STRUCTURED AND CONTINUOUS REINFORCEMENT LEARNING

PROPOSAL FOR AN FWF PROJECT

RONALD ORTNER

Contents

1. Introduction

In reinforcement learning (RL) an agent tries to learn optimal behavior in an unknown environment by evaluating feedback—usually some quantifiable and comparable reward—to his actions. As the learner's actions may pay off not immediately, he may have to accept short-term discouraging feedback to achieve a long-term goal state with large positive feedback. Thus, in typical RL applications like robotics, control, or game playing [e.g., 25, 65] the learner will get

This proposal can be considered as a continuation of the Erwin Schrödinger scholarship FWF: J 3259-N13 "Structure in Reinforcement Learning".

rewarding feedback only when a given task is finished after a series of coordinated actions which individually give no or even misleading feedback.

Typically, the learner's environment is assumed not to be deterministic so that executing the same action in the same state may produce different results. Such a randomized setting gives rise to the so-called *exploration-exploitation* problem: The learner has no guarantee that the strategy that performed best so far, is indeed the best strategy in the long term. Thus the learner has to find a reasonable payoff between *exploiting* a well working strategy and *exploring* other possibilities that may give higher reward. However, this includes the risk of following a suboptimal strategy. The learner may of course first conduct an extensive exploration of the whole environment, after which an optimal strategy can be identified with high probability, which in turn can be executed in a subsequent exploitation phase. However, separating the exploration from the exploitation is rather costly, as it will take a long time to obtain a sufficiently precise approximation of the environment. Thus, we are rather interested in algorithms that learn *online* and handle the exploration-exploitation problem implicitly. To measure the performance of an online RL algorithm we consider its *regret*, that is, the loss the algorithm suffers when compared to an optimal strategy.

The major practical problem that prevents implementation of RL algorithms for many potential applications is that real world problems usually induce RL representations with large, often continuous state space, while typical algorithms are not efficient in environments with large state space. One of the reasons for this is that, unlike humans, RL algorithms are not able to exploit the environment's structure, e.g. similarities in the state space. The main goal of the project is to develop tools how to deal with structured and continuous environments.

In the precursor project "Structure in Reinforcement Learning" (FWF:J 3259-N13), together with scientists from the SequeL team at Inria Lille (cf. Section 7 below) we were able to define very general similarity structures for reinforcement learning problems in finite domains and to achieve improved theoretical regret bounds when the underlying similarity structure is known. The developed techniques and algorithms also gave the first theoretical regret bounds for reinforcement learning in continuous domains. The proposed project wants to take the research on continuous reinforcement learning —a setting which is of particular importance for applications— a step further, not only by improving over the known bounds, but also by development of efficient algorithms. Moreover, we also want to investigate in more general settings where the learner does not have direct access to the domain information, but only to a set of possible models. Also for this setting, the precursor project has produced first theoretical results, assuming finite domains and that the set of possible models contains the correct model. In the proposed project, we aim

at generalizing this to infinite domains and loosening the assumption on the model set, which shall not necessarily contain the correct model, but only a good approximation of it.

2. State of the Art

2.1. **Setting.** Reinforcement learning problems usually are formally represented as Markov decision processes [65]. In a *Markov decision process* (MDP) M with state space S and action space A, a learner starting in a given initial state $s_1 \in S$ chooses at each time step an action from A. When executing action a in state s, the learner receives a random reward r with mean $r(s,a)$ according to some distribution on a bounded interval (which we assume to be [0,1] in the following). Further, according to the transition probabilities $p(s'|s,a)$, a random transition to a state $s' \in S$ occurs. If the learner's choice of actions at each time step t is uniquely determined by the history so far, this establishes a stochastic process described by the states s_t visited at time step t, the actions a_t chosen by the learner at step t, and the rewards r_t obtained ($t \in \mathbb{N}$). Usually, one is interested in maximizing the expected average reward

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r_t\right]$$

an algorithm can gain within T steps. For MDPs with finite state and action space, the limit of this expression for $T \to \infty$, called the *average reward*, can be maximized by an appropriate stationary *policy* $\pi : S \to A$ that defines an optimal action for each state [55]. Assuming that all states are reachable from any other state, the *optimal average reward*

$$\rho^* := \max_{\pi:S\to A} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[r(s_t, \pi(s_t)\right]$$

is independent of the initial state s_1.

A special case are MDPs wit a single state, so-called *multi-armed bandits*. In this setting the learner chooses at each time step an arm a from a set A (corresponding to the action space in the MDP setting), which gives a random reward with bounded mean $r(a)$. Thus, in the multi-armed bandit setting, only the rewards are of interest.

Online Learning and Regret. We are interested in algorithms that perform well when comparing their rewards to a strategy that always chooses the best action. That is, we compare the accumulated rewards collected by the algorithm with those to be expected from an optimal policy.

Thus, we define the *regret* of an algorithm $\mathcal{A}$ after T steps to be

$$T\rho^* - \sum_{t=1}^{T} r_t.$$

In the meantime, regret bounds for (undiscounted) RL are a well-established area of research. For the sake of completeness however, we briefly indicate the relation to similar theoretical research in RL. For a detailed discussion we refer to Section 1.1 of [22].

First, for a lot of RL algorithms, ranging from earliest work like Q-learning [72] to more recent approaches like the knowledge gradient [21] one can show *asymptotical optimality*, that is, the algorithm converges to an optimal policy when the time horizon goes to infinity. While these algorithms also often work very well in practice, those theoretical guarantees are rather weak compared to regret bounds, as asymptotical optimality does not tell anything about the algorithms' finite time behavior.

PAC bounds are stronger, bounding the number of steps until an algorithm performs probably approximately optimally. Such bounds are given e.g. for the well-known algorithms E^3 [29], and R-Max [10] and can actually be translated into regret bounds (and vice versa), cf. Corollary 3 of [22].

Similar to PAC bounds, *sample complexity bounds* as introduced in [27] for R-Max count the total number of suboptimal steps of an algorithm. These bounds are also quite popular for the discounted RL setting [63, 66, 33]. Finally, we would like to mention that, while there are also regret bounds for discounted RL [60, 61], these bounds are usually weaker than those for the undiscounted setting, as the discounted regret is measured along the trajectory taken by the learning algorithm.

2.2. **Upper Confidence Bound Algorithms and Regret Bounds.** One way to handle the mentioned exploration-exploitation dilemma is the use of confidence intervals, as introduced for the multi-armed bandit setting by Lai and Robbins [32] and later simplified by the UCB algorithm of Auer et al. [4]. For both algorithms logarithmic regret bounds have been shown. These bounds are however problem-dependent, that is, they depend on further parameters of the bandit problem at hand. There are also alternative, problem-independent bounds on the regret of order $\sqrt{T}$. More precisely, there is a lower bound of $\Omega(\sqrt{|A|T})$, which is met (apart from logarithmic factors) by various algorithms [5, 24, 2].

In the general MDP setting, regret bounds which are *asymptotically* logarithmic in the horizon T were first shown for the so-called *index policies* of Burnetas and Katehakis [14], cf. also the improvements given by Tewari and Bartlett [68]. As the bandit algorithms mentioned before, these index policies also use confidence bounds to choose actions optimistically.

The first *online* regret bounds that hold for arbitrary horizon T were shown for the UCRL algorithm, first introduced in Auer and Ortner [3] and later refined by Jaksch et al. [22]. Similarly to the UCB algorithm for the bandit setting, the UCRL algorithm[1] first considers a set of plausible models defined via confidence intervals. Among the plausible models, the algorithm then chooses the *optimistic* model that promises the highest average reward and a respective optimal policy. Jaksch et al. [22] give a regret bound of $O(D|S|\sqrt{|A|T})$ and an alternative "gap"-dependent bound of $O\Big(\frac{D^2|S|^2|A|\log T}{g}\Big)$, where D is the *diameter* of the MDP defined by the expected number of steps it takes at most to reach any state from any other state, and the gap g is the difference between the optimal average reward ρ^* and the average reward of the best suboptimal policy. A complementary lower bound of $\Omega(\sqrt{D|S||A|T})$ is known as well [22].

For the REGAL algorithm [7], a modification of UCRL, alternative regret bounds where shown where the diameter in the UCRL bounds is replaced with a smaller transition parameter D_1 (corresponding to an upper bound on the span of the *bias* [55] of an optimal policy). However, this bound can be achieved only when the learner knows an upper bound on D_1. Otherwise, the learner has to guess such an upper bound, giving a slightly worse bound with respect to the size of the state space. Similarly to UCRL, REGAL chooses an optimistic MDP and a respective optimal policy, however the set of plausible MDPs is restricted to MDPs with bias span bounded by D_1. This implies another problem with REGAL: Unlike UCRL, where optimistic MDP and optimal policy can be computed using so-called *extended value iteration*— no efficient procedure is known to compute the optimistic MDP with bounded bias span needed by REGAL.

2.3. **Exploiting Structural Information.** There are many real world problems that could in principle be represented as an MDP. However the respective MDP representation would in most cases have a large state space and a large action space, so that typical algorithms like UCRL could not be employed, since complexity and regret grow linearly or even polynomially with the number of states and actions.

We are convinced that one of the main reasons for this problem is that —unlike human learning— RL algorithms are not able to exploit underlying symmetries and similarities the learning problem Thus, the main focus of the proposed project is the investigation of such similarity structures for MDPs and their exploitation by respective learning algorithms, with an emphasis on continuous (state and action) RL.

We will concentrate on the following three topics: First, we want to make progress with RL in continuous state space. The second major topic (and this can be also considered as a method to deal with continuous state space) are more general RL settings where the right

[1]In the following, I do not distinguish between the very similar UCRL algorithm of [3] and the UCRL2 algorithm of [22].

state representation has to be learned as well. Finally, we also aim at other generalizations of continuous RL settings. As the precursor project (FWF:J 3259-N13) dealt with the same topics, in the following we not only give an overview of the proposed research, but also of what has been achieved during my 10 month stay at Inria Lille.

3. Continuous Reinforcement Learning Problems

Many real world problems demand continuous state or action spaces, and one of the big challenges for RL is to deal with such continuous domains. Continuous RL settings are usually examined under strong structural assumptions. The most natural and also most common assumptions express the idea that close states (according to some metric of the underlying space) exhibit similar behavior. Usually, such assumptions are formalized as Lipschitz (or more general Hölder) continuity of reward and transition functions.

Continuous-Armed Bandit Problems. The simplest continuous RL problem is the 1-dimensional continuum-armed bandit, where the learner has to choose arms from a bounded interval. This setting has been investigated in detail. There are regret bounds for algorithms which work with a fixed discretization [30, 6], but also adaptive discretizations have been proposed and analyzed [31, 12]. In either case, regret bounds have been derived that depend on parameters of the problem, like the Hölder parameter of the reward function. However, while the continuum-armed bandit is a continuous reinforcement learning setting, it is no special case of the general continuous state MDP setting we are interested in (and rather corresponds to a continuous action space), and hence cannot be directly compared to it.

Continuous State MDPs with Lipschitz Condition. A lot of work on continuous RL is confined to rather particular settings, primarily with respect to the considered transition model. In the simplest case, the transition function is considered to be deterministic [44], and mistake bounds for the respective discounted setting have been derived by Bernstein and Shimkin [8]. We note that for *finite* MDPs with deterministic transitions, regret bounds can be achieved which are closer to the bandit than the general MDP setting, cf. [49]. Another common assumption is that transition functions are simple (e.g. linear) functions of state and action plus some noise. For such settings sample complexity bounds have been e.g. given in [62, 11]. Further, $\tilde{O}(\sqrt{T})$ bounds for the regret after T steps are shown in [1] for the adaptive control of linear quadratic systems.

There is also some research considering more general transition dynamics under the assumption that close states behave similarly, as will be considered here. While most of this work is purely experimental [23, 70], there are also some contributions with theoretical guarantees. Thus,

Kakade et al. [26] consider PAC-learning for continuous RL in metric state spaces, under the assumption that generative sampling is possible. The proposed algorithm is a generalization of the E^3 algorithm [29] to continuous domains. A respective adaptive discretization approach is suggested by Nouri and Littman [46]. The PAC-like bounds derived however depend on the (random) behavior of the proposed algorithm.

Work in the Precursor Project. In the precursor project (FWF:J 3259-N13), borrowing the idea of discretizing and applying an algorithm with known regret bounds from similar continuum-armed bandit algorithms [30, 6], we were able to give an algorithm in the spirit of UCRL [22] 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 UCRL. The derived bounds of $\tilde{O}(T^{3/4})$ for this algorithm are the first sublinear upper bounds on the regret in such a general RL setting and were accepted to NIPS [50].

Research Plan and Methodology. Although [50] was an important first step, there are still a lot of open questions and room for improvement. First, the achieved regret bound of $\tilde{O}(T^{3/4})$ (for the Lipschitz case) does not meet the preliminary lower bound of $\Omega(\sqrt{T})$ we obtained and is probably not best possible. Computationally, there are two issues concerning the UCCRL algorithm suggested in [50], similar to the REGAL algorithm discussed at the end of Section 2.2. First, it needs (an upper bound on) the bias span of the underlying MDP as an input. Not having this information, the algorithm can try to guess it keeping the bounds of the same order, yet accepting large additive constants. Second, it is not clear whether the optimistic MDP and the optimal policy UCCRL employs in each episode are efficiently computable. Like REGAL, UCCRL only considers plausible MDPs whose bias span is upper bounded by the real one. It is an open problem we want to address whether extended value iteration or other efficient algorithms can be adapted to this setting. In that respect, we also plan to consider alternative settings in which the learner tries to generalize from a set of expert trajectories or starts learning with such information given to speed up learning.

Further, while I expect this to be difficult, an improvement over the bounds for the original UCRL algorithm (diameter, state space) would also result in an improvement for the respective continuous algorithm. Further investigation into the gap between lower and upper bound on the regret of UCRL are not the main point of this project, however progress in that direction may well be the side result of research conducted within the proposed project.

While the fixed discretization approach gave the first results, in general it will not be sufficient to deal with continuous MDP learning problems. Thus, instead of that static approach, we also

want to work on algorithms which adapt the employed discretization to the collected data, as the Zooming algorithm [31] does in the continuous bandit setting. In that case, progress would be made by generalizing notions like the *zooming dimension* to general continuous RL problems. Similar notions have also been employed for sample complexity bounds in discounted RL [15].

Last but not least, while the focus of the project is on theoretical research, we hope to learn from collaboration with the Intelligent Autonomous Systems Lab of TU Darmstadt (see Section 7 below) about application of RL methods to robotics. While we do not expect that the methods we develop for proving theoretical bounds can be employed unmodified, we hope that some ideas can be adapted to work also for practical problems.

More general topics that are hoewever also relevant for continuous RL will be discussed in the two subsequent sections.

4. Selecting State Representations

The following is quite a different approach to RL as introduced by Maillard et al. [38]. 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 RL 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 RL when the Lipschitz/Hölder parameters are not known to the learner by considering each representation to be a discretization of the parameter space [50]. The considered setting also generalizes other approaches like context trees [41] or probabilistic deterministic finite automata [71], which try to extract high-level discrete features from the observations provided by the (possibly continuous and high-dimensional) environment.

The original paper [38] 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 Ucrl algorithm for each representation and uses the upper bound on the regret known for Ucrl [22] as a reference value. Representations that do not achieve sufficiently high rewards are discarded. The Blb algorithm and its analysis come with some disadvantages. First, Ucrl is used by Blb in a "black-box" fashion. Second, in order to apply the Ucrl regret bounds, BLB needs to guess the *diameter* [22] 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.

Work in the Precursor Project. Within the precursor project (FWF:J 3259-N13), we could improve over the original results of [38] in the following way. First, using an UCRL-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 first result was accepted to ICML 2013 [39], while the second one has recently been accepted for AISTATS 2013 [45].

Research Plan and Methodology. 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. Note that this setting under the assumption of having an infinite number of models is the most general setting of continuous state RL one can think about: One has the underlying continuous state MDP, yet due to reasons of computability, one prefers to work with a suitable discretization of the MDP, yet it is not clear in advance what would be a good discretization. Considering any discretization to be an approximate model, the goal is to bound the regret with respect to the best model or (this is more challenging) the optimal policy in the original continuous state MDP.

Currently, we can deal with an infinite state space [45], yet get worse regret bounds than in the finite model case. Still, the approach in [45] takes the BLB algorithm [38] as a starting point whose regret bounds are also of order $T^{2/3}$. Thus, we expect that by using similar techniques to generalize the OMS algorithm of [39] to the infinite models case.

For the step from having an exact Markov model to having only approximations at our disposal, we will try to use techniques introduced in [51] and [47] where we also had to deal with learning in MDP using approximations, yet in a different context. Concerning the approximations, there are three scenarios (of increasing difficulty) we want to examine: First, each model (of finitely many) comes with a predefined approximation error of that model, and there is one suitable model approximating the true MDP model with exactly that error, second the approximation error has to be guessed for each model, third extend this to the case of infinitely many models. As all algorithms so far have only dealt with the setting where the underlying true MDP has finite state and action space, it will be the first step to consider the setting of finding the best approximate state representation under this same assumption. Later, we will try to adapt our results to the case of continuous state and action space.

Another open question that shall be addressed is the dependence on the model parameters, that is, size of state and action space etc. While the original BLB algorithm proposed in [38] has worse dependence with respect to the horizon, the dependence with respect to the state space is much better than for the OMS algorithm of [39] that gives optimal dependence on the horizon. That is, while the regret of [38] only depends on the state space of the true model, for [39] the

state spaces of all models appear as a factor in the regret bound. It is an open question we want to address whether it is possible to have regret bounds that are optimal with respect to the horizon and only depend on the true state space.

5. Generalizations

While the emphasis of the project lies on continuous RL, progress may be made on more general settings as also exemplified in the previous section on selecting state-representations. Thus, in general it may be useful to generally deal with structures in (discrete or continuous) RL problems and ways how to exploit them. Particularly interesting structures are similarities between states or also actions. Often, such similarity structures lead to some kind of aggregation of similar states to meta-states. This topic has already been extensively studied in the literature and has found various applications and extensions, e.g. aggregation in factored MDPs [20, 16, 28], 2-player zero-sum games [54], multiagent settings [73], and theory of social choice [53] to name a few. A lot of work on aggregation, similarity and state abstraction in MDPs considers the approximate dynamic programming view, when the underlying MDP is known. Specific topics of this work include e.g. approximate value or policy iteration [58]. For an overview see e.g. [13, 42, 36]. In practical problems "neat" aggregations are not always possible. Still, by aggregating states which are sufficiently similar, the problem at hand can be approximated reasonably well. Respective bounds on the loss by aggregation have been given in [18, 19, 67, 57, 48]. Complexity is also an issue here. While finding a minimal perfect aggregation is a problem that can be solved in polynomial time [20, 43], Even-Dar and Mansour [18] have shown that finding an optimal *approximate* aggregation is hard even when the MDP is known.

There is also work on learning in MDPs with exploitation of an underlying (similarity or other) structure. Thus, Leffler et al. [34] consider a (discrete) state space which is partitioned into sets of states of the same type, where states of the same type are assumed to have the same transition dynamics. That way, less experience is needed until good performance is achieved when compared to standard algorithms. Also, RL in factored MDPs has been considered in [64, 37, 17], including sample complexity bounds for a variant of the R-Max algorithm Li [35]. Concerning learning in structured MDPs, unlike in aggregation where the aim is to simplify the MDP, our interest lies rather in speeding up learning, which not necessarily allows simplification of the underlying MDP.

Work in the Precursor Project.

Online Aggregation. If the underlying symmetries of the MDP are not known beforehand, finding an aggregation of the MDP *online* while learning becomes an interesting topic, have been

discussed in the literature only to a limited extent [40, 59, 9], and usually without any theoretical guarantees. In the precursor project (FWF:J 3259-N13) we introduced a modified UCRL algorithm [22], which employs confidence intervals for calculating an aggregation of the estimated model MDP before computing an optimistic policy. 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 UCRL. We also derived regret bounds for the proposed algorithm that also depend on the quality of the employed aggregation algorithm. For a suggested simple aggregation algorithm (a modification of an algorithm suggested in [18]) one obtains regret bounds as for UCRL [22], just with an additional factor $\sqrt{S}$. These results have been accepted for publication in the *Annals of Operations Reserach* [47].

Similarity Exploitation and Restless Bandits. Concerning the development and exploitation of new similarity structures the so-called *restless bandit* setting proved to be a rich test bed for the development of similarity structures in RL. 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). The problem can be turned into an MDP with countably infinite state space however, where some of the transitions are known and the reward and transition functions exhibit certain structural symmetries known to the learner. Generalization of this structure has led to the definition of 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 [20, 18] and MDP homomorphisms [56].

For learning in ε-structured MDPs we were able to modify the UCRL algorithm [22] to ε-structured MDPs and to show enhanced regret bounds with improved dependence on the size of MDP. That is, the regret now depends on the number of colors instead of the number of states and actions of the MDP. These results were accepted to ALT [51] and have been invited to *Theoretical Computer Science* [52]. Further, the methods developed for the analysis of the restless bandit setting eventually also led to the already mentioned first online regret bounds for continuous RL [50].

Research Plan and Methodology. First, as a continuation of the work done in the precursor project, there are still some open questions concerning the restless bandit setting. Thus, while the regret bounds are optimal with respect to the horizon, there is still a large gap between the lower and the upper bound on the regret, when considering all parameters of the setting. Also, the current analysis does not make use of all the structural information present in the setting. Further investigation into this may lead to more general patterns that may turn out to be useful in other cases as well.

On the other hand, consideration of more complex POMDP settings is of interest as well. Although the approach used in the restless bandit will probably not generalize to arbitrary POMDPs, it would be at least interesting to see how far it works, and whether further modifications allow generalization to wider sets of POMDPs.

Another topic will be the generalization of the continuous RL setting to RL in arbitrary topological or metric spaces. For the continuum-armed bandit setting there are analogous generalizations where arms are chosen from arbitrary topological or metric spaces [31, 12]. Regret bounds then often depend on properties of this underlying space like the *zooming dimension* introduced in [31] or the *near-optimality dimension* of [12]. We want to investigate whether notions and algorithms can be extended to the general RL setting, trying to modify methods developed for the continuous RL setting under Lipschitz assumption.

A related generalization of the continuous RL setting we want to investigate is the following: Instead of considering Lipschitz continuous MDPs, one assumes more general classes of MDPs and tries to obtain regret bounds that depend on some structural parameter of the class of MDPs to be learned. Here, as a first step one has to come up with possible notions of measuring the "size" of the set of MDPs analogous to the VC dimension in supervised classification problems.

6. The Local Environment

The Chair of Information Technology is the youngest of the four chairs at the *Department of Mathematics and Information Technology* of the Montanuniversität Leoben. Peter Auer has been the head of the institute since its beginning in 2002, while I (the applicant) joined soon afterwards in 2003, now filling a permanent position as Associate Professor. There is only one other researcher employed by the university, Martin Antenreiter. Until recently, two further PostDocs Odalric-Ambrym Maillard (now at the Technion in Haifa) and Shiau Hong Lim (now at the national University of Singapure) have been employed by an EC project (CompLACS, EC project FP7/2007-2013, n° 270327). A new project employee (PostDoc) shall be recruited by summer 2013.

The institute's members do research in the field of machine learning with focus on reinforcement learning on the one hand and computer/cognitive vision on the other hand. The institute has been involved in a series of national (NFN Cognitive Vision, FWF project S9104-N13 SP4) and international (LAVA, EC project IST-2001-34405, PinView, EC project FP7/2007-2013, n° 216529; PASCAL Network of Excellence, EC project IST-2002-506778; PASCAL2, EC project FP7/2007-2013, n° 216886) research projects, and another FP7 project (ComPLACS, n° 270327). Within these projects, connections to national and European research partners have been established, from which also the proposed project will benefit.

The institute can provide all necessary resources needed for hosting another researcher, that is, a desk and a computer with internet access. Further, a fast server for performing simulation experiments is at the institute's disposal.

7. Collaborations: Inria Lille, TU Darmstadt

The SequeL group at Inria Lille - Nord Europe (Parc Scientifique de la Haute-Borne, 40 avenue Halley, 59650 Villeneuve d'Ascq) is one of the most important centers for reinforcement learning reasearch in Europe. Researchers at SequeL originate from different fields, such as computer science, applied mathematics, and signal processing. Beside the SequeL group there are further groups at Inria working on control theory, optimization, and statistical learning, respectively, thus offering a perfect environment for collaboration on the proposed research topics.

The applicant has spent 10 months of the previous year at Inria Lille financed by the FWF, working on the precursor project "Structure in reinforcement learning" (FWF:J 3259-N13). Most of the key publications emerging from the precursor project are collaborations with people from SequeL. These publications are also the starting point for the proposed project, so that it is of course planned to continue this successful cooperation.

Also, the other partners in the current EC project CompLACS are candidates for closer collaboration. This holds in particular for the Intelligent Autonomous Systems Lab of TU Darmstadt, which does research in development of learning models and algorithms for robotic systems. While the focus of our project is on theoretical results, we still think that the ideas developed can be useful for the real world problems dealt with at TU Darmstadt. Thus, it is planned that the candidate filling the project position will not only spend some time at SequeL Lille but also at TU Darmstadt to learn about the specific problems of applying the theoretical findings.

8. Project Duration, Personnel, Research Plan, and Dissemination

In the best case, the project position shall be filled with a PostDoc who shall be able to make progress on the proposed topics within two years, starting late 2013 or early 2014. The requested funding below is based on this best case. However, since experience from other projects has shown that it is often difficult to attract suitable PostDocs to Leoben, I would like to leave open the option of hiring a very good PhD student with sufficient previous knowledge in reinforcement learning instead. Due to the lower salary, a PhD student then could work for three years on the project, which is enough time to finish a PhD and also for catching up in knowledge in comparison with a PostDoc. In any case, the position will be advertised internationally to find the best possible candidate for the project.

I propose to start working on selection of approximate state representations, as I think that here first results can be achieved quickest. Otherwise, the topics can be worked on rather independently, so that there is no necessary order in which they should to be addressed. Thus, it will also depend on the preferences of the project employee what will be done first. Before half-time of the project when first results are available, the project employee shall visit Inria Lille and TU Darmstadt, cf. the previous section on collaborations.

Concerning the dissemination of the results, depending on the kind of results these will be first submitted to suitable conferences like COLT, ICML, or NIPS in order to make findings available as soon as possible. We also aim for journal publication of extended versions of suitable material in high quality peer reviewed journals.

9. Requested Funding

The requested funding consists of the cost for a PostDoc (two years) and the travel costs for a stay of three weeks at Inria Lille and TU Darmstadt, the latter calculated (and rounded down) according to the rates of the RGV (Reisegebührenvorschrift). Travel costs for conference visits are to be paid from the 5% general costs added.

(in Euro)	first year	second year
PostDoc (40h)	60,610.–	60,610.–
travel costs (Inria Lille)	900.–	
travel costs (TU Darmstadt)	1,000.–	
subtotal	62,510.–	60,610.–
Total incl. 5% general costs	129,276.–	

References

[1] Y. Abbasi-Yadkori and C. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. *COLT 2011, JMLR Proceedings Track*, 19:1–26, 2011.

[2] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *colt2009. Proc. 22nd Annual Conference on Learning Theory*, pages 217–226, 2009.

[3] P. Auer and R. Ortner. Logarithmic online regret bounds for reinforcement learning. In *Adv. Neural Inf. Process. Syst. 19*, pages 49–56, 2007.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256, 2002.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.

[6] P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007*, pages 454–468, 2007.

[7] P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 25–42, 2009.

[8] A. Bernstein and N. Shimkin. Adaptive-resolution reinforcement learning with polynomial exploration in deterministic domains. *Mach. Learn.*, 81(3):359–397, 2010.

[9] D. P. Bertsekas and D. A. Castañon. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE Trans. Autom. Control*, 34(6):589–598, 1989.

[10] R. I. Brafman and M. Tennenholtz. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

[11] E. Brunskill, B. R. Leffler, L. Li, M. L. Littman, and N. Roy. Provably efficient learning with typed parametric models. *J. Mach. Learn. Res.*, 10:1955–1988, 2009.

[12] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization of χ-armed bandits. In *Adv. Neural Inf. Process. Syst. 22*, pages 201–208, 2010.

[13] L. Buşoniu, B. D. Schutter, and R. Babuška. Approximate dynamic programming and reinforcement learning. In *Interactive Collaborative Information Systems*, pages 3–44. 2010.

[14] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Math. Oper. Res.*, 22(1):222–255, 1997.

[15] L. Busoniu and R. Munos. Optimistic planning for markov decision processes. *JMLR Proceedings Track*, 22:182–189, 2012.

[16] T. Dean, R. Givan, and S. M. Leach. Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In *UAI '97: Proc. 13th Conference*

on Uncertainty in Artificial Intelligence, pages 124–131, 1997.

[17] C. Diuk, L. Li, and B. R. Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proc. 26th Annual International Conference on Machine Learning, ICML 2009*, page 32, 2009.

[18] 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.

[19] N. Ferns, P. Panangaden, and D. Precup. Metrics for finite Markov decision processes. In *UAI '04, Proc. 20th Conference in Uncertainty in Artificial Intelligence*, pages 162–169, 2004.

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

[21] P. I. F. I. Ryzhov, W. B. Powell. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1), 2012.

[22] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

[23] N. K. Jong and P. Stone. Model-based exploration in continuous state spaces. In *Abstraction, Reformulation, and Approximation, 7th International Symposium, SARA 2007*, pages 258–272. Springer, 2007.

[24] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Gap-free bounds for stochastic multi-armed bandit. In *17th World IFAC Congress*, pages 11560–11563, 2008.

[25] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning: A survey. *J. Artif. Intell. Res.*, 4:237–285, 1996.

[26] S. Kakade, M. J. Kearns, and J. Langford. Exploration in metric state spaces. In *Machine Learning, Proc. 20th International Conference, ICML 2003*, pages 306–312, 2003.

[27] S. M. Kakade. *On the Sample Complexity of Reinforcement Learning.* PhD thesis, University College London, 2003.

[28] M. A. S. Kamal and J. Murata. Reinforcement learning for problems with symmetrical restricted states. *Robot. Auton. Syst.*, 56(9):717–727, 2008.

[29] M. J. Kearns and S. P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49:209–232, 2002.

[30] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Adv. Neural Inf. Process. Syst. 17*, pages 697–704, 2005.

[31] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proc. 40th Annual ACM Symposium on Theory of Computing, STOC 2008*, pages 681–690, 2008.

[32] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6:4–22, 1985.

[33] T. Lattimore and M. Hutter. Pac bounds for discounted mdps. In *Proc. 23rd Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

[34] B. R. Leffler, M. L. Littman, and T. Edmunds. Efficient reinforcement learning with relocatable action models. In *Proc. 22nd AAAI Conference on Artificial Intelligence*, pages 572–577, 2007.

[35] L. Li. *A Unifying Framework for Computational Reinforcement Learning Theory.* PhD thesis, Rutgers University, 2009.

[36] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proc. 9th International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

[37] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl. Knows what it knows: a framework for self-aware learning. *Mach. Learn.*, 82(3):399–443, 2011.

[38] O.-A. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Adv. Neural Inf. Process. Syst. 24*, pages 2627–2635, 2012.

[39] O.-A. Maillard, P. Nguyen, R. Ortner, and D. Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *JMLR Proceedings Track 28 : Proc. 30th International Conference on Machine Learning*, pages 543 – 551, 2013.

[40] S. Mannor, I. Menache, A. Hoze, and U. Klein. Dynamic abstraction in reinforcement learning via clustering. In *Machine Learning, Proc. 21st International Conference, ICML 2004*, 2004.

[41] R. A. McCallum. *Reinforcement Learning with Selective Perception and Hidden State.* PhD thesis, Department of Computer Science, University of Rochester, 1996.

[42] R. Munos. Approximate dynamic programming. In O. Sigaud and O. Buffet, editors, *Markov Decision Processes in Artificial Intelligence*, chapter 3, pages 67–98. 2010.

[43] S. M. Narayanamurthy and B. Ravindran. On the hardness of finding symmetries in Markov decision processes. In *Machine Learning, Proc. 25th International Conference, ICML 2008*, pages 688–695, 2008.

[44] G. Neumann, M. Pfeiffer, and W. Maass. Efficient continuous-time reinforcement learning with adaptive state graphs. In *Machine Learning: ECML 2007, 18th European Conference on Machine Learning*, pages 250–261, 2007.

[45] P. Nguyen, O.-A. Maillard, D. Ryabko, and R. Ortner. Competing with an infinite set of models in reinforcement learning, 2013. accepted for AISTATS 2013.

[46] A. Nouri and M. L. Littman. Multi-resolution exploration in continuous spaces. In *Adv. Neural Inf. Process. Syst. 21*, pages 1209–1216, 2009.

[47] R. Ortner. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. *Ann. Oper. Res.*, 2012. URL `doi:10.1007/s10479-12-1064-y`. doi:10.1007/s10479-12-1064-y, to appear.

[48] R. Ortner. Pseudometrics for state aggregation in average reward Markov decision processes. In *Proc. 18th ALT*, pages 373–387, 2007.

[49] R. Ortner. Online regret bounds for Markov decision processes with deterministic transitions. *Theor. Comput. Sci.*, 411(29–30):2684–2695, 2010.

[50] R. Ortner and D. Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Adv. Neural Inf. Process. Syst. 25*, pages 1772 – 1780, 2012.

[51] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless Markov bandits. In *Proc. 23rd Conference on Algorithmic Learning Theory, ALT 2012*, pages 214–228, 2012.

[52] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless Markov bandits, 2013. submitted to Theoretical Computer Science.

[53] D. C. Parkes and A. D. Procaccia. Dynamic social choice: Foundations and algorithms. Unpublished typescript, August 2010.

[54] M. Post and O. Junge. Exploiting symmetries in two player zero-sum Markov games with an application to robot soccer. Unpublished typescript.

[55] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

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

[57] Z. Ren and B. H. Krogh. State aggregation in Markov decision processes. In *Proc. 41st IEEE Conference on Decision and Control, Volume 4*, pages 3819–3824. IEEE, 2002.

[58] B. V. Roy. Performance loss bounds for approximate value iteration with state aggregation. *Math. Oper. Res.*, 31(2):234–244, 2006.

[59] S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. In *Machine Learning, Proc. 11th International Conference, ICML 1994*, pages 284–292, 1994.

[60] A. L. Strehl and M. L. Littman. A theoretical analysis of model-based interval estimation. In *Machine Learning, Proc. 22nd International Conference, ICML 2005*, pages 857–864, 2005.

[61] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *J. Comput. System Sci.*, 74(8):1309–1331, 2008.

[62] A. L. Strehl and M. L. Littman. Online linear regression and its application to model-based reinforcement learning. In *Adv. Neural Inf. Process. Syst. 20*, pages 1417–1424, 2008.

[63] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. Pac model-free reinforcement learning. In *Machine Learning, Proc. 23rd International Conference, ICML 2006*, pages 881–888, 2006.

[64] A. L. Strehl, C. Diuk, and M. L. Littman. Efficient structure learning in factored-state MDPs. In *Proc. 22nd AAAI Conference on Artificial Intelligence*, pages 645–650, 2007.

[65] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[66] I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proc. 27th International Conference on Machine Learning (ICML-10)*, pages 1031–1038, 2010.

[67] J. Taylor, D. Precup, and P. Panangaden. Bounding performance loss in approximate MDP homomorphisms. In *Adv. Neural Inf. Process. Syst. 21*, pages 1649–1656, 2009.

[68] A. Tewari and P. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Adv. Neural Inf. Process. Syst. 20*, pages 1505–1512. 2008.

[69] J. N. Tsitsiklis and B. Van Roy. Feature-based methods for large scale dynamic programming. *Mach. Learn.*, 22(1–3):59–94, 1996.

[70] W. T. B. Uther and M. M. Veloso. Tree based discretization for continuous state space reinforcement learning. In *Proc. 15th National Conference on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98*, pages 769–774, 1998.

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

[72] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, 1989.

[73] M. Zinkevich and T. R. Balch. Symmetry in Markov decision processes and its implications for single agent and multiagent learning. In *Proc. 18th International Conference on Machine Learning, ICML 2001*, pages 632–639, 2001.