Final Report:
Structure in Reinforcement Learning

J 3259-N13

Ronald Ortner

January 2, 2013

1 Online Aggregation

As planned, in the beginning of the project I've been concentrating on the topic of online aggregation for undiscounted reinforcement learning in Markov decision processes (MDPs). I've started research on online aggregation already back in Austria, so that I could quickly conclude work by proving regret bounds for a modified UCRL2 algorithm [2], which employs confidence intervals for calculating an aggregation of the estimated model MDP before computing an optimistic policy.

More precisely, given an unknown MDP the proposed algorithm UCAgg maintains confidence intervals for rewards and transition probabilities in order to define a set of plausible MDPs just like UCRL2. However, before selecting an optimistic plausible model and a respective optimal policy maximizing the optimal average reward, UCAgg tries to aggregate states for which the confidence intervals for rewards and transition probabilities intersect and are of comparable length. (This latter condition can be shown to be necessary.) That way, policy computation can be performed on a smaller MDP and therefore more efficiently. Indeed, the proposed algorithm has also been tested in experiments and showed improved performance when compared to UCRL2.

Concerning theoretical guarantees, it was possible to derive bounds on the regret the UCAgg algorithm suffers with respect to an optimal policy. These regret bounds depend on the quality of the employed aggregation algorithm, expressed by the *connectivity* of the aggregation returned. Intuitively, if the connectivity is low, then the aggregation algorithm aggregates only very similar states, while the higher the connectivity the more dissimilar states will be aggregated. As a simple aggregation algorithm a modification of the approximate aggregation algorithm of [1] can be used whose connectivity is $\sqrt{S}$ (where S is the number of states in the MDP). Thus, when combining this aggregation algorithm with UCAgg one obtains regret bounds as for UCRL2 [2], just with an additional factor $\sqrt{S}$. Since these results neither depend on an improved algorithm for state aggregation nor on more general bounds for the error caused

by state aggregation, further research on these topics had lower priority. Actually, my attempts to find an improved approximate aggregation algorithm did not produce any notable results. However, since at the core of this problem there is a clique problem (cf. [1]) which is not only NP-hard but also hard to approximate, improvements are maybe difficult to achieve.

These results on online aggregation have been accepted for publication in the *Annals of Operations Research* [6]. In February, I've presented this work also in the SequeL weekly group meeting, the seminar of the host institute.

2 Similarity Exploitation and Restless Bandits

Originally, the joint work on the restless bandit problem with Daniil Ryabko, Peter Auer, and Rémi Munos was planned to be rather a side subject, but eventually it turned out to spark ideas highly relevant for the core topics of the project.

In the restless bandit setting, the learner faces a multi-armed bandit problem with each arm having internal states evolving (independently of the learner's actions) according to a Markov process unknown to the learner. The rewards of each arm are stochastic and depend on the state of the arm when choosing it. Since it is assumed that the learner can only observe the state of the arm sampled, this is a *partially observable* Markov decision process (POMDP). However the problem can be turned into an MDP with countably infinite state space, where some of the transitions are known and the reward and transition structure exhibit certain structural symmetries known to the learner. More precisely, one keeps in mind for each arm (i) which state it was in when it was chosen last time and (ii) how many steps ago that happened. A state in the considered MDP representation then contains this information for each single arm. However, the rewards and transition probabilities for choosing an arm i only depend on the available information for arm i and are independent of the states of the other arms. Therefore, the same parameters can be sampled from different states in the MDP representation.

We have generalized this property by defining so called ε*-structured* MDPs in which additional to the standard MDP setting there is a *coloring function* available which assigns the same *color* to state-action pairs with ε-close rewards and transition probabilities. This notion of ε-structured MDPs is quite general and subsumes for example (approximate) state aggregation [5, 1] and MDP homomorphism [12].

For learning in ε-structured MDPs we were able to modify the UCRL2 algorithm [2] to ε-structured MDPs and to show enhanced regret bounds with improved dependence on the size of MDP (now depending on the number of colors instead of the number of states and actions of the MDP). These results fit perfectly in the suggested investigation into new similarity structures for MDPs with the restless bandit setting a first and natural application, giving the first distribution-independent regret bounds for that problem.

Our results have been accepted to ALT 2012 [7] and have been invited for a special issue of *Theoretical Computer Science* [11].

3 Continuous Reinforcement Learning

First attempts to achieve regret bounds in a simplified undiscounted continuous reinforcement learning setting (under deterministic transitions, piecewise linear, or even piecewise constant transitions) did not bring any countable success except under very strong additional assumptions, mainly due to problems with points of discontinuity. However, briefly after having finished work on the restless bandit problem, in joint work with Daniil Ryabko we noticed that the methods developed for analyzing the restless bandit setting [7] could be modified to deal with general continuous reinforcement learning problems under Lipschitz (or more generally Hölder) conditions on rewards and transition probabilities. In particular, in this latter setting discretizations of the continuous state space exhibit similar properties as ε- structured MDPs (cf. above), and the main difficulty was to adapt algorithms and proofs from the discrete to the continuous setting.

Thus, modifying the *colored UCRL2* algorithm employed in the restless bandit setting (for ε-structured MDPs), we were able to give an algorithm in the spirit of UCRL2 [2] which employs confidence intervals for *aggregated* rewards and transition probabilities to determine an optimistic policy. For deriving regret bounds we combined aggregation techniques with the original proof of the regret bounds for UCRL2, actually quite similar to the methods applied for showing the regret bounds for online aggregation and colored UCRL2, respectively. The derived bounds of $O(T^{3/4})$ for this algorithm are the first sublinear upper bounds on the regret in such a general reinforcement learning setting.

Results were accepted for NIPS 2012 [8] and have been presented at the conference in December 2012. In October, I've also presented these results in the SequeL weekly group meeting.

4 Selecting State Representations

In the second half of my stay — after having contributed to each topic of my original proposal — I have been mainly working on a different approach to reinforcement learning problems as introduced in [3]. The learner interacts with an environment by receiving rewards and observations in return for choosing actions, just as in the MDP case. However, it is not assumed that the state space of the MDP is known to the learner nor that the states are directly observable. Rather, the learner has a set of representations which map histories of past interactions to a discrete state space, only some of which result in the true underlying MDP process.

This setting is particularly interesting for continuous reinforcement learning problems: First of all, any representation can be considered to be a possible

discretization of an underlying continuous state space. Also, we actually employ this technique for continuous reinforcement learning when the Lipschitz/Hölder parameters are not known to the learner by considering each representation to be a discretization of the parameter space [8].
The considered setting also generalizes other approaches like context trees [4] or probabilistic deterministic finite automata [13], which try to extract some high-level discrete features from the given observations provided by the environment, which may be high-dimensional and/or continuous.

The original paper [3] showed regret bounds (with respect to the optimal policy for the true MDP representation) of order $O(T^{2/3})$ for the so-called BLB algorithm. The BLB algorithm employs the UCRL2 algorithm for each representation and uses the upper bound on the regret known for UCRL2 as a reference value. If a representation does not achieve sufficiently high rewards, it is discarded.

The BLB algorithm and its analysis come with some disadvantages. First, UCRL2 is used by BLB in a "black-box" fashion. Second, in order to apply the UCRL2 regret bounds, BLB needs to guess the *diameter* [2] of each representation, which comes at the cost of an additive constant exponential in the diameter of the underlying true MDP. Finally, the analysis is restricted to finite sets of representation functions.

Together with Daniil Ryabko, Odalric-Ambrym Maillard (who was a PhD at Inria Lille before joining our department in Leoben as a PostDoc), and Phuong Ngyuen (who is a PhD student of Marcus Hutter and has been visiting Inria Lille for two months), we could improve over the original results in the following way. First, using an UCRL2-like optimistic model selection approach we were able to improve the bounds from $O(T^{2/3})$ to $O(\sqrt{T})$ without the need to guess the MDP's diameter. Secondly, we could generalize the original BLB algorithm to work with an infinite set of representations.

The formerly mentioned results were recently accepted to ICML 2013 [9], while the extension to the case of infinitely many representations has been submitted to AISTATS 2013 [10].

5 Outlook

Although the work done in the ten months of the project can only be considered as a first step towards the theoretical solution of general reinforcement learning problems on the one hand and applicability of reinforcement learning problems to real-world problems on the other hand, I'm convinced that this first step is an important one and will inspire future research.

Indeed, there still remains a lot of work to be done. Concerning continuous reinforcement learning, beside some more technical issues concerning the necessary input for the algorithm, the achieved regret bounds of $\tilde{O}(T^{3/4})$ (for the Lipschitz case) do not meet the preliminary lower bound of $\Omega(\sqrt{T})$ we obtained and are probably not best possible. Further, while the class of MDPs with (Hölder or Lipschitz) continuous rewards and transitions is a natural one,

it would also be interesting to have results for other or more general classes of MDPs such that derived bounds depend on some structural parameter of the class of MDPs the MDP to be learned is taken from.

Concerning the selection of state representation models, we currently aim at settings in which the learner does not have access to the underlying MDP, but that the set of state representations available to the learner only contains a good approximation.

These questions are planned to be the topic of an FWF stand-alone project the application to which is in the course of being written.

6 Inria Lille

I'd like to conclude with some more personal remarks about working at Inria Lille. I found in the SequeL team a perfect environment for my research project, and I think that the ten months spent in Lille were probably the most fruitful in my academic career so far. I really enjoyed the open atmosphere in the team, which allowed a lot of free discussion between team members on their current research. This was complemented by regular talks by external visitors as well as people from the team.

Having returned to Austria, I am not only still in contact with members of the SequeL team, there is also some ongoing joint work on some of the open problems mentioned above as well as on new, yet related topics.

References

[1] E. Even-Dar and Y. Mansour. Approximate equivalence of Markov decision processes. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, pages 581–594, 2003.

[2] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

[3] Odalric-Ambrym Maillard, Rémi Munos, and Daniil Ryabko. Selecting the state-representation in reinforcement learning. In *Advances Neural Processing Systems 24, NIPS 2011*, pages 2627–2635, 2012.

[4] R. Andrew McCallum. *Reinforcemenet Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, University of Rochester, 1996.

[5] R. Givan, T. Dean, M. Greig, Equivalence notions and model minimization in Markov decision processes., Artif. Intell. 147 (1-2) (2003) 163–223.

[6] Ronald Ortner. Adaptive Aggregation for reinforcement learning in average reward Markov decision processes. *Annals of Operations Research*, to appear, 2012.
URL: `http://dx.doi.org/10.1007/s10479-012-1064-y`

[7] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret Bounds for Restless Markov Bandits. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory, ALT 2012. Lecture Notes in Computer Science 7568*,

pages 214–228. Springer, 2012.
URL: http://dx.doi.org/10.1007/978-3-642-34106-9_19

[8] Ronald Ortner and Daniil Ryabko. Online Regret Bounds for Undiscounted Continuous Reinforcement Learning. In *Advances in Neural Information Processing Systems 25, NIPS 2012*, to appear, 2012.
URL: http://books.nips.cc/papers/files/nips25/NIPS2012_0861.pdf

[9] Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning, accepted for ICML 2013.

[10] Phuong Nguyen, Odalric-Ambrym Maillard, Daniil Ryabko, and Ronald Ortner. Competing with an Infinite Set of Models in Reinforcement Learning, submitted to AISTATS 2013.

[11] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret Bounds for Restless Markov Bandits. *Theoretical Computer Science*, invited, in preparation.

[12] B. Ravindran, A. G. Barto, Model minimization in hierarchical reinforcement learning, in: Abstraction, Reformulation and Approximation, 5th International Symposium, SARA 2002, 2002, pp. 196–211.

[13] E. Vidal, F. Thollard, C. D. L. Higuera, F. Casacuberta, and R.C. Carrasco. Probabilitic finite-state machines. *IEEE Tr. on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005.