Uncertainties based queries for Interactive policy learning with evaluations and corrections

Celemin, Carlos; Kober, Jens

DOI
10.1145/3461615.3485404

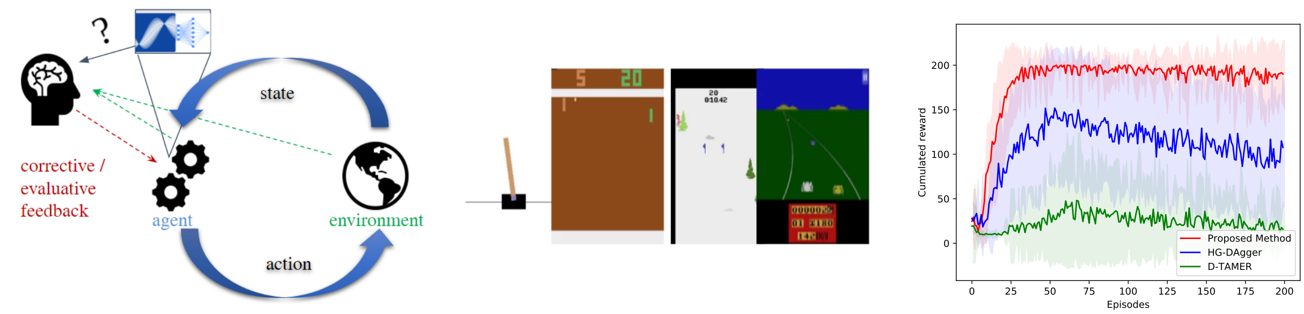
Publication date
2021

Document Version
Final published version

Published in
Proceeding ICMI 2021 Companion

Uncertainties based queries for Interactive policy learning with evaluations and corrections

Carlos Celemin
Jens Kober
c.e.celeminpaez@tudelft.nl
j.kober@tudelft.nl
Delft University of Technology
Delft, Netherlands



(a) Interactive Learning scheme. **(b) Simulated environments for validation.** **(c) Results.**

Figure 1: Proposed method, experiments, and results

KEYWORDS

Interactive Learning, corrective feedback, evaluative feedback, uncertainty, ambiguities

ACM Reference Format:
Carlos Celemin and Jens Kober. 2021. Uncertainties based queries for Interactive policy learning with evaluations and corrections . In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3461615.3485404

1 INTRODUCTION

Policy learning methods with humans in the loop have been becoming more popular in the last years within the community of Machine Learning and Robotics. There have been varied approaches for training a policy with human interventions depending on the kind of interaction the human teacher has with the robot learner. Teachers could train a policy with evaluations/rewards related to the executed actions [6, 8], or with episodic evaluations that are relative to other executions as in learning from preferences[1, 3]. Users could also directly teach how to perform with corrective demonstrations [4, 5] or relative corrections [7]. Additionally, Learning agents can model their policy uncertainty/confidence and use it to query the teacher for input about that uncertain situation [2].

In this work we propose a learning scheme that integrates human feedback in a data aggregation scheme, along with active learning queries in order to feed the system with more information whenever the policy is uncertain. The introduced learning method features two main contributions: i) interpreting and combining both evaluative and corrective feedback to shape directly the policy model with a data aggregation approach; ii) modeling epistemic and aleatoric uncertainty for actively querying the teacher either whenever the agent visit unseen states or when the agent has received ambiguous demonstrations.

2 LEARNING METHOD

The proposed interactive learning approach assumes there is a teacher who occasionally intervenes in the learning loop. It could be with corrective demonstrations, in situations in which the right action is known by her/him, or with rewards, when the action execution is less clear but the teacher has qualitative insights about the transitions of the agent.

The agent follows a stochastic policy $\pi(a|s)$ that is shaped with both kinds of feedback. Evaluative feedback is given with respect to the last action, positive rewards increase the probability of choosing the executed action $a_t(s_t) = argmax_a \pi(a|s_t)$, whereas negative rewards decrease its probability. On the other hand, corrective demonstrations a_t^h are obtained before the execution of the current policy action a_t in order to replace it. Thus, during the intervention,

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10.
https://doi.org/10.1145/3461615.3485404

the user is able to take over the operation of the agent, and the policy will be updated to increase the probability $\pi(a_t^h|s_t)$.

In order to model the uncertainties used for active learning, two different strategies are applied for measuring each of them. The epistemic uncertainty (model uncertainty) is calculated based on the variability of the prediction of an ensemble of neural networks computing $\pi(a|s)$. The aleatoric uncertainty (data uncertainty) is computed with a model that predicts the probability of choosing the wrong action, that is trained with the recorded demonstrations and the predictions of π over the states of the demonstrations. Therefore, for states with ambiguous demonstrations, this model is able to predict high probability of wrong actions because the policy is predicting only one of the demonstrated actions, i.e. having an error of prediction for some of those demonstrations.

3 EXPERIMENTS AND RESULTS

Several experiments have been carried out in order to evaluate the performance of the proposed approach. For the comparisons, it has been considered algorithms based on only either evaluative or corrective feedback. The environments used for the validation involved both simulated problems (OpenAI Gym environments), and a real robot arm KUKA iiwa 7. Additionally, the experiments involved human teachers, along with oracles for exhaustive evaluations and ablation studies. The results showed that the proposed method combining both kinds of human feedback and the queries based on the two kinds of policy uncertainty outperformed the other baselines, especially in conditions in which the teachers give noisy or mistaken feedback.

REFERENCES

[1] Riad Akrour, Marc Schoenauer, and Michele Sebag. 2011. Preference-based policy learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 12–27.

[2] Sonia Chernova and Manuela Veloso. 2009. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research* 34 (2009), 1–25.

[3] Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741* (2017).

[4] Ryan Hoque, Ashwin Balakrishna, Carl Putterman, Michael Luo, Daniel S Brown, Daniel Seita, Brijen Thananjeyan, Ellen Novoseller, and Ken Goldberg. 2021. LazyDAgger: Reducing Context Switching in Interactive Imitation Learning. *arXiv preprint arXiv:2104.00053* (2021).

[5] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. 2019. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8077–8083.

[6] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*. PMLR, 2285–2294.

[7] Rodrigo Pérez-Dattari, Carlos Celemin, Javier Ruiz-del Solar, and Jens Kober. 2018. Interactive learning with corrective feedback for policies based on deep neural networks. In *International Symposium on Experimental Robotics*. Springer, 353–363.

[8] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*.