Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# An Advising Framework for Multiagent Reinforcement Learning Systems

**Felipe Leno da Silva, Ruben Glatt, Anna Helena Reali Costa***
Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil
{f.leno,ruben.glatt,anna.reali}@usp.br

## Abstract

Reinforcement Learning has long been employed to solve sequential decision-making problems with minimal input data. However, the classical approach requires a long time to learn a suitable policy, especially in Multiagent Systems. The *teacher-student* framework proposes to mitigate this problem by integrating an advising procedure in the learning process, in which an experienced agent (human or not) can advise a student to guide her exploration. However, the teacher is assumed to be an expert in the learning task. We here propose an advising framework where multiple agents advise each other while learning in a shared environment, and the advisor is not expected to necessarily act optimally. Our experiments in a simulated Robot Soccer environment show that the learning process is improved by incorporating this kind of advice.

## Introduction

Reinforcement Learning (RL) (Littman 2015) is a widely used tool to autonomously learn how to solve sequential decision problems, but RL agents are known to take a long time to reach convergence. The *teacher-student* framework (Taylor and others 2014) is one approach to alleviate this problem. A more experienced agent (teacher) advises actions to a learning agent (student), which results in faster learning. However, the *teacher-student* framework assumes that teachers follow a fixed (and good) policy. This means that, in order to apply this idea in a Multiagent RL domain, advising relations could only be established after teachers converged to a fixed policy. Agents could play both the roles of advisor and advisee during the learning process, as they may have explored different areas of the state-action space at a given time step. In this case, the advisor's current policy is most likely not optimal. Hence, agents must be able to evaluate how confident they are in their current policy to receive and give advice. We here propose a new framework for advice taking in which multiple simultaneously learning agents can share advice between them. To the best of our knowledge our proposal is the first policy advice framework intended to accelerate learning in a Multiagent System (MAS) composed of simultaneously learning agents.

*We are grateful for the support from CAPES, CNPq (grant 311608/2014-0), Google, Nvidia corporation, and São Paulo Research Foundation (FAPESP), grant 2015/16310-4.


## Preliminaries

RL solves sequential decision-making problems modeled as *Markov Decision Processes* (MDP). An MDP is described by the tuple $\langle S, A, T, R \rangle$, where $S$ is the set of environment states, $A$ is the set of available actions, $T$ is the transition function, and $R$ is the reward function. The agent goal is to learn an optimal policy $\pi^*$, that maps the best action for each possible state. However, learning $\pi^*$ may take a very long time, and the *teacher-student* framework alleviates this problem by receiving advice from a more experienced teacher (Taylor and others 2014). At every learning step, the teacher observes the student's current state and may provide a suggested action. However, advice is limited by a *budget* $b$. After $b$ is spent, the teacher is unable to provide further advice. Hence defining *when* to give advice is critical to accelerate learning. In the *Importance Advising*, the advice is triggered by the teacher when the importance metric $I(s)$ is above a predefined threshold:

$$I(s) = \max_a Q_{teacher}(s,a) - \min_a Q_{teacher}(s,a). \quad (1)$$

However, notice that Equation (1) is only efficient if the teacher has a fixed policy, because if she is still learning the Q-values estimates may be unreliable. As multiple agents in the same environment may be learning together, this importance metric is likely to be misleading in our setting.

## Proposal

We are interested in MAS composed of multiple autonomous agents simultaneously learning in a shared environment. Although we focus on RL agents, our framework is formulated more general and applicable for agents using any learning algorithm. Unfortunately, identifying which of the agents have a good policy is not easy, since some (or all) of them may be simultaneously learning in the same environment. Hence, instead of providing a fixed teacher like in the previous works, we propose to build *ad hoc advisor-advisee* relations. These relations are established for a single step according to each agent's confidence in her own policy for the current state. At each step, before choosing their action, agents evaluate a confidence function $\Upsilon$ and calculate a probability for broadcasting a request for advice to all reachable agents. All agents are restricted by the budgets $b_{ask}$

and $b_{ad}$ that are, respectively, the maximum number of times an agent can receive/give advice. In case the agent receives more than one advice for a given step, the executed action is chosen through a majority vote. The prospective advisor decides to answer or not to an advice request according to her confidence function $\Psi$. Thus, we propose a novel confidence metric to be applied in our setting, estimating the quality of the current policy, which is not done in the original *teacher-student* framework. This metric can be used for any kind of agent, regardless of which learning algorithm is used. With the assumption that the agent is learning in the environment and her policy is improving with the learning process, we calculate the confidence as:

$$\Upsilon_{visit}(s) = (1+\upsilon)^{-\sqrt{n_{visits}(s)}} \quad (2)$$

where $n_{visit}(s)$ is the number of times the agent visited the state $s$ and $\upsilon$ is a scaling variable. The intuition behind this equation is that the agent's policy becomes more reliable as she repetitively explores the state, thus the probability for asking for advice is lower when the state visit counter becomes higher. As the opposite is valid for prospective advisors, this confidence function can also be used for the advisor as $\Psi_{visit}(s) = 1 - \Upsilon_{visit}(s)$. Using ($\Upsilon_{visit}$, $\Psi_{visit}$), we derive the **Visit-Based Advising** for our setting.

## Experimental Evaluation

In our experiments we compare the *Ad Hoc Visit-Based Advising* (AdHocVisit) with an adaptation of the *Teacher-Student Framework* to our setting (Teacher-Student) and the regular learning without advice (NoAdvice).

Our experimental domain is the *Half Field Offense* (HFO) Robot Soccer simulator (Hausknecht and others 2016). In our HFO setting, three learning agents try to score goals against a high-skilled goalkeeper. A learning episode starts with the agents and ball initiated in a random position in the field, and ends when either the offense agents scored a goal, the defending agent caught the ball, the ball leaves the field, or a time limit is exceeded. In order to evaluate the learning speed of the agents with each of the frameworks, we trained the agents for 5000 episodes.The results here discussed are averages over 50 executions of this procedure. We evaluate the *Goal Percentage* metric, that is, the percentage of episodes in which a goal was scored. Figure 1 shows the improvement of the learning process and Figure 2 shows the spent budget. *AdHocVisit* is the top algorithm most of the time with a slightly superior performance, surpassing the *NoAdvice* after roughly 1000 learning episodes. *Teacher-Student* presented a poor learning speed until roughly episode 800, interval in which all of its budget was inefficiently spent with misleading advice. After that, *Teacher-Student*'s performance is comparable to *NoAdvice*, which means that it brought no benefits in this experiment. On its turn, *AdHocVisit* presented a better asymptotic performance, while expending thoughtfully the available budget. This shows that using the number of state visits to compute the confidence metric is reliable in the HFO domain. We now intend to combine our confidence metric $\Psi_{visit}$ with $I(s)$, in a way to consider both the expected policy quality (number of visits) and state importance (differences in Q-values). *AdHocVisit* achieved an improvement over the regular learning of roughly 6% in the asymptotic performance, which is a relevant improvement for this complex task. The main conclusion of this experiment is that our advising method outperformed regular learning even though the agents had no previous knowledge. Our results indicate that the *ad hoc* advice is a promising advising framework.


Figure 1: The goal percentage during learning.

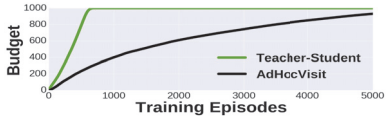


Figure 2: The spent budget for each algorithm.

## Conclusion and Further Works

We here propose a new advising framework in which multiple agents can simultaneously learn and advise each other, even when all agents start with no previous knowledge. Rather than defining a fixed teacher for a given student, the agents can establish *ad hoc* relations only for the states in which their current policies are expected to be useful for others, which is defined through confidence functions. Our experiments in a complex Robot Soccer task showed that our framework is promising but can be further improved. The next step is to improve our confidence function by combining it with the original *teacher-student* importance function. The *ad hoc* advising is a first step towards the Transfer Learning framework described in (Silva and Costa 2016).

## References

Hausknecht, M., et al. 2016. Half field offense: An environment for multiagent learning and ad hoc teamwork. In *Adaptive Learning Agents (ALA)*.

Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. *Nature* 521(7553):445–451.

Silva, F. L., and Costa, A. H. R. 2016. Transfer learning for multiagent reinforcement learning systems. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 3982–3983.

Taylor, M. E., et al. 2014. Reinforcement learning agents providing advice in complex video games. *Connection Science* 26(1):45–63.