STAND-ALONE PROJECT

FINAL REPORT

P 26219-N15

Project number

Project title	Structured and Continuous Reinforcement Learning Strukturiertes und kontinuierliches Verstärkungslernen
Project leader	**Ronald Ortner**
Project website	**http://personal.unileoben.ac.at/rortner/fwf.htm**

I. Summary for public relations work

1. Zusammenfassung für die Öffentlichkeitsarbeit

Im sogenannten *Verstärkungslernen* (engl. *reinforcement learning*) hat ein Lerner die Aufgabe, in einer ihm unbekannten Umgebung optimales Verhalten zu erlernen. Dies kann z.B. das Erreichen eines bestimmten Ziels oder das Ausführen einer komplexen Aufgabe sein. Der Lernprozess wird dabei allein durch das Feedback der Umgebung gesteuert. D.h. der Lerner kann die Reaktion der Umgebung auf sein Verhalten beobachten und erhält etwa für das erfolgreiche Absolvieren einer Aufgabe eine Belohnung in Form einer Auszahlung. Da eine solche Auszahlung auch verzögert erfolgen kann, beispielsweise erst nach einer koordinierten Abfolge von bestimmten Aktionen, um ein Ziel zu erreichen, muss der Lerner oft kurzfristig negative Auszahlungen in Kauf nehmen, um nach erfolgreicher Absolvierung einer Aufgabe mit einer hohen positiven Auszahlung belohnt zu werden.

Für diese Problemstellung gibt es zwar bereits einige Lernalgorithmen, für die gezeigt werden kann, dass sie im Prinzip jede Aufgabenstellung nach einer gewissen Zeit lösen können (sofern die Aufgabenstellung gewisse Bedingungen erfüllt), in der Praxis sind diese Algorithmen aber meist nicht einsetzbar, da die computergerechte Darstellung praktischer Problemstellungen so komplex ist, dass Lernalgorithmen selbst für einfache Aufgaben viel zu lange zur Lösung benötigen würden.

Im vorliegenden Projekt wurden u.a. Lernalgorithmen für kontinuierliche Zustandsräume entwickelt, die in der Praxis höchst relevant sind, für die es bisher aber kaum theoretische Ergebnisse gab. Dabei konnte ein neuer Algorithmus entwickelt werden, für den in Aufgabenstellungen in Umgebungen, die sich „brav“ verhalten, gezeigt werden kann, dass er schneller lernt als bisher bekannte Algorithmen.

Weiters beschäftigte sich das Projekt mit der Frage, inwiefern Algorithmen einfachere Darstellungen einer Aufgabenstellung während des Lernprozesses selbst erlernen und verwenden können. Dabei hat der Algorithmus mehrere solcher Darstellungen zur Auswahl, weiß aber nicht, welche für das Erreichen seines Lernziels am besten geeignet ist. Im Rahmen des Projekts konnte gezeigt werden, dass Algorithmen auch dann erfolgreich lernen können, wenn keine der Darstellungen völlig korrekt ist, und es nur mindestens eine gute Approximation der Lernumgebung gibt. Dabei ist es für den Lernerfolg interessanterweise gar nicht nötig, dass der Lernalgorithmus diese „gute“ Darstellung identifiziert. Tatsächlich ist dies im allgemeinen viel schwieriger und in manchen Fällen gar unmöglich.

2. Summary for public relations work

In reinforcement learning a learner wants to learn optimal behavior in an unknown environment. For example, the goal of the learner could be to reach a certain location or state, or to solve a complex task. The learning process itself is governed only by feedback of the environment. That is, the learner can observe the reaction of the environment to his actions and e.g. obtains a reward for solving a given task. Since the solution of a task may require the execution of a longer sequence of coordinated actions, the learner must be able to learn also from delayed feedback, for example by accepting short-term discouraging feedback to achieve a long-term goal giving high reward.

Problem settings of this kind are in principle solvable by existing reinforcement learning algorithms, which can even be shown theoretically to be able to solve any task, provided that the task has certain properties (like that it is possible to recover from mistakes). However, at the same time these algorithms are hardly applicable to real world problems. This is mainly due to the fact that the representation of even the simplest problems gives rise to huge state spaces, so that algorithms cannot solve these problems in reasonable time.

In the project at hand we managed to develop reinforcement learning algorithms for problems with continuous state space, which are of particular importance in the context of applications but for which there have been only few theoretical results available so far. It could be shown that in environments that behave „nicely“ the new algorithm can provably learn faster than known algorithms.

Another question that was dealt with in the project was whether a learning algorithm can learn to use simpler representations in the learning process. More precisely, the learner is given a set of possible representations, some of which are suitable, while others can even be misleading. In this setting it could be shown that a learning algorithm developed in the project can successfully learn even if there is no completely correct representation at its disposal. Instead, it is sufficient that there is at least one representation that is a good approximation of the environment. It is particularly interesting that for successful learning it is not necessary to identify this representation, which can be more difficult and sometimes is even impossible.

II. Brief project report

1. Report on research work

1.1 Information on the development of the research project

The original project proposal is still available on the project website http://personal.unileoben.ac.at/rortner/fwf/Proposal-FWF-P26219-N15.pdf

As described in the proposal, the main goals were to make progress on reinforcement learning (RL) in continuous domains on the one hand and on learning state representations on the other hand. In both settings, we were particularly interested in strong theoretical guarantees for the algorithms to be developed. More precisely, we aimed at improving known bounds on the *regret* that compares the performance of a learning algorithm to the optimal policy after any T steps. Concerning the representation learning setting, another important topic was to generalize the setting to the case were the set of representations given to the learner does not contain any Markov model.

Beyond these two main topics we were interested in generalizations where structure helps to improve the learning process. Here due to personal interests of the involved personnel, in particular of Chao-Kai Chiang, who was the employed PostDoc in the second year of the project, research headed in a slightly different direction with an emphasis on bandit problems, yet leading to nice results in the investigated settings (see below) that fit the project topic of structure and reinforcement learning quite well, yet with a different flavor.

1.2 Most important results and brief description of their significance (main points)

At the beginning of the project, we concentrated on the topic of selecting state representations where the learner does not have direct access to the state space, but is only given a set Φ of some possible state representation models. Here we were able to improve and generalize the known bounds as follows: First, we managed to drop the assumption that the model set Φ contains a model that induces a Markov decision process (MDP). Instead, we considered models that approximate an MDP. Thus, we have not only introduced a suitable notion of *ε-approximate state representations* but also improved over previous bounds on the error made by (MDP) ε-approximations. These bounds show that when working with an ε-approximate model one has to accept a (per-step) error of up to εD, where D is the diameter of the true underlying MDP. This is reflected in the regret bounds we have derived in this setting. Our suggested algorithm is able to deal with approximate models by estimating for each model its approximation error. Furthermore, in case Φ contains a model inducing an MDP, we were able to improve the previously known bounds with respect to the dependence on the state space S to attain bounds of order $|S|(|\Phi|\ T)^{1/2}$, where T is the horizon, i.e., the number of steps taken by the algorithm. These results (published as [6][1]) lay an important foundation for further work in reinforcement learning with huge state spaces. In particular, the extension to approximate representations make the techniques applicable for aggregation approaches and a respective application to discretization in continuous environments seems reachable as well. Furthermore, the improvement of the bounds with respect to the size of the state space shows that it is likely that the known bounds for countable sets of representations can be improved as well.

Soon afterwards, based on an idea of the project PostDoc K.Lakshaman and in collaboration with Daniil Ryabko of Inria Lille, we could improve the known regret bounds for reinforcement learning in continuous state space, in case when transition probabilites are sufficiently smooth. Using non-parametric kernel density estimation (instead of histograms in a discretized state space) for estimating the transition probability distributions, we were able to derive regret bounds that depend on the smoothness of the transition probability distributions. In particular, under the assumption that the transition probability functions are smoothly differentiable, the regret bound can be improved from the previously known best bound of $O(T^{3/4})$ to $O(T^{2/3})$ for reinforcement learning in *1*-dimensional state space and horizon T. For reinforcement learning in higher dimensional state space we were able to achieve similar improvements. These results (published and presented at ICML 2015 [5]) are not only the best known bounds for continuous RL with smooth transition functions, they also are the first regret bounds for a kernel-based method in RL. Moreover, the paper also gives some new concentration bounds for kernel-estimation with non-iid random variables which are of independent interest.

In the second year, Chao-Kai Chiang (a former PostDoc in our EC project CompLACS) took over the position of K.Lakshmanan. Beside starting to work on regret bounds for a model-free RL algorithm (see below) he also continued some work with Peter Auer on finding an algorithm that works for stochastic as well as for non-stochastic bandit problems. The resulting paper not only complements previously known regret bounds, but also settles an open question regarding the general trade-off in regret between the stochastic and the non-stochastic setting for any possible algorithm. That is, no algorithm that generally obtains $O(\log T)$ regret after any T steps in the stochastic setting can also achieve $O(T^{1/2})$ regret in non-stochastic settings. These results have been accepted and presented at COLT 2016 [1].

[1] For references see Section 1 in the Attachments III.

Inspired by a visit of Madalina Drugan (now at TU Eindhoven) at our institute in autumn 2014, we also got interested in the structure of multi-objective bandits, the result of which was a paper that determines the sample complexity for the identification of the Pareto front in such problems. The obtained lower and the upper bounds depend on some natural geometric properties of the considered point set. The paper was published and presented at AISTATS 2016 [2].

In cooperation with (current and former) members of Inria Lille, we also improved the known sample complexity bounds for pure exploration in combinatorial bandits in the fixed confidence as well as in the fixed budget setting, exploiting some combinatorial properties and introducing a new structural parameter in the bounds. The respective paper was published and presented at AISTATS 2016 as well [3].

While these latter settings of multi-objective reinforcement learning and combinatorial bandits were not directly considered in the original project proposal, I think they fit the topic of the project on structure and reinforcement learning quite nicely.

Last but not least, together with Odalric-Ambrym Maillard (now at Inria Saclay, France), Shie Mannor and Timothy Mann (both at Technion) we extended reinforcement learning to an active learning setting where the learner can ask for additional information in the form of roll-outs of a policy. This shall allow learning in general multi-chain MDPs with potentially infinite diameter, where without access to some external information, it is in general impossible to converge to an optimal policy. We introduced the notion of *active regret*, where each step in a requested roll-out contributes the same to the regret as a step in which an action is chosen but were no reward is gained. In this setting, we derived bounds on the active regret that resemble bounds for standard RL showing that the additional overhead caused by roll-outs is in general negligible. A paper containing the results is in preparation [7].

There is also a lot of work that is still ongoing, some of which is intended to be turned into a paper within the next few months, while more challenging topics will be the topic of a follow-up project proposal soon to be submitted to FWF. Here we just briefly mention a few research directions. First, visits to Inria Lille in September 2015 and April 2016 gave the opportunity to continue the work with Daniil Ryabko on selecting state representations and the restless bandit problem. Concerning the former, there are some ideas how to extend the current results to the setting of countable sets of representations obtaining not only an improved regret rate of $O(T^{1/2})$ but also avoiding the large additive constants of previous work by choosing a more sophisticated penalization term in the algorithm. With respect to the restless bandit problem, we think that application of the results in the state representation setting can not only help to improve currently known bounds but also to generalize results from the Markov setting to a setting were it is only assumed that the underlying processes are mixing. We also work on the development of an efficient algorithm for computation of the optimal policy in an MDP under the constraint that the bias is bounded by some constant. This is of importance for our RL algorithms in continuous state space, which rely on such a sub-procedure. Finally, together with Peter Auer we have started to develop a model-free general reinforcement algorithm with the goal of extending it to continuous domains without relying on some discretization of the state space.

1.3 Information on the execution of the project, use of available funds and (where appropriate) any changes to the original project plan relating to the following:

The project started in April 2014 and ended in May 2016. The major part of the available funds was used as planned for funding a PostDoc position for 2 years. Unlike originally planned, the position was split up (see next section for details). The remaining funding has been used as planned for covering travel costs for visits to the collaborating institute Inria Lille, France. The planned visit (of Chao-Kai Chiang) to TU Darmstadt also took place, but was funded by the CompLACS project.

2. Personnel development – Importance of the project for the research careers of those involved (including the project leader)

For our small department, projects like the FWF project are vital to attract young reseachers to the institute. The Montanuniversität Leoben is a university of technology with specific curricula concentrating on mining, metallurgy, and materials. Thus, there are no mathematics or computer science students so that young academics always have to be attracted from outside. For me personally, this was the first project with employees under my supervision and hence an important experience.

Concerning the project employees, Chao-Kai Chang has been hired for a PostDoc position funded by the European CompLACS project (the respective work package led by the department head Peter Auer) beginning of 2014. Around the same time K.Lakshmanan applied for the project position and we decided to offer K.Lakshmanan a one year contract (hoping to get some additional funding for a second year) and giving Chao-Kai Chiang the option to stay for a second year at the institute funded by the project after termination of the CompLACS project in Februray 2015 (bridging the few weeks in between by money from the university).

K.Lakshmanan has already been working on reinforcement learning problems as a PostDoc at IIT-Bombay after obtaining a PhD from IISc Bangalore under supervision of Shalabh Bhatnagar. During the first months of the project K.Lakshmanan quickly familiarized himself with the theoretical foundations of regret bounds for reinforcement learning and had the idea to employ nonparametric estimation techniques to improve the currently known regret bounds for reinforcement learning in continuous state space. In May 2015 K.Lakshmanan joined the National University of Singapore as a PostDoc.

Chao-Kai Chiang took over the project position in April 2015. He received his PhD at National Taiwan University (NTU) in 2014 and then worked as a PostDoc for a year at NTU and Academia Sinica, before joining the CompLACS project. His original research interests were more directed towards bandit settings and while he also started to work on RL in MDP settings, he kept working on bandit problems, luckily also involving Peter Auer more into the project, which resulted in the COLT paper about a common algorithm for stochastic and nonstochastic bandit problems. Chao-Kai Chiang has joined the University of California in Los Angeles in the meantime.

Last but not least, Madalina Drugan (now at Technische Universiteit Eindhoven) should be mentioned, who visited in autumn 2014 and initiated our interest in multi-criteria RL, which resulted not only in a nice paper but also in an application for a Marie Curie scholarship to fund a longer stay at our institute.

3. Effects of the project beyond the scientific field

Some of the project results concerning representation learning seemed to be interesting in the context of philosophy of science. An interdisciplinary paper [4] presenting these results to an audience outside computer science has been published in *Minds and Machines*, a journal that is at the intersection of artificial intelligence and philosophy and deals with philosophical aspects of computer science.

4. Other important aspects

Examples:

- Project-related participation in national and international scientific / scholarly conferences, list of most important lectures held;
- Organisation of symposiums and conferences;
- Prizes/awards;
- Any other aspects.

The results of the project were not only presented at the venues of the conferences of the respective accepted papers (i.e, ALT 2014, ICML 2015, COLT 2016, and AISTATS 2016), additionally the following talks on results of the project were given at the following workshops:

- R.Ortner: Selecting Near-Optimal Approximate State Representations in Reinforcement Learning, Large-scale Online Learning and Decision Making Workshop (LSOLDM) 2014, Windsor, UK, 11.09.2014
- R.Ortner: Selection of State Representations and Aggregations in Reinforcement Learning, NIPS Workshop: From Bad Models to Good Policies (Sequential Decision Making under Uncertainty), Montreal, Canada, 12.12.2014
- R.Ortner: Regret Bounds for Optimistic Algorithms in Multi-armed Bandits and MDPs, Machine Learning Summer School Chalmers 2015. Göteborg, Sweden, 30.04.2015

In March 2015 K. Lakshmanan gave two talks at National University of Singapore and Tata Research TRDCC on "Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning". Last but not least, we also presented a poster introducing the project at *WerWasWo*, an event of the university where institutes and researchers can present their work to the public.

III. Attachments

1. Scholarly / scientific publications

Publications may only be listed if they relate directly to the project. **Up to three of the most important publications** should be highlighted.

Please indicate at the end of every peer-review publication (in brackets) the Open Access (OA) type as following:

- Gold OA = published in Open Access Journal, with or without an author fee (see register of all Open Access Journals http://www.doaj.org/)
- Hybrid OA = published in a subscription journal but Open Access by an author fee (see http://en.wikipedia.org/wiki/Hybrid_open_access_journal)
- Green OA = self-archived electronic copy of the final "accepted manuscript" which might include an embargo period (see: http://www.fwf.ac.at/en/research-funding/open-access-policy/)
- Other OA = any other type of Open Access
- No OA = not published Open Access

1.1 ***Peer-reviewed publications / already published*** (journals, monographs, anthologies, contributions to anthologies, proceedings, research data, etc.)

►[1] Peter Auer, Chao-Kai Chiang: An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits, JMLR Workshop and Conference Proceedings Volume 49: Proceedings of the 29th Conference on Learning Theory, COLT 2016, pp. 116-120.
http://jmlr.org/proceedings/papers/v49/auer16.pdf (Gold OA)

[2] Peter Auer, Chao-Kai Chiang, Ronald Ortner, Madalina Drugan: Pareto Front Identification from Stochastic Bandit Feedback, JMLR Workshop and Conference Proceedings Volume 51: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016.
http://jmlr.org/proceedings/papers/v51/auer16.pdf (Gold OA)

[3] Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner, Peter Bartlett: Improved Learning Complexity in Combinatorial Pure Exploration Bandits, JMLR Workshop and Conference Proceedings Volume 51: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016.
http://jmlr.org/proceedings/papers/v51/gabillon16.pdf (Gold OA)

[4] Ronald Ortner: Optimal Behavior is Easier to Learn than the Truth, *Minds and Machines* 26(3), 243-252.
DOI 10.1007/s11023-016-9389-y
http://link.springer.com/content/pdf/10.1007%2Fs11023-016-9389-y.pdf (Gold OA)

►[5] K.Lakshmanan, Ronald Ortner, and Daniil Ryabko: Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning, JMLR Workshop and Conference Proceedings Volume 37: Proceedings of The 32nd International Conference on Machine Learning, ICML 2015.
http://jmlr.org/proceedings/papers/v37/lakshmanan15.pdf (Gold OA)

►[6] Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko: Selecting Near-Optimal Approximate State Representations in Reinforcement Learning, Proceedings of the 25th International Conference on Algorithmic Learning Theory, ALT 2014. Lecture Notes in Computer Science 8776, Springer 2014, pp. 140-154. http://personal.unileoben.ac.at/rortner/Pubs/ApproxStateRep.pdf (Green OA)

1.2 ***Non peer-reviewed publications / already published*** (journals, monographs, anthologies, contributions to anthologies, research reports, working papers / preprints, proceedings, research data, etc.)

There were no project publications that were not peer-reviewed.

1.3 ***Planned publications***

(journals, monographs, anthologies, contributions to anthologies, proceedings, research data, etc.)

[7] Odalric-Ambrym Maillard, Timothy A.Mann, Ronald Ortner, and Shie Mannor: Active Roll-outs in MDP with Irreversible Dynamics, in preparation (to be submitted to Journal of Machine Learning Research)

2. Most important academic awards

There were no awards related to the project.

3. Information on results relevant to commercial applications

There were no patents, copyrights or other types of commerical applications of results of the project (which was not to be expected from a theory project).

4. Publications for the general public and other publications

A poster has been presented introducing the project at *WerWasWo*, an event of the university where institutes and researchers can present their work to the public.

	national	**International**
Self-authored publications on the www		
Editorial contributions in the media		
(Participatory) contributions within science communication		
Popular science contributions	1	

5. Development of collaborations

Indication of the most important collaborations (no more than 5) that took place (i.e. were initiated or continued) in the course of the project. Please provide the name of the collaboration partner (name, title, institution) and a few words about the scientific content. Please **categorise** each collaboration arrangement as follows:

N				Nationality of collaboration partner (please use the ISO-3-letter country code)
	G			Gender **F** (female) **M** (male)
		E		Extent **E1** **low** (e.g. no joint publications, but mention in acknowledgements or similar); **E2** **medium** (collaboration e.g. with occasional joint publications, exchange of materials or similar, but no longer-term exchange of personnel); **E3** **high** (extensive collaboration with mutual hosting of group members for research stays, regular joint publications, etc.)
			D	Discipline **W** **within the discipline** (within the same scientific field) **I** **interdisciplinary** (involving two or more disciplines) **T** **transdisciplinary** (collaborations outside the sciences)

Note: General scientific contact and occasional meetings should not be considered collaborations for the purposes of this report.

N	G	E	D	Name	Institution
RUS	M	E3	W	Daniil Ryabko, Fellow	Inria Lille – Nord Europe, FRA
ROU	F	E2	W	Madalina Drugan, Dr	TU Eindhoven, NLD
FRA	M	E2	W	Odalric-Ambrym Maillard, Fellow	INRIA Saclay – Île-de-France, FRA
FRA	M	E2	W	Victor Gabillon, Dr	Queensland University of Technology, AUS
DEU	M	E1	I	Jan Peters, Prof.	TU Dortmund, DEU

6. Development of human resources in the course of the project

There were two PostDocs employed during the project, who left the institute after the end of the project and now work as PostDocs at other universities.

	In progress	Completed	Gender	
			f	m
Full professorship				
Venia thesis (*Habilitation*) / Equivalent senior scientist qualification				
Postdoc	2			m
Ph.D. theses				
Master's theses				
Diploma theses				
Bachelor's theses				

7. Applications for follow-up projects

7.1 Applications for follow-up projects (FWF projects)

Please indicate the project type (e.g. stand-alone project, SFB, DK, etc.)

Project number (if applicable)	
Project type	stand-alone
Title / subject	Exploration and Reinforcement Learning in Continuous Environments

Status	granted □	pending □	in preparation ☒
Application reference (if a patent is applied)			

7.2 Applications for follow-up projects (Other national projects)

none

7.3 Applications for follow-up projects (international projects)

Madalina Drugan with whom we collaborated during the project has recently applied for a Marie Curie fellowship to fund a longer stay at our institute.

Country			
Funding agency	Please choose an item: EU ▼		
Project number (if applicable)			
Project type	Marie Curie fellowship		
Title / subject	Reinforcement learning and evolutionary computation for scalable learning and optimization algorithms		
Status	granted □	pending ☒	in preparation □
Total costs (granted)			

IV. Cooperation with the FWF

Please rate the following aspects with regard to your interaction with the FWF. Please provide any **additional comments (explanations)** on the supplementary sheet with a reference to the corresponding question/aspect.

Scale:

-2	highly unsatisfactory
-1	unsatisfactory
0	appropriate
+1	satisfactory
+2	highly satisfactory
X	not used

Rules
(i.e. guidelines for: funding programme, application, use of resources, reports)

Rating

Application guidelines	Length	**+1**
	Clarity	**+1**
	Intelligibility	**+1**

Procedures (submission, review, decision)

	Advising	**+2**
	Duration of procedure	**+2**
	Transparency	**+1**

Project support

Advising	Availability	**+2**
	Level of detail	**+2**
	Intelligibility	**+1**

Financial transactions (credit transfers, equipment purchases, personnel management)	**+2**

Reporting / review / exploitation

	Effort	**0**
	Transparency	**+1**

Support in PR work / exploitation	X

Comments on cooperation/interaction with the FWF:

Just a minor suggestion: In the section on collaborations, if the nationality of the involved researchers is that important, I guess it should be ok to give the complete nationality to spare people looking up ISO country codes.