Research Project Proposal: A Non-Cooperative approach in Configurable Markov Decision Process

Alessandro Concetti, alessandro.concetti@mail.polimi.it

1. Introduction to the problem

Reinforcement Learning (RL) [14] is one of the three main branches of Machine Learning (ML), alongside supervised learning and unsupervised learning. While the goals of supervised and unsupervised learning can be summarized respectively in *"learning a model"* and *"learning a representation of data"*, the goal of RL is *"learning to control"*. Hence, RL studies how software agents ought to take actions in an environment in order to maximize a cumulative reward coherent with its goal. The theoretical framework used to represent learning scenarios in RL is Markov Decision Process (MDP) [13], through which the interaction of an agent with the environment can be modeled. The agent performs actions in the environment following a policy $\pi(.|s)$, which returns a probability distribution over the set of possible actions A given the current state s. The goal of RL is to learn the optimal policy π^* that maximizes the expected future reward.

In many practical situations, the environment where the agent performs actions can have some configurable parameters. To deal with those cases, a new framework called Configurable Markov Decision process (Conf-MDP) has been introduced in [9] extending the classical MDP with the possibility of configuring the environment. This novel framework has recently been enhanced by a new learning algorithm called REMPS which has been crucial to make this new framework suitable for real-world applications and opened the door to the use of Conf-MDP for many practical purposes. In [8], for instance, Conf-MDP has been leveraged to identify the policy space of an agent acting in the environment showing the possible benefits that this new framework could bring to the field of Imitation Learning [11]. For all these reasons, Conf-MDP is considered to be a very promising research area that deserves to be pushed further.

In practice, a Conf-MDP can be seen as a framework composed of two entities acting in the environment: an agent who learns the optimal policy and a supervisor whose aim is to configure the parameters in order to optimize the agent's learning process. The supervisor and the agent share the same goal: *reach the optimal policy as fast as possible*. Hence, there is no reason to support a multi-agent approach since the supervisor and the agent act at two different levels and the supervisor may be transparent to the learning agent. However, if the hypothesis of pure cooperation between the supervisor and the agent was broken, all these considerations would fail. When the two entity try to optimize two different reward functions, the Conf-MDP cannot be leveraged in its canonical formulation but it could be reasonable to shift to a non-cooperative multi-agent approach. Hence, the roles of the supervisor and the agent should be redefined: the supervisor, called *configurator* so far, does not know the reward function R_A of the agent and tries to model it through one of the main Inverse Reinforcement Learning (IRL) techniques [11]. Based on the estimate of R_A, the configurator modifies the environmental parameters in order to optimize its reward function R_C. The agent performs actions in the environment optimizing R_A but it could be aware of the presence of a non-cooperative configurator, leading to possible strategic behaviors [6].

Recent successes in adversarial machine learning as [10] or [5] have meant that the researchers' interest in adversarial models has significantly increased in recent years. In addition, there are many real-world applications that highlight the importance of a paradigm-shifting in Conf-MDPs, e.g. e-commence pricing or supermarket product placement. In the latter example, the configurator aims to maximize the supermarket's revenue while agents, namely customers, aim to maximize their shopping satisfaction. In this context, the configurator models the customers' reward function and changes the placement of the products on the shelves in order to maximize supermarket's revenue taking into account customer satisfaction.

2. Main related works

In the first work about Conf-MDP [9], in addition to the formalization of the new framework, a learning algorithm called Safe Model-Policy Iteration (SMPI) has been presented to jointly and adaptively optimize the policy and the environment configuration guaranteeing a monotonic improvement. Subsequently, the learning process has been enhanced by [7] in which a new algorithm, inspired by [12], called Relative Entropy Model-Policy Search (REMPS) has been proposed to deal with real-world continuous Conf-MDPs. The latest work about Conf-MDP [8] aims at pointing out the potential that this framework has in the field of Imitation Learning. In particular, it has been leveraged to ease the policy space identification of an agent acting in a configurable environment, proving to be very effective in distinguishing the non-controllable parameters from the useless controllable parameters. However, Conf-MDP cannot yet cover other real-world scenarios where the configurator and the learning agent do not optimize the same reward function. There are also other works that inspire the non-cooperative paradigm shift in Conf-MDP. e.g. main successes in the field of Multi-agent Reinforcement Learning [1, 4, 3] and theoretical results in Game Theory and Mechanism Design [6]. Stackelberg leadership model [15], particularly, is a 2-agent GT framework composed by a leader, who plays the first move, and a follower, who plays its rational response. Recently, T. Fiez, B. Chasnov, and L. J. Ratliff have shown in [2] important convergence results of the learning dynamics in Stackelberg Games, investigating the relationship between Nash and Stackelberg equilibria in zero-sum games and providing a new gradient-based update for the leader. Stackelberg Games share many elements with the proposed research project. Indeed, the configurator could be seen as the leader and the learning agent as the follower who takes note of the new configuration imposed and performs actions in the environment.

3. Research plan

The goal of the research is to extend the Conf-MDP framework to address scenarios where the configurator and the learning agent do not share the same goal. The contribution that the proposed project intends to provide is not only theoretical but also algorithmic and experimental. Indeed, the theoretical results will be leveraged to implement a learning algorithm that will be tested and validated both from a theoretical and experimental point of view.

The research can be decomposed into multiple tasks as can be seen in the Gantt diagram in Figure 1. The first task consists of the preliminary study and analysis of the state of the art. Then, the core of the project can be divided into two main phases. Finally, the project ends with the experimental phase and paper writing. To well understand the two-phase division of the project core, we have to delve deeper into the interaction between the agent and the configurator. As briefly explained in Section 1, the learning agent may or may not be aware of the presence of a non-cooperative configurator. This awareness could lead the agent to strategic behavior, making the interaction between the two entities much more difficult to model because the agent could perform some actions only to deceive the configurator ending up in a new better configuration w.r.t. R_A. In *phase 1* we consider an agent unaware of the game focusing more on the configurator, who has to model R_A and tune the parameters optimizing its own reward function R_C taking into account the estimate of R_A. In *phase 2* instead we consider an agent aware of the game by adapting what we develop in the first phase to deal with those scenarios. The structure of the two phases is exactly the same. The first step is the *theoretical analysis*, in which we will focus on understanding the problem and come up with some ideas for a solution approach. The second step is the *algorithm implementation*, which will partially overlap with the theoretical analysis so as to check the effectiveness of the theoretical part and possibly improve it. The experimental phase, instead, is not divided according to the agent awareness and the experiments will be performed in both cases in order to empirically confirm the theoretical results and measure the performance of the algorithm.

The project ends with the writing of a paper that will be submitted to one of the main conferences in the field (probably ICML 2021 whose deadline will be in February 2021), so that through peer review other researchers can objectively evaluate the outputs of the research project.

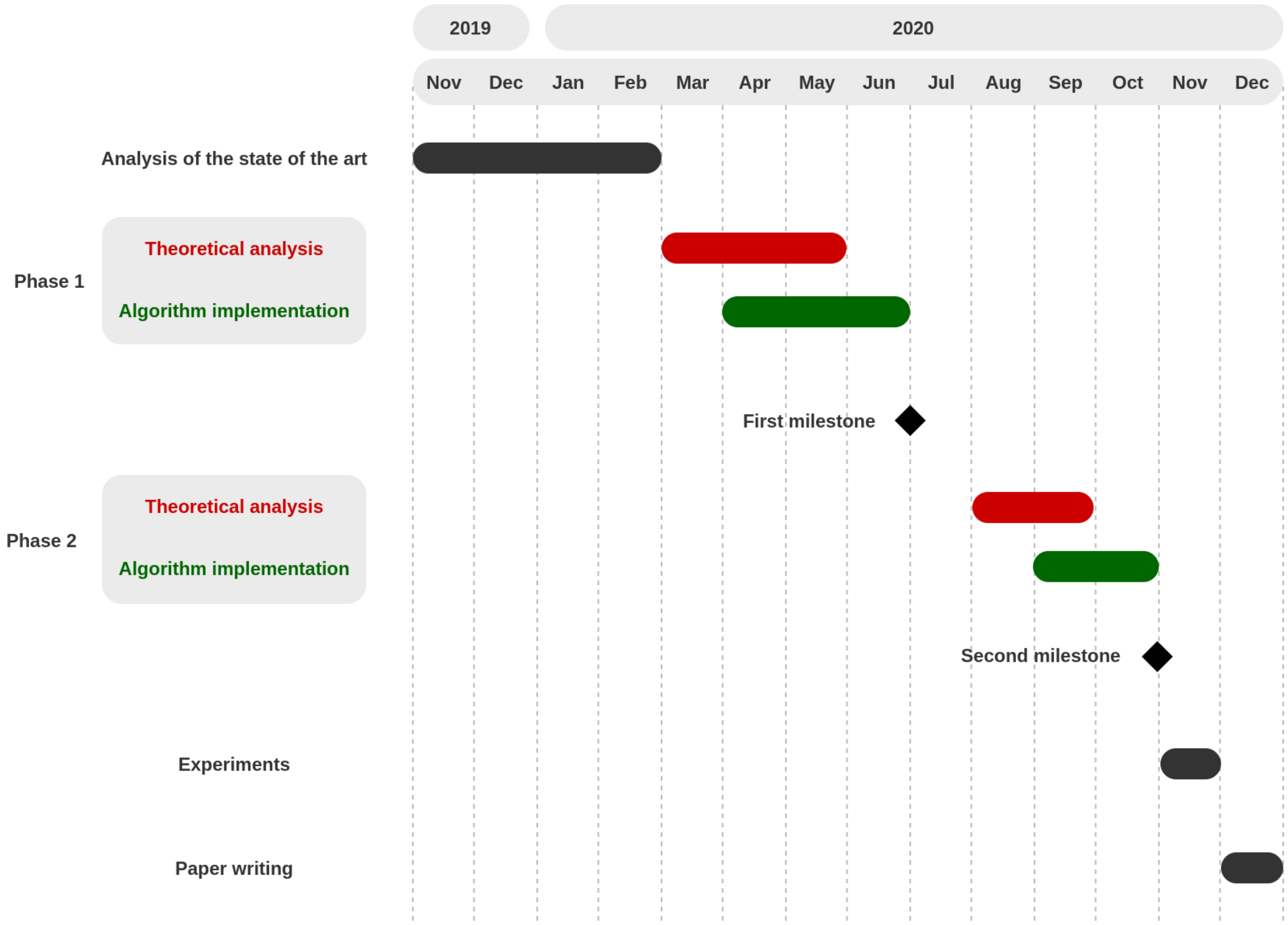

Figure 1: Gantt diagram

References

[1] Bu, L., Babu, R., De Schutter, B., et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38*, 2 (2008), 156–172.

[2] Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217* (2019).

[3] Hernandez-Leal, P., Kaisers, M., Baarslag, T., and de Cote, E. M. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183* (2017).

[4] Hernandez-Leal, P., Kartal, B., and Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems 33*, 6 (2019), 750–797.

[5] Ho, J., and Ermon, S. Generative adversarial imitation learning. In *Advances in neural information processing systems* (2016), pp. 4565–4573.

[6] Jackson, M. O. Mechanism theory. *Available at SSRN 2542983* (2014).

[7] Metelli, A. M., Ghelfi, E., and Restelli, M. Reinforcement learning in configurable continuous environments. In *International Conference on Machine Learning* (2019), pp. 4546–4555.

[8] Metelli, A. M., Manneschi, G., and Restelli, M. Policy space identification in configurable environments. *arXiv preprint arXiv:1909.03984* (2019).

[9] Metelli, A. M., Mutti, M., and Restelli, M. Configurable markov decision processes. *arXiv preprint arXiv:1806.05415* (2018).

[10] Mirza, M., and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[11] 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.

[12] Peters, J., Mulling, K., and Altun, Y. Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010).

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

[14] Sutton, R. S., and Barto, A. G. Reinforcement learning: An introduction.

[15] Von Stackelberg, H. *Market structure and equilibrium*. Springer Science & Business Media, 2010.