Human Corrective Advice in the Policy Search Loop

Carlos Celemin
Electrical Engineering department, AMTC
University of Chile
carlos.celemin@ing.uchile.cl

Guilherme Maeda
Intelligent Autonomous Systems Laboratory
TU Darmstadt
maeda@ias.tu-darmstadt.de

Jens Kober
Cognitive Robotics department
TU Delft
J.Kober@tudelft.nl

Javier Ruiz-del-Solar
Electrical Engineering department, AMTC
University of Chile
jruiz@ing.uchile.cl

Keywords—Reinforcement Learning, learning from demonstration, interactive machine learning, movement primitives.

I. Introduction

Machine Learning methods applied to decision making problems with real robots usually suffer from slow convergence due to the dimensionality of the search and difficulties in the reward design. Interactive Machine Learning (IML) or Learning from Demonstrations (LfD) methods are usually simple and relatively fast for improving a policy but have the drawback of being sensitive to the inherent occasional erroneous feedback from human teachers. Reinforcement Learning (RL) methods may converge to optimal solutions according to the encoded reward function, but they become inefficient as the dimensionality of the state-action space grows.

Thus, this paper exploits the synergistic combination of RL with IML strategies. Human knowledge on the task is used to speed up the RL process, at the same time RL is used to provide more stability and robustness to the sporadic erroneous human feedback (humans are not perfect and prone to fail in repetitive tasks). Existing work on the combination of RL with human reinforcements [1]–[3] has shown benefits of the user's knowledge for speeding up the learning process while keeping the convergence properties of RL algorithms.

Policy Search RL has shown more appropriate for tackling high-dimensional robotic problems than value based RL [4]. Therefore, this work proposes the use of learning methods based on Policy Search (PS) techniques that additionally makes use of available human knowledge for reducing the number of trials, which is one of the main constraints of robot learning in the real-world. Here, corrective feedback in the action domain advised by human teachers is used instead of human reinforcements. In the proposed approach, human knowledge is provided to the PS learning agents with corrective advice using the COACH algorithm [5], which has shown to outperform pure autonomous RL agents and pure interactive learning agents based on human reinforcements.

This hybrid scheme of learning is applied to learn tasks modeled as Markov Decision Processes (MDP), and also problems with robot arms using policies represented by motor primitives.

II. Background

The proposed learning approach is a simultaneous combination of PS and the IML framework COACH, which are briefly described below.

Algorithm 1: Model Free Policy Search.

1: **repeat**
2: **Explore:** Execute M roll-outs using policy π_k
3: **Evaluate:** Obtain outcomes of trajectories or actions
4: **Update:** Compute π_{k+1} given the roll-outs and evaluations
5: **until** Policy converges $\pi_{k+1} \approx \pi_k$

A. Policy Search

PS is a branch of RL where parametrized policies are learnt directly in the parameter space, based on the cost given by the reward function, without a value function. The general structure of a PS method is presented in Algorithm 1, which includes three main steps: First, the exploration step creates samples of the current policy for executing each roll-out or episode. Second, the evaluation step quantifies the quality of the executed roll-outs according to the reward/cost function. Finally, the update step uses the evaluation of the roll-outs to compute the new parameters of the policy. This update can be based on policy gradients, expectation-maximization, information theoretic, or stochastic optimization approaches.

B. Learning from Human Corrective Advice

COrrective Advice Communicated by Humans (COACH) was proposed for training agents interactively during task execution [5]. In this framework, human teachers suggest corrections for the performed actions immediately after their execution with vague binary signals. The advice is a relative change of the action's magnitude, which is used for updating the policy with stochastic gradient descent. The binary signals are "to increase" or "to decrease" the executed action and could be independently given for each of the degrees of freedom that compose the action vector.

III. Policy Search guided with simultaneous human corrective advice

In this work, we combine PS with human advice, where the human teacher is able to correct the policy at each time step, whereas the PS only updates the policy model after each iteration of M trials based on the performance measurements of every roll-out. This combination, illustrated in Fig.1, can be seen as a regular PS algorithm with modified exploration, in which COACH is run every roll-out and the human corrections are incorporated as exploration noise. The set of M roll-outs, where some include human corrections, are then evaluated and used in the update step of the PS procedure.

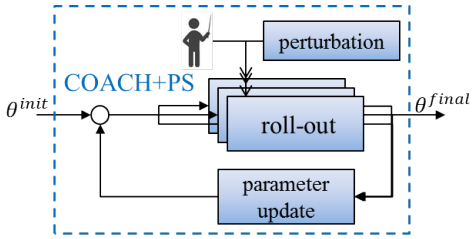

Fig. 1. Learning Simultaneously with COACH+PS.

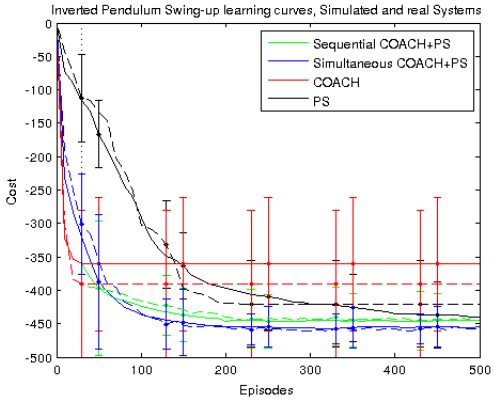

Fig. 2. Learning curves of the experiments for the inverted pendulum swing-up problem with the simulated system (normal lines) and the real system (dashed)

IV. Experiments and Results

Our approach has been tested in several tasks, some of them are briefly presented in this work. First, an experiment for learning a task in an MDP setting on an inverted pendulum is presented, followed by experiments in learning movement primitives of a robot arm.

A. Learning with MDP: Inverted Pendulum Swing-Up case

We compared different instances of our approach on this well-known nonlinear control problem in the RL literature: (i) using only a pure PS agent, (ii) a controller using only human feedback under the original COACH formulation, (iii) a controller where COACH is used to derive the initial policy which is subsequently refined using PS (namely, Sequential COACH+PS); and (iv) the Simultaneous COACH+PS presented in the previous section where the human has direct access to provide feedback on the roll-outs of PS. The results in Fig.2 show that the PS convergence was the slowest, while COACH had the fastest improvement in the early stage of the process. The proposed, simultaneous hybrid scheme showed the best balance in terms of velocity of convergence and final cost obtained.

B. Learning Movement Primitives

The validation of the hybrid approach for learning movement primitives was carried out by replicating the experiments of the original PI^2 paper [6]. The experiment consisted of learning robot arm reaching movements (similar to human reaching movements) with a total duration of 0.5 seconds. The task had the condition of reaching a specific via-point

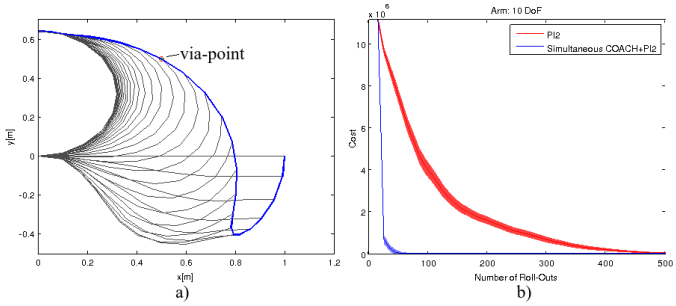

Fig. 3. Case of 10 DoF robot arm: a) movement through the via-point, b) learning curve

at $t = 0.3s$, and was evaluated for arms with 1, 2, 10, and 50 degrees-of-freedom (DoF). The experiments were executed first with the original PS algorithm PI^2, and followed by our hybrid approach combining PI^2 and COACH.

The results in Fig.3 show the 10 DoF arm case. Note that the use of the human corrective advice in the PI^2 algorithm speeds up the learning process more than 10 times when converging towards the lowest policy cost.

V. Conclusion

Vague corrections provided by human teachers usually result in fast, but sub-optimal learning, whereas PS relies on the definition of cost functions that are not very explicit or intuitive to the users' understanding. Therefore, this paper proposed the combination of human support to PS learning. From the point of view of interactive machine learning, these hybrid strategies provide more robustness to the convergence, since the sensibility to noisy or mistaken human corrections is diminished. Moreover, the quality of the policies is improved with the cost based corrections of PS which perform fine tuning of the policies taught by the users. The experiments with the swing-up task (MDP), and with the arm movement task (motor primitives) showed that the hybrid learning scheme can benefit from the advantages of both kinds of learning strategies. Experiments showed that the combination was capable of speeding up the convergence of a PS learner up to 30 times.

References

[1] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Aaai*, vol. 6, 2006, pp. 1000–1005.

[2] A. C. Tenorio-Gonzalez, E. F. Morales, and L. Villaseñor-Pineda, "Dynamic reward shaping: training a robot by voice," in *Ibero-American Conference on Artificial Intelligence*. Springer, 2010, pp. 483–492.

[3] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and mdp reward," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 475–482.

[4] M. P. Deisenroth, G. Neumann, J. Peters *et al.*, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.

[5] C. Celemin and J. Ruiz-del Solar, "Interactive learning of continuous actions from corrective advice communicated by humans," in *Robot Soccer World Cup*. Springer, 2015, pp. 16–27.

[6] E. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral control approach to reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. Nov, pp. 3137–3181, 2010.