Environment Guided Interactive Reinforcement Learning: Learning from Binary Feedback in High-Dimensional Robot Task Environments

Extended Abstract

Isaac Sheidlower
Tufts University
Medford, Massachusetts
Isaac.Sheidlower@tufts.edu

Elaine Schaertl Short
Tufts University
Medford, Massachusetts
elaine.short@tufts.edu

Allison Moore
Tufts University
Medford, Massachusetts
Allison.Moore@tufts.edu

Abstract

Continuous Action-space Interactive Reinforcement learning (CAIR) is the first continuous action-space interactive reinforcement learning algorithm that can out-preform state-of-the-art reinforcement learning algorithms early on in training. We test CAIR in two simulated robotics environments with intuitive and easy to design heuristic teachers.

Keywords

Interactive Reinforcement Learning; Human Robot Interaction

ACM Reference Format:
Isaac Sheidlower, Elaine Schaertl Short, and Allison Moore. 2022. Environment Guided Interactive Reinforcement Learning: Learning from Binary Feedback in High-Dimensional Robot Task Environments: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 Introduction

Interactive Reinforcement Learning (IntRL) is an extension of Reinforcement Learning (RL) that allows human or AI teachers to provide learning agents with feedback on their actions. This feedback allows the learning agent to both learn its task quicker and adapt its behavior to the preferences of the teacher. We propose Continuous Action-space Interactive Reinforcement learning (CAIR), the first continuous action-space IntRL algorithm that can outperform state-of-the-art RL algorithms on high-dimensional robot tasks. CAIR does this by learning both from a teacher and environmental reward, allowing a teacher to generally guide the agent towards the desired behavior, while the environmental reward helps stabilize learning from potentially noisy binary feedback on a high-dimensional action space. We present promising results in two simulated robotics environments using intuitively designed heuristic teachers that reflect easily understood aspects of the task. In this way, we can approximate strategies that a human teacher may have in mind when teaching a CAIR agent.

IntRL frameworks, such as TAMER [2, 10, 13] and COACH [3, 11], have successfully accelerated agent learning, but their success has largely been limited to tasks with discrete action spaces. This limitation restricts both the type of tasks an IntRL agent can learn and the level of precision their behavior can achieve within those tasks [7, 14]. Meanwhile, traditional RL has been increasingly focused on continuous action-space environments with the development of algorithms such as SAC, TD3 and HER [1, 6, 9]. While there has been some work done on IntRL in continuous action-spaces, none of the proposed approaches have been shown to compete with or surpass current RL algorithms in non-trivial tasks [5, 12]. CAIR seeks to bridge the gap between RL and IntRL, empowering people to teach agents through feedback in continuous action-space settings.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

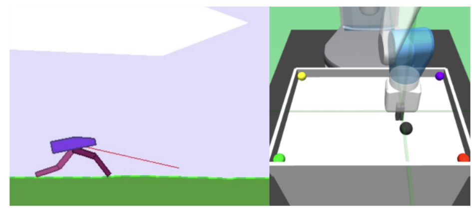
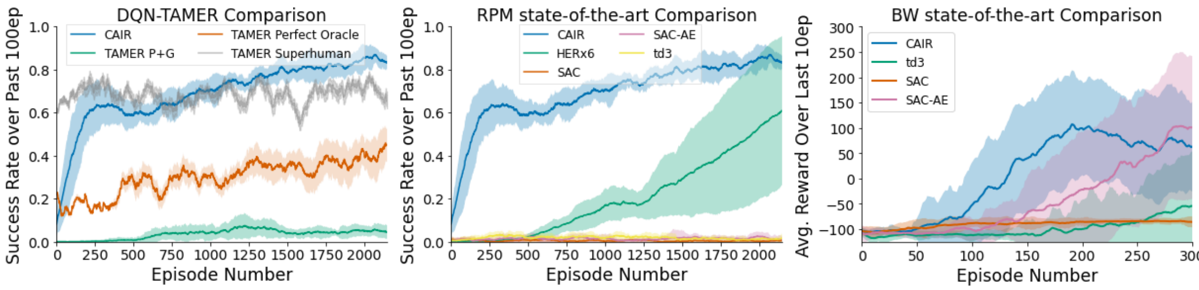


Figure 1: Left: BipedalWalker-v3, Right: Robot Push Multi

2 Method

CAIR learns in continuous observation and action-space environments. We model the environment as a standard MDP with: transition $T : (S, A) \rightarrow S$, reward $R : S \rightarrow r$, and policy $\pi : S \rightarrow A$. CAIR builds off of two prior learning paradigms Policy Shaping (PS) and Soft Actor-Critic (SAC), to enable IntRL in continuous action spaces. Policy Shaping, an IntRL algorithm, (PS) treats learning from a teacher providing binary feedback (in the form of +1/-1, good/bad signals) as separate from learning from the environmental reward and maintains a parameter C that reflects how consistent the agent believes the teacher's feedback to be [8]. SAC is an off-policy model-free actor-critic based RL algorithm that can learn in continuous action spaces and finds optimal policies by leveraging a maximum entropy approach. CAIR maintains two SAC networks, *Teach* and *Env*, which learn from teacher feedback and environmental reward respectively. CAIR then outputs the agents policy via sampling from a weighted average from *Teach* and *Env*'s policy distributions defined by:

$$\Delta_\pi = maximum(tanh(KL(\langle \mu_{teach}, \sigma_{env} \rangle, \langle \mu_{env}, \sigma_{env} \rangle)), 0)$$
$$\Delta_\pi = min(\Delta_\pi, \kappa)$$
$$\mu_{CAIR} = \Delta_\pi * \mu_{teach} + (1 - \Delta_\pi) * \mu_{env}$$
$$\sigma_{CAIR} = \sigma_{env}$$
$$a_{CAIR} \sim Normal(\mu_{CAIR}, \sigma_{CAIR})$$



(a) (b) (c)

Figure 2: Comparison between CAIR, Deep TAMER, and state-of-the-art RL algorithms.

Where κ is a hyper-parameter between 0 and 1 that reflects the maximum amount of trust the agent will put into the teachers policy at any given time (similar to C in PS [8]).

We test CAIR in two simulated robotics environments (Figure. 1). Robot Push Multi (RPM) is a sparse reward environment wherein the robot will receive a reward of 0 whilst the ball is at one of the four goals in the corners and a reward of -1 otherwise. Bipedal Walker (BW) is a dense reward environment, reward given to the agent is based on distance traveled, if the agent has crashed, and a slight negative reward for applying torque to its joints.

We developed heuristic teachers to provide feedback to CAIR that are meant both to be easy to design and reflect human perceptible/understandable properties of the task. For brevity, we present only the best performing heuristic teachers here. For RPM, we used a "push" heuristic teacher, which gives good feedback for pushing the ball and negative feedback otherwise. This heuristic is both intuitive and could be a teaching strategy adopted by a human teacher. For BW we used a "seeable" teacher. "Seeable" refers to fact that a human observer may not be able to tell when the walker is going forward or falling in very short time-scales. The seeable teacher provides positive feedback when the agent is moving forward at a *human perceptible* speed, has not fallen, and has one leg off the ground. Again, this heuristic teacher uses a strategy that a human could similarly approximate.

3 Experiments

We compared CAIR to both current RL algorithms and the IntRL algorithm DQN-TAMER. Results can be found in Figure. 2. For RPM, we compared CAIR against both single agent RL algorithms, as well as a multi-agent parallelized version of HER which has 6 concurrent learners (Figure. 2, b). Nevertheless, CAIR preforms much better than other approaches (reaching a 50% success rate in about 25 minutes). To compare with DQN-TAMER we had to discretize the environment's action space (Figure. 2, a). We also trained a "perfect oracle," a fully trained RL agent with an optimal policy that provides positive feedback if the learner's actions match its own. While perfect oracles do not reflect how humans actually teach robots [4], it provides the best possible scenario for the DQN-TAMER agent. "TAMER Superhuman" is a perfect oracle which provides feedback at every time step (note: this involves critiquing each individual robot action which is not possible for a person without significantly slowing down the robot). We can see that DQN-TAMER, even in the best case, gets out-preformed by CAIR as time goes on since CAIR acting in a continuous action-space does not have precision limitations. Also, note that DQN-TAMER when using an intuitive heuristic teacher "P+G," which is the same as the push heuristic but also provides positive feedback when the ball is at the goal, struggles to learn a policy much better than random.

In BW (Figure. 2, c), CAIR shows great learning improvement particularly early on in training, but is eventually out-preformed by some algorithms. We suspect that this is because the seeable teacher's feedback strategy is very effective in getting the agent to start walking, however since it does not adapt over time it will eventually give almost exclusively positive feedback once the agent is consistently moving forward. Though a primary drawback of using easy to define heuristic teachers is that they do not adapt over time as a human teacher would, the early gains of using CAIR are still very promising. And since CAIR learns from the teacher and the environment separately, CAIR can learn an optimal policy if a teacher stops engaging with the agent.

4 Discussion and Conclusion

While CAIR demonstrates great improvements in simulation, CAIR must be tested on a real robot. Furthermore, though the development of heuristic teachers as a way of evaluating IntRL algorithms is itself a contribution, CAIR must be tested with human teachers. When testing with human teachers, we plan on also introducing a new method for providing binary feedback called *toggle feedback*. Toggle feedback provides positive feedback until told otherwise and similarly for negative feedback. This allows a human teacher to provide dense feedback without the strain of having to constantly press a button or repeat an utterance.

In conclusion, we proposed CAIR, the first IntRL algorithm that can achieve state of the art performance in continuous action-space tasks. We presented results in two robotics environments using easy to define heuristic teachers. We plan on continuing to work with CAIR and investigate continuous action-space IntRL algorithms to keep people in the learning loop.

Acknowledgements

This work was funded in part by the National Science Foundation (IIS 2132887).

References

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2018. Hindsight Experience Replay. *arXiv:1707.01495 [cs]* (Feb. 2018). http://arxiv.org/abs/1707.01495

[2] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. 2018. DQN-TAMER: Human-in-the-Loop Reinforcement Learning with Intractable Feedback. *arXiv:1810.11748 [cs]* (Oct. 2018). http://arxiv.org/abs/1810.11748

[3] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L. Littman. 2019. Deep Reinforcement Learning from Policy-Dependent Human Feedback. *arXiv:1902.04257 [cs, stat]* (Feb. 2019). http://arxiv.org/abs/1902.04257

[4] Christian Arzate Cruz and Takeo Igarashi. 2020. A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, Eindhoven Netherlands, 1195–1209. https://doi.org/10.1145/3357236.3395525

[5] Carlos Celemin and Javier Ruiz-del Solar. 2019. An Interactive Framework for Learning Continuous Actions Policies Based on Corrective Feedback. *Journal of Intelligent & Robotic Systems* 95, 1 (July 2019), 77–97. https://doi.org/10.1007/s10846-018-0839-z

[6] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. *arXiv:1802.09477 [cs, stat]* (Oct. 2018). http://arxiv.org/abs/1802.09477

[7] Chris Gaskett, David Wettergreen, and Alexander Zelinsky. 1999. Q-Learning in Continuous State and Action Spaces. In *Advanced Topics in Artificial Intelligence (Lecture Notes in Computer Science)*, Norman Foo (Ed.). Springer, Berlin, Heidelberg, 417–428. https://doi.org/10.1007/3-540-46695-9_35

[8] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2625–2633.

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *arXiv:1801.01290 [cs, stat]* (Aug. 2018). http://arxiv.org/abs/1801.01290

[10] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: the TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*. ACM Press, Redondo Beach, California, USA, 9. https://doi.org/10.1145/1597735.1597738

[11] James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, David Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. *arXiv:1701.06049 [cs]* (Jan. 2017). http://arxiv.org/abs/1701.06049

[12] Ngo Anh Vien, Wolfgang Ertel, and Tae Choong Chung. 2013. Learning via human feedback in continuous state and action spaces. *Applied Intelligence* 39, 2 (Sept. 2013), 267–278. https://doi.org/10.1007/s10489-012-0412-6

[13] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *arXiv:1709.10163 [cs]* (Jan. 2018). http://arxiv.org/abs/1709.10163

[14] Ching-An Wu. 2019. *Investigation of Different Observation and Action Spaces for Reinforcement Learning on Reaching Tasks.* http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-271182