Research Project Proposal: Sample complexity of transfer learning in reinforcement learning under a parameterized setting

Riccardo Poiani, poiani.riccardo@gmail.com

1. Introduction to the problem [max 1 page]

Reinforcement learning (RL) is the area of machine learning that studies, with the help of statistical tools, sequential decision-making problems in which an agent acts in an environment with the goal of maximizing the accumulation of a reward signal. This problem is usually formulated as a *Markov decision processes* (MDP). Training agents to act optimally taking smart decisions can be a slow and, sometimes, dangerous process (e.g., a physical robot may get damaged by doing very bad actions). To solve this issue, many approaches make use of *transfer learning* (TL) algorithms: when some kind of knowledge is available from similar tasks that have already been solved, it can be used to speed up learning in a new, unknown, and related task. Moreover, the study of RL algorithms, nowadays, often focuses on providing worst-case upper bounds on how many samples the agent needs to reach a behaviour of a desired quality: this measure is usually referred to as *sample complexity*.

When dealing with transfer learning in reinforcement learning, it is possible to face many different settings and scenarios that may differ in the allowed differences between tasks, knowledge transferred and performance metrics. The field has many problems of interest; ranging from, but not limited to, algorithms that are able to transfer automatically (i.e., without the need for human intervention), scalability to complex domains, and theoretical performance guarantees of the proposed solutions.

Despite the success of RL algorithms in practical cases, existing solutions are still not able to deal with very complex problems. In this case the transfer learning approach, which has already proven to work very well in practice, is essential for bringing RL into the real world. As an example scenario, it is possible to consider a robot that has to learn to swing a bat of a certain length and mass: if it already knows how to swing a different bat, it is possible to reuse this previous knowledge to speed up the learning process significantly, since the two tasks are very similar. The approach has also proved beneficial in more complex settings, such as the *autonomous driving* problem: here, the problems mentioned at the beginning are even more relevant.

A typical transfer-learning setting in RL is when an agent acts in an environment whose dynamics are regulated by some unknown parameters: this problem is often referred to as *parameterized Markov decision process*. The general problem is to study the sample complexity of agents that exploit the environment parametrization in order to improve performance. This problem can be addressed in two ways: with and without a *generative sampling model*. In the former, the agent can access samples in all the possible states and actions by querying a generative sampling model; in the latter, instead, this model is not available and the agent needs to navigate the environment step-by-step to identify what the true parameter is.

As already mentioned, transfer learning and sample complexity already received a lot of attention from the research community; indeed, the empirical benefits on the performance of these approaches have already been demonstrated. An open problem in the current literature is how to explore new tasks when transferred structure information is given. Since this is complex and little understood at the moment, an approach in which a generative model is available can help make progress in understanding how an agent can exploit the transferred knowledge.

2. Main related works [max 0.5 page]

Some works have been carried out for learning in classical RL problems in the presence of a generative oracle. [6] proposes an algorithm that samples uniformly from each state-action pair, and proposes an analysis of its sample complexity that is found to match a lower bound they suggest. However, to prove this match they need an assumption that restricts the possible values on the required accuracy of the method. This last problem has been overcome in [13] with the use of variance reduction techniques. Nevertheless, even if their result is an improvement, they still do not match completely the lower bound. The work of [14] studies the problem of planning in sequential decision problems under the generative settings, achieving the optimality in the worst case. Differently from the previous two algorithms, this one works in an online setting.

For what concerns the absence of a generative model, many algorithms have been proposed that work in different transfer problems. The algorithm presented in [9] studies the sample complexity when the true environment is known to belong to a finite class of N arbitrary models, proposing tight bounds but an impractical algorithm. [5] studies the finite parameterized setting and, by means of *sequential probability ratio test*, finds a bound to the sample complexity that does not depend on the state-action space, but on the size of the set of possible models. Contextual MDPs, presented in [7], are similar to the parameterized setting, and are further discussed in [11] and [1]. Hidden parameter MDPs presented in [4] and improved in [12][8], consider families of models with low-dimensional latent factors. In particular [8] is able to deal with non-linear mechanics by using Bayesian neural networks [10]. Indeed, their algorithm works empirically well, but, due to the complexity of the solution, sample complexity bounds are missing. The work presented in [3], instead, under certain assumptions, shows that their algorithm is able to reduce significantly the sample complexity. Finally, [16] considers transferring value functions and derives a finite-sample analysis, while [15] adopts an innovative and robust approach in transferring samples.

3. Research plan [max 1.5 page]

The goal of the research lies in understanding the sample complexity of the parameterized transfer learning setting described above. In particular, being able to comprehend which samples are more informative than others to identify the true parameter, and in which states higher accuracy is needed, is the main point in this problem. In particular, to better decompose this general goal, we can split it into two components. In the first one, the presence of a generative model is assumed and the goal is to provide an algorithm (possibly an online one) that gives accurate, optimal and detailed sample complexity bounds for the parameterized transfer learning setting described above. In particular, this part should help to fill the gap in the current literature mentioned above. Moreover, the presence of a generative oracle is a key component that allows us to focus merely on information regarding the structure in each state-action pair of the problem, without taking into account the more complex problem of navigating the environment. Thus, with respect to the non-generative setting it simplifies the problem, by putting a sort of threshold on what is possible to achieve, and, moreover, it may allow taking insights on how to develop algorithms for the more complex case that will be the focus of the second part of this research. Indeed, in the second part, the aim is to extend the output of the first work to the non-generative case, in order to improve the state-of-the-art sample complexity bounds of more practical scenarios.

The first part of the research will be mainly theoretical since the oracle that gives samples in any state-action pair is almost always missing in practical applications of the considered setting. The second part is, instead, both theoretical and experimental. Indeed, the aim is to produce solutions to transfer in parameterized settings, and, thus, it makes sense to compare empirical performance with state-of-the-art algorithms. This should be carried out without neglecting theoretical upper-bounds to the sample complexity.

The research plan takes inspiration from the split of the research goals into the two explained sub-goals, which, due to their nature, will be carried out sequentially, starting from the simpler generative case. We allocate about three months for the first goal, and four months for the second part, starting from November 2019.
In a tentative schedule, for what concerns the *first phase*, we will start dedicating 2 months to the theoretical analysis of the problem, sketching down different meta-algorithms of increasing complexity, trying to reach

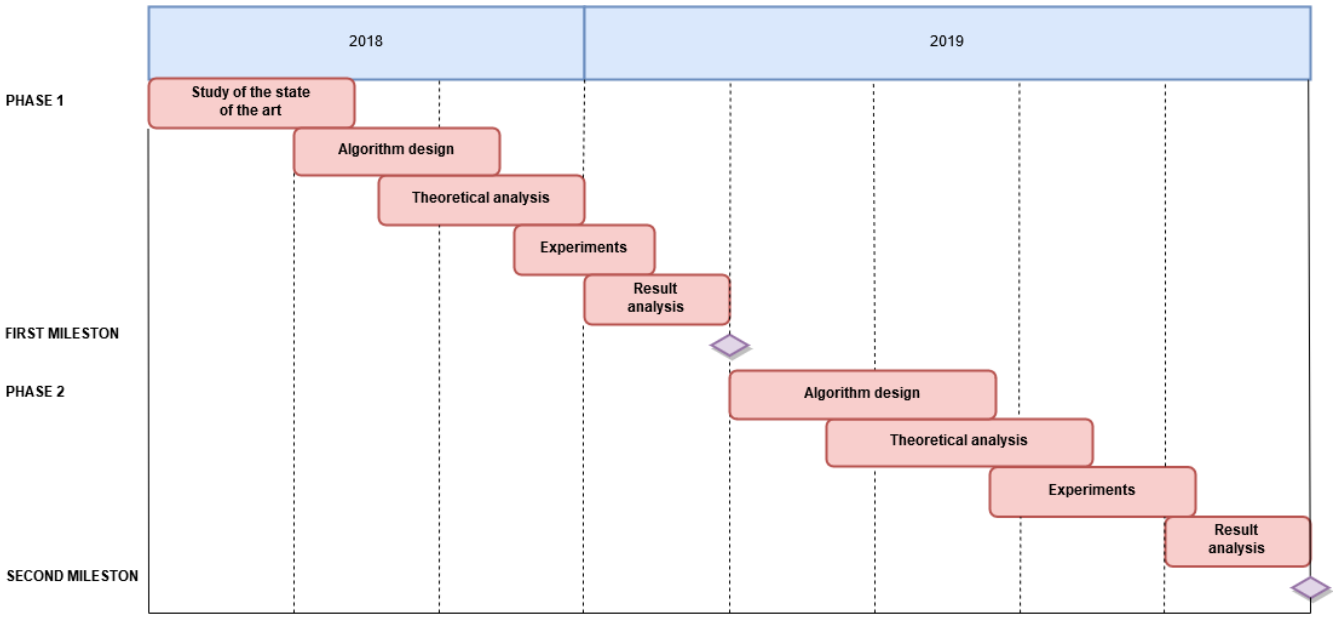

Figure 1: *Gantt chart* summarizing the research plan. Each vertical dashed bar represents a month. Each violaceous rhombus refers to a milestone. The first step, included for completeness, regarding the study of the state of the art has already been carried out.

optimal results in terms of sample complexity. Moreover, a *setting-dependent lower bound* on the sample complexity will also be analysed. In the case in which the proposed solution could only deal with a finite set of possible parameterized models, in order to also tackle the case in which this structure is infinite, a *discretization* approach that preserves the theoretical property of the algorithm will be proposed. Even if the focus of this first part is mainly theoretical, the empirical validation won't be completely neglected, and, when the theoretical analysis is almost over, a month will be allocated for it. Indeed, we will try to propose an RL problem in which the new algorithm can be adopted. In particular, the *Python* programming language will be used together with its vast ecosystem of reinforcement learning libraries, such as *OpenAIGym* [2], to simulate possible environments. After this, analysing the results and writing a document that summaries the work done will lead to the completion of a *first milestone*. This paper will be submitted to a top-rank conference such as *ICML*, whose submission deadline is on 7 February.
Regarding the *second phase*, as already mentioned, results will try to be extended to the non-generative setting. Again, at the beginning, the firsts 2-3 months will be dedicated to the theoretical analysis of the sample complexity of meta-algorithms that possibly take inspiration from the result of the first sub-goal. After that, experiments will be carried out and compared to state-of-the-art algorithms that solve the same type of problem. Finally, the *second milestone* will be reached with the analysis of the results and another document that may be submitted to a relevant conference, such as *NeurIPS*, whose submission deadline (around the end of May) matches with the end of our work. A summary of the research plan is summarized in Figure 1.

For what concerns the evaluation of our works, for the first sub-goal, since it is mainly theoretical, the existing sample complexity bounds of the classical RL setting will be used together with the lower bound that we intend to propose. We expect to reach at least the result of the classical RL case since the transfer setting should be able to benefit from the additional information at disposal. Theoretical bounds of state-of-the-art algorithms will be used also as metrics for the output of the second sub-goal, together with empirical comparisons. Indeed, as already mentioned, here the practical focus is more relevant than in the first case.

References

[1] Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of contextual mdps using spectral methods.

[2] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

[3] Brunskill, E., and Li, L. Sample complexity of multi-task reinforcement learning, 2013.

[4] Doshi-Velez, F., and Konidaris, G. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations, 2013.

[5] Dyagilev, K., Mannor, S., and Shimkin, N. Efficient reinforcement learning in parameterized models: Discrete parameter case. pp. 41–54.

[6] Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning 91*, 3 (Jun 2013), 325–349.

[7] Hallak, A., Castro, D. D., and Mannor, S. Contextual markov decision processes, 2015.

[8] Killian, T., Daulton, S., Konidaris, G., and Doshi-Velez, F. Robust and efficient transfer learning with hidden-parameter markov decision processes, 2017.

[9] Lattimore, T., Hutter, M., and Sunehag, P. The sample-complexity of general reinforcement learning, 2013.

[10] MacKay, D. J. C. A practical bayesian framework for backpropagation networks. *Neural Comput. 4*, 3 (May 1992), 448–472.

[11] Modi, A., Jiang, N., Singh, S., and Tewari, A. Markov decision processes with continuous side information, 2017.

[12] Rasmussen, C. E., and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[13] Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2018), SODA '18, Society for Industrial and Applied Mathematics, pp. 770–787.

[14] Szörényi, B., Kedenburg, G., and Munos, R. Optimistic planning in markov decision processes using a generative model. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1035–1043.

[15] Tirinzoni, A., Rodriguez Sanchez, R., and Restelli, M. Transfer of value functions via variational methods. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 6179–6189.

[16] Tirinzoni, A., Sessa, A., Pirotta, M., and Restelli, M. Importance weighted transfer of samples in reinforcement learning, 2018.