Research Project Proposal: Exploiting Environment Configurability for Policy Space Identification

Guglielmo Manneschi, guglielmo.manneschi@mail.polimi.it

1. Introduction to the problem

Reinforcement Learning (RL) [10] is the branch of Machine Learning focused on decision making. The theoretical framework used to represent the typical learning scenario is the Markov Decision Process (MDP) [8]. An MDP is a stochastic control process that allows to model situations in which an agent has access to a set of observable features, can choose what to do from a set of actions, and receives rewards in the measure of how well it is doing a certain task.

The agent acts following a policy $\pi(\cdot|s)$ that specifies the probability distribution over the possible actions given that the agent is in state s. In the classical application, the goal is to find the optimal policy π^*, i.e. the one that gives the agent the greatest expected sum of future rewards. To find this policy, the agent can search inside a space of functions called *policy space* and indicated with Π_Θ. The policy space is defined by the set of observable features the agent has access to, that represent its perception of the environment and of itself. The size of this space is an index of the learning capabilities of the agent: the greater it is, the more complex behaviours it will be able to show. In fact a particular policy depends on a linear or nonlinear combination of these features specified by a vector of parameters $\theta \in \Theta$.

In general the policy space of an agent may not be known, and its inference can be interesting in many situations. A common problem in the RL literature is Imitation Learning, the task of learning through the observation of the behaviour of an expert. This method allows a faster convergence to an optimal solution and can be tackled with two different methods: Behavioural Cloning (BC) and Inverse Reinforcement Learning (IRL). The first tries to directly reproduce the policy used in the observed interactions with the environment, while the latter aims at understanding the reward function used by the expert. Both these approaches however usually require a representation of the policy space of the observed agent to work properly, hence the importance of its identification.

Another motivation to support the importance of this topic is related to the choice of the environment in which to put a learning agent. In fact, it is often possible to choose a specific environment in a learning experiment, and knowing the policy space of the agent offers the possibility of choosing the right level of difficulty for its specific capabilities.

The problem of identifying the policy space of an agent is not trivial. Since this space depends on the features the agent has access to, a possible approach could consist in their detection. Having access to a greater set of features, we may find out which ones are necessary to explain the behaviour of the agent. However in the classical framework with a fixed environment, many observable state features might be useless to the agent in the definition of the optimal policy, and hence undistinguishable from the ones that are not actually perceived. To solve this issue we would need to put the agent into different environments, so to solicit the use of the features which we are uncertain about.

Configurable Markov Decision Processes (Conf-MDPs) [3] are an extension of MDPs that add the possibility to configure the environment, and hence can be used to distinguish observable features from useless ones. For example, in the task of driving a car it is possible to change how the car reacts to the agent's actions, such as increasing the maximum speed, e.g. to discover if an agent can perceive obstacles at higher speeds. We can imagine the presence of a supervisor modifying the probability distributions that define the model's dynamics or the initial states distribution.

Conf-MDP is a novel approach in RL problems, based on the observation that many real-world problems allow the configuration of some environmental parameters. Additionally, modeling the policy space would allow to search for the environment that is most suited for a particular agent, which in turn can be found by using Conf-MDPs.

2. Main related works

This research project touches essentially two areas of the RL literature: the Configurable Markov Decision Process (Conf-MDP) framework [3] and the Imitation Learning field [6].

In [3] the formulation of the Conf-MDP can be found. Given a policy space Π and a model space $\mathcal{P}$, the performance measure $J_{\mu}^{P,\pi}$ is defined taking into account both the agent's policy π and the selected environment configuration P. In this way, it is possible to optimize the parameters to find the model and the policy that give the highest performance. The algorithm is a safe approach in this search of a policy-model couple, i.e., it ensures a monotonic improvement on the performance via a dissimilarity penalization that let only updates to a near region of the search space. Recently, other works tackled a similar problem by considering explicitly the cost of environment modifications [9].

How this framework can be applied to other RL problems is still an open issue, for example to Inverse Reinforcement Learning (IRL), the problem of recovering the reward function used by an agent by having access only to its trajectories. Contrary to what is done in Behavioural Cloning [2], which aims at finding the policy used by the agent in a specified environment, IRL is a more general representation of the goal of an agent which is resilient to environment changes.

The first approaches appeared in the literature require an interaction of the agent with the environment [5, 1]. Recently, a different approach emerged to solve this issue, for instance in [7], where the algorithm is based on gradient minimization and does not need to solve the forward problem. The price of not interacting with the environment is the need of a representation of the policy space of the agent [7, 4, 11].

3. Research plan

The goal of the research is to identify the policy space of the agent by using Conf-MDPs to gather additional knowledge coming from the configuration of the environment.

The contribution of this research will be threefold. There will be a theoretical contribution, mainly focused on the problem analysis and formalization. On the basis of these results, an algorithmic contribution will consist in the proposal and implementation of one or more algorithm to address the problem. These algorithms will be validated both on a theoretical and experimental view point, based on specifically defined measures of performance.

The research is made up of multiple tasks. The first phase (P1) is the analysis of the state of the art of the problem, and the identification of related works in the literature. This phase will be followed by a theorethical analysis of the problem (P2), in which we will focus on understanding the problem and come up with some ideas for a solution approach. The third phase is the implementation of the solution approach in the form of some algorithms (P3), this task will partially overlap the theorethical analysis, so to check the effectiveness of the theorethical part and possibly to gather additional insights on the problem at hand. Finally, an experimental evaluation of the proposed solution will be done (P4), both to confirm the theorethical results and to measure the performances of the algorithms, and the writing of the paper will be completed (P5). A Gantt diagram of the tasks is provided in Figure 1.

To evaluate the outputs of the research we will submit a paper to one of the main conferences in the field. Through the peer review process other researchers will give us an external evaluation of the work. This process will possibly lead to a publication.

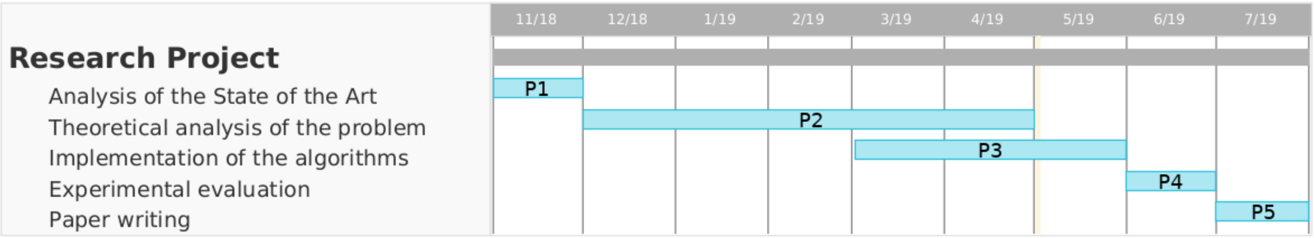

Figure 1: Gantt diagram of the research project.

References

[1] Abbeel, P., and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning* (2004), ACM, p. 1.

[2] Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and autonomous systems 57*, 5 (2009), 469–483.

[3] Metelli, A. M., Mutti, M., and Restelli, M. Configurable markov decision processes. In *35th International Conference on Machine Learning* (2018), vol. 80, PMLR, pp. 3491–3500.

[4] Metelli, A. M., Pirotta, M., and Restelli, M. Compatible reward inverse reinforcement learning. In *Advances in Neural Information Processing Systems* (2017), pp. 2050–2059.

[5] Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *Icml* (2000), vol. 1, p. 2.

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

[7] Pirotta, M., and Restelli, M. Inverse reinforcement learning through policy gradient minimization. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).

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

[9] Silva, R., Melo, F. S., and Veloso, M. What if the world were different? gradient-based exploration for new optimal policies. *EPiC Series in Computing 55* (2018), 229–242.

[10] Sutton, R. S., and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

[11] Tateo, D., Pirotta, M., Restelli, M., and Bonarini, A. Gradient-based minimization for multi-expert inverse reinforcement learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (2017), IEEE, pp. 1–8.