-Bianchi, J Shawe-Taylor, and P Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, pages 1–7, 6 2012. 2

Articles in Conference Proceedings

[4] A Boularias, J Kober, and J Peters. Relative entropy inverse reinforcement learning. In G Gordon, D Dunson, and M Dudík, editors, *14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 182–189, Ft. Lauderdale, FL, USA, 4 2011. MIT Press. 2

[5] A Boularias, O Kroemer, and J Peters. Structured apprenticeship learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2012)*, Bristol, UK, 2012. 2

[6] C Daniel, G Neumann, and J Peters. Hierarchical relative entropy policy search. In *15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9, La Palma, Canary Islands, Spain, 4 2012. 1

[7] MP Deisenroth and CE Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In L Getoor and T Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 465–472, Bellevue, Washington, USA, 2011. Omnipress. 1

[8] P Hennig. Optimal reinforcement learning for Gaussian systems. In J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 325–333, Granada, Spain, 2011. 1

[9] O Kroemer and J Peters. A non-parametric approach to dynamic programming. In J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1719–1727, Granada, Spain, 2011. 1

[10] J Peters, K Mülling, and Y Altun. Relative entropy policy search. In M Fox and D Poole, editors, *24th National Conference on Artificial Intelligence (AAAI-10)*, pages 1607–1612, Atlanta, GA, USA, 7 2010. AAAI Press. 1

[11] Y Seldin, P Auer, F Laviolette, J Shawe-Taylor, and R Ortner. PAC-Bayesian analysis of contextual bandits. In J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1683–1691, Granada, Spain, 2011. 2