FROM SATISFICING TO OPTIMIZATION IN ONLINE REINFORCEMENT LEARNING

PROPOSAL FOR AN FWF PROJECT

RONALD ORTNER

Contents

Scientifc Abstract

1. Wider research context

Reinforcement learning (RL) has been successful in many applications, but theory has not been able to guarantee reliability and robustness of the used algorithms. In our opinion, one of the reasons for this is that RL theory focuses on optimization, while most RL problems in practice are task-oriented and optimality does not play any role. We aim at a restart of RL theory by replacing the optimality paradigm by a satisficing criterion.

2. Objectives

We expect that the paradigm shift from optimization to satisficing will alleviate the development and analysis of algorithms. Based on approaches and results of a precursor project, we aim at continuing the development of an alternative RL theory for satisficing. The goal is to provide satisficing algorithms for which we can prove that these are more efficient in terms of (sample) complexity than RL algorithms that aim for optimal behavior.

3. Approach

For a start we will stick to the approach of the precursor project, where we considered regret as performance measure and showed that satisficing regret (with respect to a given satisfaction level) is significantly smaller than standard regret (with respect to the optimum). While the respective performance bounds achieved so far are independent of the horizon, they still depend on the size of the problem (i.e., number of arms in the bandit setting, size of state-action space for MDPs). As a next step we want to replace the size by more specific problem dependent parameters, also considering more suitable alternative measures that unlike regret are not worst-case based.

4. Innovation

The results of the precursor project have shown that an achievable satisfaction level provides an ‘Archimedean point’ that allows to learn with only constant regret, something that is not possible within the classic optimization setting. Accordingly, although satisficing is a concept that has hardly been seriously considered in the context of RL, we believe in its potential to provide a powerful alternative paradigm for RL.

5. Primary researchers involved

Ronald Ortner (PI)

1. Introduction

The standard model for reinforcement learning (RL) problems is the Markov decision process (MDP) setting. Originally introduced by Richard Bellman in the 1950s, the operations research community developed algorithms and theory for MDPs, interested in the case when an MDP is explicitly given. In the 1980s, the AI community adopted the MDP framework as model for learning in an unknown environment described by an MDP, whose parameters now have to be learned by the learning agent. Theory and algorithms for MDPs that have been provided by operations research came in handy, as these could be also be used in the reinforcement learning context. The paradigm of optimization, that is, having the goal of maximizing reward, was naturally adopted and not further challenged by AI researchers.

In this project we are particularly interested in this aspect: What changes when we weaken the goal of optimality and are happy with algorithms that perform just good enough instead of optimally? How big is the gain with respect to complexity and performance? These questions have already been addressed in a precursor project (cf. Sec. 4 below), which could give some answers but of course generated new questions that we want to investigate in this project.

In brief, we could show that in satisficing problems one can have performance guarantees that are not achievable in the classic optimization setting. This has been accomplished for the simple multi-armed bandit setting just as for the general MDP setting. What remains to be determined is the precise form of these performance bounds. That is, currently we have upper bounds for our algorithms and only prelimiary lower bounds that do not match. It is not even clear what are the parameters tight bounds will depend on. Beside closing this gap, an advanced question is to investigate the boundary between satisficing and optimization, in particular in the form of algorithms that approach optimization by repeated satisficing with adapting the satisfaction level.

2. Setting

Reinforcement learning (RL) is a broad field with many different approaches each having its own focus. We are interested in *online* reinforcement learning, where the learning agent is presented a learning environment and has to learn to perform well by learning through trial and error without having any prior knowledge. Learning tasks are considered to be non-episodic with no automatic resets to initial states, and the performance criterion will be total resp. average reward. In spite of the popularity of episodic RL in recent years, we consider this setting to be the most natural and suitable for generic RL tasks.

Since the origins of RL exciting applications (starting with Backgammon via autonomous driving to recent successes like Alpha-GO and partially also ChatGPT) have been developped in parallel with theoretical contributions which not only provided guarantees for existing approaches and showed general limits that cannot be overcome by any algorithm, but also provided various ideas that were applicable in practice. The first guarantees that were provided by RL theory were convergence results that showed that an algorithm would converge to an optimal solution in the long run (i.e., when the number of steps $T \to \infty$). While such results are valuable as a justification for the employed algorithms, they have little power in practice, as in any application the number of learning steps is limited. Even results providing a rate that promises fast (i.e., exponential) convergence cannot amend this in general, as the involved constants are often very large or even unknown.

Originally established in the simplest RL setting, the multi-armed bandit problem, so-called *regret bounds* consider an online learning setting, in which the performance of the learner is compared to that of an optimal policy after any finite number of steps T. Theoretical results that bound the regret can really quantify finite-time performance of an algorithm. On the one hand, upper bounds provide information about the worst case behavior of a specific algorithm, while lower bounds show that for some problem settings all algorithms have to respect a certain learning complexity that cannot be improved on.

For the general RL setting with an underlying MDP crucial results have been derived in [18], showing an upper bound on the regret of $\tilde{O}(DS\sqrt{AT})$ after any T steps for RL in an MDP with S states, A actions and diameter (i.e., the maximal expected transition time between two states) D. An accompanying lower bound was of order $\Omega(\sqrt{DSAT})$ [18]. Some improvements on the upper bound have been made in the meantime. [14] improved the dependence on the diameter to $\sqrt{D}$, while [15] showed that the diameter in the original bounds can be replaced by the bias span $H := \mathrm{span}(\lambda)$ of an optimal policy when H is known. Moreover, the algorithm also works more generally for weakly communicating MDPs. Under the same assumption [40] could show a best possible upper bound of $\sqrt{HSAT}$. The assumption of knowing the span could be removed recently in [7], finally closing the gap between upper and lower bound on the regret.

Within the proposed project we want to target the question whether and to what extent these bounds can be improved when considering *satsificing* instead of *optimization.* That is, instead of comparing to the optimal policy (as in the classic regret), we consider the loss with respect to a given satisfaction level σ, which we call *satisficing regret*, cf. Sec. 4 for details.

3. Related Work

3.1. **Satisficing for Bandits.** While satisficing objectives have been considered before, most respective investigations have been made in the simple bandit setting. Just like for the classic multi-armed bandit setting there are different questions considered.

The pure-exploration setting when the goal is to identify *all* arms above the satisfaction level is considered in [24, 28, 20, 5]. A related problem is to identify all ε-good arms for a given ε [25]. A variant that is closer to our setting is the problem of identifying one or more generally k arms among the top m arms. Sample complexity bounds for pure exploration are given in [9, 10]. In these settings one is interested in *sample complexity* or bounds on the *simple regret* for a given confidence or a certain sample budget. We note that in the online setting it is not sufficient to pick after an inital exploration phase an arm that is likely to be above the satisfaction level: Subsequent exploitation will suffer linear regret due to the small error probability that the chosen arm is not satisficing.

Concerning online regret bounds, following [21] which proposed a simple index policy for satisficing and evaluated it experimentally, [38] shows that the (classic) regret is finite if the satisfaction level is chosen to be between the reward of the best and the second-best arm.

The notion of *expected satisficing regret* introduced by [33] coincides with ours in a special case, but considers more generally a Bayesian setting that also takes into account the learner's belief that some arm is satisficing. Various bounds on the expected satisficing regret are given in [33], including a lower bound as well as upper bounds for problems with Gaussian reward distributions when using adaptations of the UCL algorithm [32], cf. Sec. 4 for a brief discussion.

A different kind of regret is suggested in [1]. It is assumed that the learner obtains a reward of 1 if the *actual* reward of the chosen arm is above a given threshold (which may vary with time), otherwise the reward is 0. The given regret bounds resemble those in the standard setting (i.e., are logarithmic in the horizon T) but with the mean rewards replaced by more parameters of the reward distributions.

The idea of satisficing has also been considered as a complementary feature of the classic multi-armed bandit setting. Under the notion of *conservative bandits*, the learner aims at learning the optimal arm but at the same time not falling below a baseline level. An algorithm for this setting is suggested in [39], which gives regret bounds similar to the classic setting and shows that with high probability the algorithm stays above the baseline.

A related notion is that of *lenient regret*, which considers the loss with respect to $\mu_* - \epsilon$ for some $\epsilon > 0$. The setting suggested in [26] is more general as different loss functions can be considered, but for a special case corresponds to the satisficing regret we propose yet not with respect to a given fixed level but defined via the allowed deviation ε from the optimal mean reward. Beside convergence guarantees for a variant of Thompson sampling it is shown that when $\mu_* > 1 - \epsilon$ the lenient regret is constant.

Curiously, the most relevant result has been derived in a different context, that is, in the setting of multi-objective bandit problems. In [17] regret bounds are derived what also provide results for the satisficing regret, the precise form of the respective bound is discussed in Sec. 4 below.

3.2. **Satisficing for MDPs.** For the general MDP setting there is not much work on satisficing. For robust MDPs the problem of finding a satisficing policy is considered in [34]. This approach not only provides an alternative to the usual worst-case analysis in this setting, it may also be relevant in the learning context for our project, cf. Sec. 6 below. For Q-learning a satisficing variant is presented in [16], for which however only some experimental evaluation is given.

A notion of satisficing regret is introduced in [4] for an episodic and Bayesian RL setting. It is however rather concerned with approximating an optimal policy and has little to do with satisficing in our context. Formalized using rate distortion theory, the paper proposes a generalization of the PSRL algorithm [37] and gives bounds on the classic (Bayesian) regret which generalize respective results for bandits [35, 2, 3] and in a special case correspond to those available for PSRL [29].

More relevant to our setting is work on constrained MDPs (CMDPs). Typically, the goal in the CMDP setting is to optimize a reward function as in standard MDPs, but on the other hand satisfy some cost constraint for an additional cost function. Concerning the latter one can consider the respective regret with respect to constraint violations, which basically corresponds to our regret with respect to a satisfaction level σ.

Most work on RL in CMDPs considers the discounted [12] or the episodic setting [31, 19, 23, 13]. However, there is some relevant work considering the average reward setting as we do. Usually, some strong assumptions are made, such as that transition probabilities are known [41], when it is easy to guarantee that constraints are fulfilled. Ergodic CMDPs are considered in [36], however the regret from constraint violations after T steps is of order $T^{2/3}$ and thus quite large.

The state of the art is [11], whose algorithm on ergodic MDPs has regret (with respect to the optimal average reward) of order $\sqrt{T}$ and a constant violation of the cost constraint. For weakly communicating MDPs, bounds of order $T^{2/3}$ are shown for regret as well as cost constraint violation. Both bounds are high probability bounds, but the considered reward and cost functions are assumed to be deterministic. For a brief discussion of these bounds see Sec. 5 below.

4. The Precursor Project

In the precursor project (TAI 590-N) we were able to come up with the first results that were as surprising as encouraging. First concentrating on the simpler bandit setting, using results of [8] we could show that when given a satisfaction level σ that can be met (that is, the mean reward of the optimal arm is above σ), then the (pseudo-)regret with respect to σ after any T steps is bounded by a constant which is independent of T. That is, the regret does not grow with the horizon, which is long known to be unavoidable in the standard (optimization) case [22]. Our results were complemented by a more general algorithm which can be shown to have constant regret if the requested satisficing level σ is below the optimal mean reward of an arm, while having standard logarithmic gap-dependent bounds otherwise. These results were presented at EWRL 2022 and have been published as [27]. Some preliminary improvements we have are discussed in Sec. 5 below.

In the more general MDP setting, we were able to obtain analogous results. First, if the optimal policy has average reward above σ, then the regret with respect to σ is bounded by a constant independent of the horizon T. Second, we could provide an algorithm which has constant regret with respect to σ if the optimal average reward $\rho^* > \sigma$ and obtains logarithmic regret bounds similarly as UCRL2 [18] otherwise. These results were first presented at EWRL 2023 and then submitted to *Mathematics of Operations Research.* The submission received encouraging reviews and a revision is currently under review.

5. Project Goals

5.1. True Shape of σ-regret.

5.1.1. *The Bandit Setting.* Although much simpler than the MDP setting, there are still quite a few open questions on satisficing in the multi-armed-bandit setting. The classic *(pseudo-)regret* after T steps is defined as

$$R_T := \sum_{t=1}^{T} \big(\mu_* - \mathbb{E}\left[\mu_{A_t}\right]\big), \tag{1}$$

where $\mu_* := \max_i \mu_i$ is the maximal mean reward of an arm.

A straightforward adaptation of the regret when considering satisficing instead of optimization is the (satisficing) σ-regret defined in analogy to (1) as

$$R_T^\sigma := \sum_{t=1}^{T} \max\big\{\sigma - \mathbb{E}\left[\mu_{A_t}\right], 0\big\}.$$

This natural notion of regret has been investigated in the precursor project [27]. The respective bounds that were derived are of the form

$$R_T^\sigma \leq \text{const} \sum_{i:\Delta_i^\sigma > 0} \Big(\frac{1}{\Delta_i^\sigma} + \frac{2\Delta_i^\sigma}{|\Delta_*^\sigma|^2} \Big), \tag{2}$$

where $\Delta_i^\sigma = \sigma - \mu_i$ for arms i with mean reward below σ and $\Delta_*^\sigma = \mu_* - \sigma$. These quantities are similar to the gaps $\Delta_i := \mu_* - \mu_i$ that appear in (classic) regret bounds [6]. Our bound improves over the results of [33], which only showed bounds logarithmic in T. An alternative constant bound on the σ-regret of order $\log(1/\Delta_i^S)/\Delta_i^S$ for each arm i below σ is given in the already mentioned [17] and in some cases better, in some cases worse than the bound of [27]. In any case, the bound of [17] does not depend on the gap Δ_*^σ of the optimal arm to the level σ.

Concerning lower bounds we were at first only able to show that in special cases when $\Delta_i^\sigma = \Delta_*^\sigma$ the upper bound is tight. In general, it seems not intuitive that the dependence of the regret with respect to Δ_*^σ is as in (2). So our first research goal is to determine the correct dependence of the σ-regret on the quantities Δ_i^σ and Δ_*^σ and to accordingly provide tight bounds on the σ-regret.

For the bandit setting with just two arms we were already able to make some progress and could show an upper bound on the σ-regret of order $\log(1/\Delta_*^\sigma)/\Delta$, where Δ is the classical gap between the good (i.e., optimal) and the bad arm (which is assumed to be below σ). This result will be presented at EWRL 2024. We also think that we can show that a lower bound of $\log(\Delta/\Delta_*^\sigma)/\Delta$ holds, but could not yet resolve the contradiction to the upper bound given in [17]. Generalizing these bounds to the case of arbitrarily many arms has turned out to be not quite straightforward, but would of course be one of our first goals of the project.

5.1.2. *The MDP Case.* Unlike the bandit setting where the regret only depends on the mentioned gaps, the results that we have for satisficing in MDPs have quite a few additional parameters. These are mainly mixing times that determine how long it takes for some policy until the empirical distribution approaches the respective stationary distribution induced by the policy. While it seems intuitive that some additional parameters such as the diameter will occur, which also appears in lower and upper bounds on the regret in the standard optimization setting, it is unclear whether additional mixing time related constants of the MDP are actually necessary for satisficing regret. In addition, while in the bandit setting we have lower bounds at least for a few special cases, for the MDP setting we do not have any results beyond the MAB setting. Thus, while we cannot expect to solve the problem of determining the true shape of the σ-regret in the MDP case within this project (as the respective question for the optimization setting has been closed only recently [7] after more than ten years of research), we at least aim at getting a bit closer by determining what are the appropriate parameters and providing respective lower bounds.

5.2. **Planning in MDPs for Satisficing and Alternative Performance Measures.** In the algorithms we have developed so far, the planning subroutines simply use optimization. In particular, our satisficing algorithm for the MDP setting computes and executes the optimal policy of the estimated MDP in exploitation phases. Computation of an optimal policy in MDPs is usually not considered to be particularly expensive, however obviously scales with the size of state and action space. However, in large MDPs a satisficing policy could be found much faster than an optimal policy (in simple cases e.g. using known planning algorithms such as weighted A* [30]). Of course, here a payoff between complexity and regret has to be considered, as the regret guarantees for the optimal policy in the empirical MDP will be better than for a satisficing policy. It is also an open problem how to formalize this payoff in an appropriate way, a question that we will investigate.

Another aspect here is that regret is defined with respect to the worst case. Accordingly, respective bounds will necessarily depend on the whole state-action space even when aiming at satisficing instead of optimization. In order to take into account that one *can* obtain a satisficing policy much faster one would need to consider different performance measures, a direction we also plan to explore in the project. Such an alternative measure might also be interesting in the classic optimization setting, where regret bounds are known to be overly pessimistic, and an alternative measure could prove to be more relevant for applications in practice. After eliminating the dependence on the horizon in the precursor project, the goal here would be to obtain performance guaranatees that do not depend on the size of the state-action space anymore but on problem-dependent parameters instead.

5.3. **From Satisficing to Optimization.** One of the core aspects we are interested in is the boundary between satisficing and optimization. In the precursor project we have shown that satisficing can be done with just constant regret. This implies that knowing a reference value that separates the optimal from the best suboptimal average reward one can obtain optimality with just constant regret, a fact that has been observed in the bandit setting already in [8]. On the other hand, we known from lower bounds that for learning an optimal policy the regret will always depend on the horizon T.

We are interested in the phase shift and in algorithms that try to learn an optimal policy by alternating satisficing and adapting the reference value for the former. While this will not result in overall improved theoretical performance (which is impossible to achive) we expect to obtain new insights and also hope that this research direction will help to develop algorithms with improved empirical performance.

5.4. **Extension to Multi-Objective RL.** We have already mentioned in Sec. 3 that previous work on multi-objective RL implies results for satisficing RL. On the other hand, we expect that findings of the project can be generalized to or used in the multi-objective setting. A generalization of the current result to the multi-objective setting seems to be straightforward when one aims at satisficing for all objectives. More interesting is to try to generalize the bounds derived in [11] for constraint MDPs to the case of random rewards (resp. costs) and from ergodic to communicating MDPs, which we think should be an achievable goal employing the methods found in the precursor project. We note that our current results already achieve that for the case when one is only interested in satisficing the constraints.

6. Methods

In the precursor project standard methods for MDPs and RL like value iteration, UCB [6], and UCRL [18] proved to be sufficient for our purposes. In the successor project it will be necessary to develop new approaches. However, we think that the mentioned algorithms will be a good starting point. Thus, for the preliminary improvements in the two-armed bandit setting (cf. Sec. 5.1.1) an appropriate modification of the UCB algorithm could be used that also looks promising for the setting with an arbitrary number of arms. A respective adaptation of the algorithm to the MDP setting seems possible. However in this general setting, the development of new proof techniques will be more crucial.

Concerning the employment of planning algorithms for satisficing (cf. Sec. 5.2), we will take a closer look at the wide range of algorithms available in this area (such as the already mentioned weighted A*-algorithm) and try to adapt them to work for our purposes. A different approach that looks promising is that for finding a satisficing policy in the context of robust MDPs [34], which may also be applicable to our RL setting. For the topic of investigating into the limit between satisficing and optimization (cf.Sec. 5.3) we have already outlined that we will first look into algorithms that employ a satisficing approach with adaptive threshold. Finally, for the multi-objective setting (cf. Sec. 5.4) we expect a combination of UCB/UCRL with one of the developed satisficing algorithms to work out.

7. Applications and Implications

As already noted in the proposal of the precursor project, a small project like ours neither will be able to solve all problems in RL nor will it cause a "*satisficing turn*" in the discipline. However, we expect to contribute to a complementary alternative to current RL theory with the potential to reduce the gap to RL applications. The precursor and the current project can only be first steps that has to be followed by further research of more advanced topics, including continuous state-action spaces, learning of state representations, and implementing domain knowledge into the considered algorithms. Only when addressing this wide range of questions we will finally have a chance to actually bridge RL theory and applications.

For this project we will focus on RL theory, which however does not mean that we want to disregard practical applications. A student has just recently started to work on a small logistics domain about storage allocation for his Master project, where we want to see the developed satisficing algorithms in action. These experiments shall compare satisficing to optimization approaches and will provide valuable feedback concerning the boundary of the two domains (cf. Sec. 5.3).

8. Ethical and Gender-related Aspects

This is a theoretical computer science project, so there are neither any ethical, safety-related, regulatory nor any sex-specific or gender-related issues that need to be considered. In case some of the developed algorithms are tested, this will be done on artificial sample domains, which does not involve any data privacy nor safety-related or regulatory issues.

Appendix 1: References

[1] Jacob D. Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, pages 4889–4897, 2016. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/0bf727e907c5fc9d5356f11e4c45d613-Paper.pdf`.

[2] Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 373–382. PMLR, 2021. URL `https://proceedings.mlr.press/v139/arumugam21a`.

[3] Dilip Arumugam and Benjamin Van Roy. The value of information when deciding what to learn. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 9816–9827, 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/517da335fd0ec2f4a25ea139d5494163-Paper.pdf`.

[4] Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, pages 9024–9044, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/3b18d368150474ac6fc9bb665d3eb3da-Paper-Conference.pdf`.

[5] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT 2010, 23rd Conference on Learning Theory*, pages 41–53, 2010. URL `http://sbubeck.com/COLT10_ABM.pdf`.

[6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

[7] Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward mdps. *CoRR*, abs/2406.01234, 2024. doi: 10.48550/ARXIV.2406.01234. URL `https://doi.org/10.48550/arXiv.2406.01234`.

[8] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *COLT 2013, 26th Annual Conference on Learning Theory*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 122–134, 2013. URL `https://proceedings.mlr.press/v30/Bubeck13`.

[9] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of a bandit arm relative to a reward quantile. In Satinder Singh and Shaul Markovitch, editors, *Proceedings*

of the Thirty-First AAAI Conference on Artificial Intelligence, pages 1777–1783, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14335.

[10] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 991–1000. PMLR, 2019. URL http://proceedings.mlr.press/v97/chaudhuri19a.html.

[11] Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward Markov decision process with constraints. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 3246–3270. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22i.

[12] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R. Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, pages 8378–8390, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5f7695debd8cde8db5abcb9f161b49ea-Paper.pdf.

[13] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, pages 3304–3312. PMLR, 2021. URL https://proceedings.mlr.press/v130/ding21d.

[14] Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pages 2998–3008, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3a20f62a0af1aa152670bab3c602feed-Abstract.html.

[15] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 1573–1581, 2018. URL https://proceedings.mlr.press/v80/fruit18a.

[16] Michael A. Goodrich and Morgan Quigley. Satisficing Q-learning: Efficient learning in problems with dichotomous attributes. In *Proceedings of the 2004 International Conference on Machine Learning and Applications – ICMLA 2004*, pages 65–72. IEEE Computer Society,

2004. doi: 10.1109/ICMLA.2004.1383495.

[17] Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Mach. Learn.*, 110(6):1233–1266, 2021. doi: 10.1007/S10994-021-05956-1.

[18] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010. URL `https://www.jmlr.org/papers/volume11/jaksch10a/jaksch10a.pdf`.

[19] Krishna Chaitanya Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *35th AAAI Conference on Artificial Intelligence, AAAI 2021, 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The 11th Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, pages 8030–8037. AAAI Press, 2021. doi: 10.1609/aaai.v35i9.16979.

[20] Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Mach. Learn.*, 108 (5):721–745, 2019. doi: 10.1007/s10994-019-05784-4.

[21] Yu Kohno and Tatsuji Takahashi. A cognitive satisficing strategy for bandit problems. *International Journal of Parallel, Emergent and Distributed Systems*, 32(2):232–242, 2017. doi: 10.1080/17445760.2015.1075531.

[22] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6:4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

[23] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala R. Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 17183–17193, 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/8ec2ba5e96ec1c050bc631abda80f269-Paper.pdf`.

[24] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1690–1698. URL `https://proceedings.mlr.press/v48/locatelli16`.

[25] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ-good arms in stochastic bandits. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/edf0320adc8658b25ca26be5351b6c4a-Abstract.html`.

[26] Nadav Merlis and Shie Mannor. Lenient regret for multi-armed bandits. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, pages 8950–8957. AAAI Press, 2021. doi: 10.1609/aaai.v35i10.17082.

[27] Thomas Michel, Hossein Hajiabolhassan, and Ronald Ortner. Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research*, 08 2023. URL `https://openreview.net/forum?id=QnT41ZGNh9`.

[28] Subhojyoti Mukherjee, Kolar Purushothama Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented UCB. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2515–2521, 2017. doi: 10.24963/ijcai.2017/350.

[29] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710, 2017. URL `https://proceedings.mlr.press/v70/osband17a`.

[30] Ira Pohl. Heuristic search viewed as path finding in a graph. *Artificial Intelligence*, 1(3-4): 193–204, 1970. doi: 10.1016/0004-3702(70)90007-x.

[31] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, pages 15277–15287, 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf`.

[32] Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proc. IEEE*, 102(4):544–571, 2014. doi: 10.1109/JPROC.2014.2307024.

[33] Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Satisficing in multi-armed bandit problems. *IEEE Trans. Autom. Control.*, 62(8):3788–3803, 2017. doi: 10.1109/TAC.2016.2644380.

[34] Haolin Ruan, Siyu Zhou, Zhi Chen, and Chin Pang Ho. Robust satisficing MDPs. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 29232–29258, 2023. URL `https://proceedings.mlr.press/v202/ruan23a`.

[35] Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Math. Oper. Res.*, 47(4):2815–2839, 2022. doi: 10.1287/moor.2021.1229.

[36] Rahul Singh, Abhishek Gupta, and Ness B. Shroff. Learning in Markov decision processes under constraints. *CoRR*, abs/2002.12435, 2020. URL `https://arxiv.org/abs/2002.12435`.

[37] Malcolm J. A. Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning, ICML 2000*, pages 943–950. Morgan Kaufmann, 2000. doi: 10.5555/645529.658114.

[38] Akihiro Tamatsukuri and Tatsuji Takahashi. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems*, 180:46–53, June 2019. doi: 10.1016/j.biosystems.2019.02.009.

[39] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *Proceedings of Machine Learning Research*, pages 1254–1262, 2016. URL `https://proceedings.mlr.press/v48/wu16`.

[40] Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 2823–2832, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/9e984c108157cea74c894b5cf34efc44-Paper.pdf`.

[41] Liyuan Zheng and Lillian J. Ratliff. Constrained upper confidence reinforcement learning. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020*, volume 120 of *Proceedings of Machine Learning Research*, pages 620–629, 2020. URL `https://proceedings.mlr.press/v120/zheng20a`.

Appendix 2: Financial Aspects

A2.1. The Local Environment. The Chair of Information Technology is one of the four chairs at the *Department of Mathematics and Information Technology* of the Montanuniversität Leoben and has been founded in 2002. Peter Auer has been the head of the institute since then, while the applicant joined in 2003 and since 2010 fills a permanent position as Associate Professor. The only other researcher employed by the university is Martin Antenreiter.

The research focus of the institute in recent years has been on reinforcement learning, beside some involvement in machine learning related projects with the industry. The institute has also been part of several 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; ComPLACS, FP7-ICT, n° 270327, CHIST-ERA project DELTA) projects. Within these projects, we have established several connections to national and European research partners, from which the proposed project will benefit as well.

At the institute we can provide all necessary resources needed for hosting another researcher, including a desk, a computer with internet access, and a computation server for performing simulation experiments.

A2.2. Project Duration, Personnel, Research Plan, and Dissemination. We plan to use the funding to hire a PostDoc for two years, starting mid 2025. Optionally, I would like to leave open the option of hiring a very good PhD student with sufficient knowledge in reinforcement learning instead. Due to the lower salary, a PhD student then could work on the project for a bit more than another year, 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.

Concerning the research plan, it would be most natural to start with working on the generalization of the preliminary results for the bandit setting with two arms (cf. Sec. 5.1.1) to an arbitrary number of arms. Otherwise, the topics suggested in Sec. 5 in principle can be worked on in an arbitrary order. However, personally I find investigations into the boundary between satisficing and optimization the most interesting part of the project and would start to work on this topic afterwards.

Concerning dissemination, we will submit our results to suitable conferences like COLT, ICML, or NIPS in order to make our research quickly available. Extended versions will be submitted to high quality peer reviewed journals.

A2.3. Requested Funding. We request funding for research personnel to conduct research on the topics outlined in the proposal. Given the sophistication of the field, only someone with a high-level education in reinforcement learning theory will be able to successfully work on the project. Accordingly, we apply for funding a PostDoc with experience in the field who will work full-time for two years. Travel costs for conference visits are to be paid from the 5% general costs added.

(in Euro)	first year	second year
PostDoc (40h)	84,030.–	89,786.06–
Total incl. 5% general costs	182,506.86–	

Appendix 3: Curriculum Vitae

Personal data

First name and surname: Ronald Ortner
Birth place and date: Lienz, June 24th, 1973
Nationality: Austria
Address: Montanuniversität Leoben, Lehrstuhl für Informationstechnologie,
Franz-Josef-Strasse 18, A-8700 Leoben, Austria
Website: `http://personal.unileoben.ac.at/rortner/`
ORCID: `https://orcid.org/0000-0001-6033-2208`

Education

1991 – 2002 graduate and PhD studies of mathematics, graduate studies of philosophy and computer science at the University of Salzburg

Dec 1997 graduated with distinction in mathematics (Master)

Dec 1998 graduated with distinction in philosophy (Master)

Jan 2002 received the doctorate in mathematics with distinction, doctoral dissertation: "Arrangements of Pseudocircles" (awarded the "Hans Stegbuchner-Preis")

Jun 2010 habilitation in "Grundlagen der Informationsverarbeitung" (Theoretical Computer Science) at Montanuniversität Leoben

Positions

1996 – 1999 research fellow at the Institute of Philosophy of Science of the International Research Center Salzburg

1999 – 2000 alternative service at the Red Cross

2000 – 2002 employed at Infineon Technologies in Villach
(working on production logistics and line simulation)

2002 – 2003 assistant professor at the Department of Mathematics C of the Graz University of Technology, research position in FWF-project P15577 "Asymptotic Properties of Random Walks on Graphs"

2003 – 2010 assistant professor at the Department of Mathematics and Information Technology of the Montanuniversität Leoben since 2010 associate professor at the Department of Mathematics and Information Technology of the Montanuniversität Leoben

2012 Erwin Schrödinger scholarship of the FWF for 10 month visit of Inria Lille (project “Structure in reinforcement learning" FWF: J 3259-N13)

Research Interests

reinforcement learning (theory), Markov decision processes, graph theory, combinatorial geometry

Funded Projects

2012 FWF (Austrian Science Fund) project J 3259-N13 “Structure in reinforcement learning", granted 28,825.– Euro

2014 – 2016 FWF (Austrian Science Fund) project P 26219-N15 *Structured and Continuous Reinforcement Learning*, granted 130,536.– Euro

2020 – 2022 ÖAD project WTZ, Project No. SI 13/2020 *Structural and symmetry properties of graph products*, granted 7,000,– Euro

2022 – 2024 FWF (Austrian Science Fund) project TAI 590-N *Reinforcement Learning: Beyond Optimality*, granted 150,761.10 Euro

Selected Publications

(for a complete list of publications see `https://orcid.org/0000-0001-6033-2208`)

Journal Papers

[1] Ronald Ortner. Regret bounds for reinforcement learning via Markov chain concentration. *Journal of Artificial Intelligence Research*, 67: 115–128, 2020.
URL: `https://doi.org/10.1613/jair.1.11316`

[2] Ronald Ortner. Optimal Behavior is Easier to Learn than the Truth. *Minds and Machines*, 26: 243–252, 2016.
URL: `https://doi.org/10.1007/s11023-016-9389-y`

[3] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret Bounds for Restless Markov Bandits. *Theoretical Computer Science* 558, 62–76 (2014).
URL: http://dx.doi.org/10.1016/j.tcs.2014.09.026

[4] Ronald Ortner. Adaptive Aggregation for reinforcement learning in average reward Markov decision processes. *Annals of Operations Research* 208(1), 321–336 (2013).
URL: `http://dx.doi.org/10.1007/s10479-012-1064-y`

[5] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* 11: 1563–1600, 2010.
URL: `http://www.jmlr.org/papers/volume11/jaksch10a/jaksch10a.pdf`

[6] Peter Auer and Ronald Ortner. UCB Revisited: Improved Regret Bounds for the Stochastic Multi-Armed Bandit Problem. *Periodica Mathematica Hungarica* 61(1–2): 55–65, 2010.
URL: `http://dx.doi.org/10.1007/s10998-010-3055-6`

[7] Ronald Ortner and Wolfgang Woess. Non-backtracking random walks and cogrowth of graphs. *Canadian J. Math.*, 59(4):828–844, 2007.
URL: `http://www.cms.math.ca/cjm/a150010`

Conference Papers

[8] Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, Odalric-Ambrym Maillard. Regret Bounds for Learning State Representations in Reinforcement Learning, In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* pp. 12717–12727.
URL: `http://papers.nips.cc/paper/9435-regret-bounds-for-learning-state-representations-in-reinforcement-learning.pdf`

[9] Peter Auer, Pratik Gajane, Ronald Ortner. Adaptively Tracking the Best Bandit Arm with an Unknown Number of Distribution Changes, In: *Conference on Learning Theory, COLT 2019, PMLR Volume 99*, pp. 138–158.
URL: `http://proceedings.mlr.press/v99/auer19a.html`

[10] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007. Lecture Notes in Computer Science 4539*, pages 454–468. Springer, 2007.
URL: `http://www.springerlink.com/content/44011766316x5x77`