Research Project Proposal: On the sample complexity of Inverse Reinforcement Learning

Filippo Lazzati, filippo.lazzati@asp-poli.it

02/12/2022

1. Introduction to the problem [max 1 page]

% Description of the areas and research topics in which the problem is positioned.
Inverse reinforcement learning (IRL) is the problem of inferring the reward function of an agent, given its policy or observed behavior [1]. Given a certain Markov Decision Process [11] without reward function, along with a certain expert policy, IRL aims to compute a specific reward function which is feasible (namely, the expert policy is optimal in the MDP with such reward function) and satisfies specific properties. IRL, along with Behavioral cloning (BC), is one of the techniques we can use for solving the imitation learning problem, *namely the problem of efficiently learn a desired behavior by imitating an expert's behavior* [10]. Similarly to the forward Reinforcement Learning setting [12], it is crucial to efficiently explore the environment and collect the minimum amount of samples that allows solving the problem [6]. In the context of PAC learning [4], we talk about sample complexity to refer to this concept.
% Brief description of the research topic.
The research topic concerns the complexity of estimating all the reward functions that are feasible given a certain Markov Decision Problem without reward function but with an expert policy (the notion of feasible set is defined here [9]). In particular, it focuses on the computation of lower and upper PAC bounds on the sample complexity of such problem analogously to what can be found in the literature for the forward RL problem (like [2, 3]).
% Motivations to support the importance of the research topic.
The research topic has both a theoretical and practical importance. With regards to the former, it allows to characterize the complexity of the IRL problem from the point of view of the number of samples necessary to solve it in an acceptable way. We might then compare the complexity to the forward RL problem and see which of them is easier. Moreover, having a lower bound might help to assess the performance of existing and new algorithms. On the other side, PAC analysis allows to propose algorithms which are PAC optimal, and therefore it would allow to devise algorithms with worst-case theoretical guarantees. Finally, let me remark that IRL has currently a practical importance, but it has not been understood in-depth from a theoretical point of view.
% Description of the problem.
The problem of the thesis concerns the PAC sample complexity of estimating the feasible set [9] of an IRL problem. However, I will not consider, as estimation error, the distance between the inferred (at the subsequent RL step) policy and the true policy like in [9, 7], but I will focus on the straightforward Hausdorff distance between the estimated set and the true feasible set. As a metric between functions, I will use the max norm for the generative case, and the 1-norm for the forward case. The reasons behind this choice concern the complexity of the considered problem. If I use a generative sampling model, I can reach any (s, a) pair, thus I can provide max norm guarantees, while with a forward sampling model I might not be able to do so.
% Motivations to support the importance of the problem.
Up to now, IRL researchers have focused on proposing algorithms for computing only one specific feasible reward function that satisfies certain properties. Once obtained a reward function, the next step is straightforward. To perform forward RL in the MDP with such reward to determine a policy to use on our agent (which has now learned from the expert). However, few works have focused on the estimation of the entire feasible set, and in

particular on the complexity of doing so. If we were able to characterize such estimation problem, then new IRL algorithms might be devised atop such basic estimation problem, and the PAC framework is a powerful tool for characterizing it.

2. Main related works [max 0.5 page]

% Provide a concise description of the main related works and their limits.
In the context of RL, there are many works in literature that focus on the sample complexity, from many points of view. In [2], the authors focus on the problem of computing a PAC estimate of the action-value function in max norm in the case of generative model, while [3] partially relaxes the problem by providing almost matching bounds on the sample complexity of estimating the best value function in the finite-horizon problem with a forward model. If we think to bandits, [8] is a famous paper providing PAC bounds for the sample complexity of estimating the best arm. If we move to the IRL setting, as far as I know, the amount of works in this direction reduces a lot. In particular, [9] provides a characterization of the feasible set and then compute upper PAC bounds to the sample complexity of estimating the feasible set with a uniform sampling strategy under generative model in the infinite-horizon setting. [7] provides an upper bound under both generative and forward model in the finite-horizon setting. However, it should be remarked that none of these works provides lower bound results and that they analyze the distance between the estimated and the true feasible sets considering the consequences they have on the successive forward RL step. In the thesis case, the metric will directly measure the distance between such sets.

3. Research plan [max 1.5 page]

% Describe the goal of the research.
The goal of the research is to provide lower and upper (hopefully matching) bounds to the sample complexity of estimating the feasible set in the context of IRL. The distance between reward sets is measured as the Hausdorff distance using the max norm as metric between two functions:

$$h(\mathcal{R}, \hat{\mathcal{R}}) := \max\{\sup_{r\in\mathcal{R}} \inf_{\hat{r}\in\hat{\mathcal{R}}} ||r-\hat{r}||_\infty, \sup_{\hat{r}\in\hat{\mathcal{R}}} \inf_{r\in\mathcal{R}} ||r-\hat{r}||_\infty\}$$

where $\mathcal{R}$ is the true feasible set and $\hat{\mathcal{R}}$ is its estimate. The focus will be on two different settings: the infinite-horizon setting with a generative sampling model and the finite-horizon setting with a forward sampling model.
% Describe the nature of the research: theory, application, implementation, experimental, hybrid.
The research is theoretical. PAC bounds can be computed (and proved) using mathematical demonstrations that might exploit results from previous works found in literature. For the proof of the lower PAC bounds a specific problem is proposed, and the minimum amount of samples to collect to obtain (ϵ, δ)-correctness (according to the PAC framework) is computed (lower bounded). As far as the upper bounds are concerned, a specific algorithm/sampling strategy has to be proposed, and then mathematical tools like the Hoeffding's inequality [5] can be used. However, it should be noticed that, since an algorithm has to be proposed, it might be worthy to provide also its implementation and analyze its computational (not sample) complexity, and to perform its experimental validation.
% Describe the tasks in which the research is decomposed, remarking the output of each task.
The following tasks can be identified:

1. explore thoroughly the literature to identify all possible works that might be helpful in this case;
2. try to re-use the results for the forward RL problem in the case of the IRL case; hopefully, this task might provide some bounds;

3. try to devise some proofs combining mathematical tools for computing the bounds which there have not been able to provide re-using the results of RL. Hopefully after this step we will end up with all the desired bounds;
4. during the previous two steps, devising an algorithm for solving the estimation problem, and then computing its upper bound in some way and experimentally validating it;
5. once (hopefully) all the results have been determined, prepare a presentation/paper to present them.

% Provide a simple GANTT diagram of the task.
A textual representation of the GANTT diagram of the thesis prject is the following:

1. March 2022 - April 2022: Study the mathematical and algorithmic tools necessary for the understanding of the thesis' topics;
2. April 2022 - June 2022: Study and analysis of the literature to get an idea of the state-of-the-art techniques and tools in this context;
3. June 2022 - August 2022: Try to devise a mathematical proof for the lower bound to the sample complexity of IRL in the infinite-horizon case under generative model;
4. August 2022 - October 2022: Try to devise a mathematical proof for the lower bound to the sample complexity of IRL in the episodic fixed-horizon case under forward model;
5. October 2022: Try to devise an algorithm for solving the feasible set estimation task;
6. October 2022 - November 2022: Try to devise a mathematical proof for the upper bound to the sample complexity of the devised algorithm;
7. November 2022 - December 2022: Try to refine all the proved bounds to try to make them match and try to refine the algorithm too;
8. December 2022 - January 2023: Write down all the results in a scientific paper to spread them with the community.

In Fig. 1 is presented the GANTT diagram of the project; notice that the considered year is year 2022.

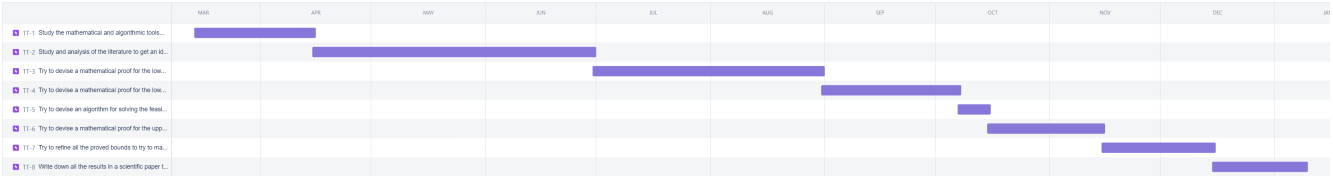

Figure 1: The GANTT diagram of the thesis project.

Where I have already carried out steps from 1 to 6 in the past months, and I am currently involved in step 7. The final goal is to send a paper to ICML 2023, whose deadline is the 9th January 2023.
% Describe the metrics to use to evaluate the outputs of the research.
In general, to evaluate the goodness of a research project, the main metrics adopted concern the amount of citations the paper receives during time and in which conferences/journals such citations take place. In other words, bibliometrics like the **impact factor** can be used to this aim. To evaluate the outputs of *this* specific research problem, I can identify two main metrics: the amount of bounds found and their tightness. With regards to the former, there is not too much to say, while as far as the second one is concerned, let me mention the fact that the tighter a bound, the better it is, since gross bounds are useless to characterize the complexity of a problem. The ideal case is that in which matching bounds are provided, namely the lower bounds match the upper bounds and therefore the sample complexity of the problem is completely characterized.

References

[1] Arora, S., and Doshi, P. A survey of inverse reinforcement learning: Challenges, methods and progress. *CoRR abs/1806.06877* (2018).

[2] Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model, 2012.

[3] Dann, C., and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning, 2015.

[4] Haussler, D. Probably approximately correct learning.

[5] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 301 (1963), 13–30.

[6] Kakade, S. On the sample complexity of reinforcement learning.

[7] Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning, 2022.

[8] Mannor, S., Tsitsiklis, J., Bennett, K., and Cesa-bianchi, N. The sample complexity of exploration in the multi-armed bandit problem.

[9] Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. Provably efficient learning of transferable rewards. In *Proceedings of the 38th International Conference on Machine Learning* (18–24 Jul 2021), M. Meila and T. Zhang, Eds., vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 7665–7676.

[10] Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics 7*, 1-2 (2018), 1–179.

[11] Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[12] Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction*, second ed. The MIT Press, 2018.