State of the Art on: PAC sample complexity of Inverse Reinforcement Learning

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

02/12/2022

1. Introduction to the research topic [max 2 pages]

% Description of the areas in which the research topic 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 [23] 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* [22]. Similarly to the forward Reinforcement Learning setting [24], it is crucial to efficiently explore the environment and collect the minimum amount of samples that allows solving the problem [13]. In the context of PAC learning [11], we talk about sample complexity to refer to this concept.
% List of the most prestigious journals and conferences related to the research topic (and the criteria according which such venues are prestigious).
The most prestigious **conferences** in this context are NeurIPS, ICML, IJCAI and AAAI. NeurIPS concerns, in particular, the topics of *General Machine Learning* and *Reinforcement Learning*, which are of interest for this thesis. ICML faces mainly the same topics as NeurIPS while also AAAI is focused on *Machine Learning* and *General Artificial Intelligence*. These conferences, along with ICLR which mainly concerns *Deep Learning*, are the most prestigious for a variety of reasons: although they are rather old and they have the biggest "h5-index", which is a (biased) measure of the amount of research in a certain direction. The main reason why they are considered among the most important is because of the amount of funds and participation from top companies and universities they receive. Indeed, the GGS rating is $A++$ for all these conferences. Prestigious **journals** are JMLR, MLJ and JAIR, which are ranked very high on Scimago.

1.1. Preliminaries

% Mathematical/modeling/algorithmic tools necessary for the understanding of the research topic. If many, report a brief description of them and provide some references.
There are many **mathematical tools** used in the field. Avoiding to refer to all the general tools like linear algebra, statistics, (convex) optimization and programming, the main specific tools are:

Concentration Inequalities which are a tool for analyzing *random fluctuations of functions of independent random variables around their means* [6];

Statistical Learning and Minimax Theory *which provides a rigorous framework for establishing the best possible performance of a procedure under given assumptions* [12];

Stochastic Processes and Martingale Inequalities which represent the analogous of concentration inequalities in the case the random variables are not independent of each other [18];

Likelihood Ratio Method which is a technique for proving the lower bound to the sample complexity of a certain quantity introduced in [20];

Information Theory which *answers two fundamental questions in communication theory: What is the ultimate data compression (answer: the entropy H), and what is the ultimate transmission rate of communication (answer: the channel capacity C)* [7];

PAC Learning which is a powerful framework for analyzing the characteristics of an algorithm and a problem [11].

With regards to the main **algorithmic tools** used, let me mention the most specific tools:

Dynamic Programming which refers to a recursive method for solving a problem by breaking it down in simpler subproblems [5];

Optimism in Face of Uncertainty which is a *heuristic* for exploration *in sequential decision-making problems* [14].

% Implementation and technological tools used in the field.
In the field, Python is definitely the most widespread programming language for the implementation of the algorithms proposed in the papers, whose code is usually shared on Github. Thanks to powerful libraries like Gym, TensorFlow and PyTorch, the implementation with Python is much simpler than with other programming languages. Moreover, Google offers the possibility of running code on its cloud through the well-known PaaS Google Colab. Another tool is, for instance, the solver Gurobi for mathematical optimization.

1.2. Research topic

% 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 [21]). 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 [4, 8]).
% 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.

2. Main related works [max 3 pages]

2.1. Classification of the main related works

% Provide and describe some dimensions to classify the main related works.
The topic of the sample complexity is a very theoretical one, and it is still mostly unexplored for IRL. Instead, for the case of RL, (almost) matching bounds have been computed for basically all the problems and quantities of interest. Related works can be classified in IRL and RL works, and then for the sampling model they use: according to [13], we can distinguish between the generative model and the forward model. Generative model means that we have an oracle that allows us to query whatever (s, a) pair we want, while the forward model requires us to explore the environment to sample the pairs. Another dimension is whether a paper provides both a lower and upper bound or only an upper bound. Finally, we can group works based on the setting they are considering, namely whether they consider the discounted/average infinite-horizon setting or the (discounted)

fixed finite-horizon setting. Notice that some combinations of problems are of little interest, like the finite-horizon setting under the generative model.
% Provide the classifications of the main related works, remarking which problems are open and which, instead, are fully assessed.
Let me start with the **Bandits** setting, which can be seen as a particular case of the RL setting. For such setting, we have matching bounds, so the problem is closed. With regards to the **Reinforcement Learning** setting, for the case of infinite-horizon under generative model we have matching bounds, while for the episodic under forward model we have almost matching bounds. For the general RL setting, a (probably) tight upper bound is found, which matches the lower bound in the finite case. It is thus clear that for what concerns the RL setting, basically all the problems are fully assessed. If we move to consider the **Inverse Reinforcement Learning** setting, then things change a lot. There have been found only the upper bounds for the infinite horizon under generative model and for the finite horizon under forward model for objectives different from those considered in this thesis, and no lower bound if provided. Therefore, the problem is still open.

2.2. Brief description of the main related works

% Provide a concise description of the main related works, emphasizing their limits, if any.
In [9] an algorithm was proposed for solving the multi-armed bandit problem with a PAC analysis for the number of time steps to identify a near-optimal arm; (ϵ, δ)-correctness was guaranteed after $O(\frac{n}{\epsilon^2} \log \frac{1}{\delta})$ time steps (n is the number of arms). Paper [20] provides a matching lower bound. Notice that [20] is the fundamental paper that introduces the *Likelihood Ratio Method* for computing lower bounds through a proof by absurd. Historically, one of the first important works on the sample complexity of RL is [13]. Moving to the RL discounted infinite-horizon setting under generative model, the paper that found matching bounds for the estimation of the action-value function in max norm is [4], where a matching bound of $\Theta(\frac{N}{\epsilon^2(1-\gamma)^3} \log \frac{N}{\delta})$, where γ is the discount factor ad $N := |\mathcal{S} \times \mathcal{A}|$ is the size of the state-action space, is proved by exploiting the likelihood ratio method. Such result improves on the previous best lower bound of $\tilde{\Omega}(\frac{N}{\epsilon^2(1-\gamma)^2})$ of [2, 10] and the previous best upper bound of $\tilde{O}(\frac{N}{\epsilon^2(1-\gamma)^4})$ proved for some algorithms like [3] for instance. If we consider the episodic fixed-horizon setting for RL, then paper [8] is the most interesting result, since the algorithm it proposes, UCFH, produces an upper bound of $\tilde{O}(\frac{|\mathcal{S}|^2|\mathcal{A}|H^2}{\epsilon^2} \log \frac{1}{\delta})$ to the number of episodes required (H is the length of each episode) to provide an ϵ-optimal estimate of the value function; moreover, [8] proves a lower bound of $\tilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|H^2}{\epsilon^2} \log \frac{1}{\delta+c})$, so it is clear that the bounds are almost matching. Previous works in this setting are not worthy to be mentioned. For the general RL problem, [17] proposes an algorithm, MERL, with a sample complexity of $\tilde{O}(\frac{N}{\epsilon^2(1-\gamma)^3} \log^2 \frac{N}{\delta\epsilon(1-\gamma)})$, and proves a matching lower bound except for logarithmic factors. As far as the IRL setting is concerned, the literature is rather poor. In [16, 15], information theoretic lower and upper bounds are proved but for very limited settings, which are not worthy to be mentioned. Instead, paper [21] analyzes the sample complexity of two sampling strategies under generative model, uniform sampling and TRAVEL, in the case of a discounted infinite-horizon IRL setting, and proving an upper bound of $\tilde{O}(\frac{SA}{\epsilon^2(1-\gamma)^4})$; however, the objective they consider is different from the Hausdorff distance with max norm between feasible sets that this thesis aims to compute. The other work, paper [19], extends the result of [21] to the finite-horizon case, by proposing algorithm AceIRL, and proving an upper bound of at most $\tilde{O}(\frac{SAH^5}{\epsilon^2})$ episodes. It should be remarked that, again, the objective is different.

2.3. Discussion

% Provide a discussion of the main open issues.
In the IRL setting, basically no lower bound to the sample complexity exists, neither in the infinite-horizon nor in

the finite-horizon episodic settings. Without any notion of lower bound, it is difficult to assess whether the upper bounds proposed are tight, and therefore how good the proposed algorithms are. Moreover, both the results [21, 19] measure the difficulty of estimating the feasible set by considering the goodness of the reward functions in the estimated set for the subsequent task of forward RL. However, no work considers the difficulty of estimating the feasible set as that of directly measuring the distance between sets using the Hausdorff distance and then the max (or one) norm.

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., Munos, R., Ghavamzadeh, M., and Kappen, H. Reinforcement learning with a near optimal rate of convergence.

[3] Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. J. Speedy q-learning. In *NIPS* (2011).

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

[5] Bellman, R. *Dynamic Programming*. Dover Publications, 1957.

[6] Boucheron, S., Lugosi, G., and Massart, P. Concentration inequalities - a nonasymptotic theory of independence. In *Concentration Inequalities* (2013).

[7] Cover, T. M., and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

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

[9] Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. pp. 193–209.

[10] Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res. 7* (2006), 1079–1105.

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

[12] John Lafferty, Han Liu, L. W. Minimax theory.

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

[14] Kamiura, M., and Sano, K. Optimism in the face of uncertainty supported by a statistically-designed multi-armed bandit algorithm. *Biosystems 160* (2017), 25–32.

[15] Komanduru, A., and Honorio, J. On the correctness and sample complexity of inverse reinforcement learning. *CoRR abs/1906.00422* (2019).

[16] Komanduru, A., and Honorio, J. A lower bound for the sample complexity of inverse reinforcement learning. *CoRR abs/2103.04446* (2021).

[17] Lattimore, T., Hutter, M., and Sunehag, P. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning* (Atlanta, Georgia, USA, 17–19 Jun 2013), S. Dasgupta and D. McAllester, Eds., vol. 28 of *Proceedings of Machine Learning Research*, PMLR, pp. 28–36.

[18] Lattimore, T., and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

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

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

[21] 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.

[22] 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.

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

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