A Robust Approach for Continuous Interactive Reinforcement Learning

Cristian Millán-Arias
Escola Politécnica de Pernambuco, Universidade de Pernambuco
Recife, Brazil
ccma@ecomp.poli.br

Bruno Fernandes
Escola Politécnica de Pernambuco, Universidade de Pernambuco
Recife, Brazil
bjtf@ecomp.poli.br

Francisco Cruz
School of Information Technology, Deakin University
Geelong, Australia
Escuela de Ingeniería, Universidad Central de Chile
Santiago, Chile
francisco.cruz@deakin.edu.au

Richard Dazeley
School of Information Technology, Deakin University
Geelong, Australia
richard.dazeley@deakin.edu.au

Sergio Fernandes
Escola Politécnica de Pernambuco, Universidade de Pernambuco
Recife, Brazil
smurilo@ecomp.poli.br

ABSTRACT

Interactive reinforcement learning is an approach in which an external trainer helps an agent to learn through advice. A trainer is useful in large or continuous scenarios; however, when the characteristics of the environment change over time, it can affect the learning. Robust reinforcement learning is a reliable approach that allows an agent to learn a task, regardless of disturbances in the environment. In this work, we present an approach that addresses interactive reinforcement learning problems in a dynamic environment with continuous states and actions. Our results show that the proposed approach allows an agent to complete the cart-pole balancing task satisfactorily in a dynamic, continuous action-state domain.

CCS CONCEPTS

• **Computing methodologies → Reinforcement learning**; *Temporal difference learning*.

ACM Reference Format:
Cristian Millán-Arias, Bruno Fernandes, Francisco Cruz, Richard Dazeley, and Sergio Fernandes. 2020. A Robust Approach for Continuous Interactive Reinforcement Learning. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20), November 10–13, 2020, Virtual Event, NSW, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3406499.3418769

1 INTRODUCTION

Reinforcement learning (RL) is an approach that tries to solve the problem of an agent interacting with the environment to learn the desired task autonomously. The agent learns from its own experience, taking actions, and discovering which ones produce the greatest reward [17]. However, in many RL implementations, the space of states and actions is usually considered a discrete domain [6, 17, 19] or a discrete representation [1, 3, 7, 15]. Discretization prevents the agent from identifying which regions of space are more important than others. Moreover, in this process, information is lost, and it is difficult to learn from past experiences [8, 20]. In large domains, the agent spends a lot of time finding an optimal policy, being impractical in real-world applications [3]. Interactive Reinforcement Learning (IRL) is an approach that allows an external trainer advises the RL agent to improve its performance [3]. Additionally, RL agents usually work in environments which are not controlled, i.e., it is not guaranteed that the environment is kept in constant condition, avoiding some external noise input. Therefore, it is essential to develop robust algorithms that help the agent to learn faster an optimal policy, and to overcome uncontrollable disturbances in large domains.

2 INTERACTIVE AND DYNAMIC APPROACH

In several occasions, letting an agent learn a task by itself involves problems from exploration and weak tendency that avoid finding the optimal policy [9]. IRL considers a knowledgeable trainer, which gives advice or guidance to the RL agent, having an effect of restricting the action selection to those related to the target object [16].

In an IRL scenario, it is desired the interaction between the external trainer and the agent be as minimal as possible. The guidance can be obtained from either an expert or non-expert trainer, artificial agents with perfect knowledge of the task; or, previously trained agent [4, 5]. There are two approaches to receiving advice from an external trainer, reward-shaping [12, 16, 19], where external trainer provides additional reward, and policy-shaping, where an external trainer modifies the action just selected by the agent [7, 13].

HAI '20, November 10–13, 2020, Virtual Event, NSW, Australia

ACM ISBN 978-1-4503-8054-6/20/11.
https://doi.org/10.1145/3406499.3418769

During the learning, the agent performs an action that stimulates the environment in some way. If the environment is governed by a parametric system, the action acts as an input that modifies the output values, but not the model of the system. Thus, there could be parameters in the system that change concerning time; such parameters can be independent of actions and states [14]. In this sense, some features of the system change independently of agent control. Consequently, an RL agent can receive different amounts of reward for the same action, during the process to gather knowledge.

Morimoto and Doya [11] present the Robust Reinforcement Learning (RRL) an approach that introduces a disturber who provides disturbance to the environment. To resist a disturbance input, it considers an additional reward that modifies the main reward of the environment.

3 OUR APPROACH: INTERACTIVE ROBUST REINFORCEMENT LEARNING

In order to include advice during learning when the agent interacts with a dynamic environment, we combine the IRL and RRL approaches to propose Interactive Robust Reinforcement Learning (IRRL), an approach that involves advice for the agent to learn a task from an environment that has dynamic features.

For IRL in continuous scenarios, we use the approach presented in Millán et al. [10]. The main idea is to include external advice as a probability function in the policy $\pi(u|x, J)$, that denotes the probability density for taking action u in the state x when the trainer provides an advice J. Furthermore, as in some steps the trainer may not provide feedback, the likelihood of receiving feedback has probability $0 < L < 1$ [7].

To address the robust approach, we include policy-gradients [18] in the Actor-Disturber-Critic (ADC) algorithm proposed by Morimoto and Doya [11]. The main idea is to consider an objective function for agent policy π, and the disturber κ. In the disturber, the cost function $\Gamma(\kappa)$ evaluates the performance of the distribution in generating disturbances that have a more significant impact on the states, and in the selection of the next action. We consider that the disturber is a probability density function $\kappa_\omega(x)$ parameterized by the weight vector $\omega \in \mathbb{R}^{N_d}$. The parameter ω is adjusted in the direction of the gradient $\nabla_\omega \Gamma(\kappa)$ to generate the highest possible disturbance:

$$\omega_{t+1} - \omega_t \approx \alpha_\omega \nabla_\omega \Gamma(\kappa),$$

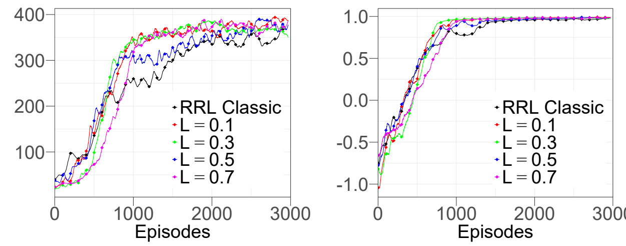
where α_ω is a learning rate of the disturber. In order to resist the disturbance, we consider the additional reward $w(\omega_t)$ of the form:

$$w(\omega_t) \longleftarrow \eta^2 \omega_t^\dagger \omega_t,$$

where $\dagger$ is the transpose of a vector and η is a parameter of robustness [11].

4 EXPERIMENTAL RESULTS

To evaluate the performance of our methodology, we apply it to the *cart-pole balancing task* [2]. In our experiments, 20 agents are trained with 3000 episodes, we also investigate the learning behavior for different values of the probability of likelihood L. The RL parameters are set with values $\gamma = 0.9$, $\sigma_x = 1$, $\sigma_j = 1$, $\alpha_\theta = 0.0001$, $\alpha_v = 0.0001$, $\alpha_\omega = 0.0001$, and $\eta = 0.45$. The friction of the cart on track is the disturbance, setting in $[0.0005, 1]$. To provide advice,



(a) Average steps (b) Average reward

Figure 1: Average steps (a) and average reward (b) over 20 runs using IRRL with different probability of likelihood L.

we use an oracle, a function that advises pushing the cart to the right or the left, whit values 1 and -1, respectively. In this sense, the advice favors the non-negative actions if advising to the right or non-positive actions if advising to the left. Fig. 1(a) represents the average steps taken by the agent to keep the pole balanced. We observe a better performance of the agents receiving advice compared to the autonomous RRL agent. In the first episodes, agents receiving a lot of advice may take longer episodes to improve their performance; however, after 1000 episodes, the average number of steps is higher than 300. With a probability of interaction $L = 0.7$, learning begins with a low performance; however, its performance improves at the same time than other probability of likelihood values L. Fig. 1(b) shows the average reward collected during learning. After 1500 episodes. all the agents converge to a reward close to 1.

5 CONCLUSION AND FUTURE WORKS

We presents an approach to implement the so-called IRRL, a combination of IRL and RRL in scenarios where states and actions are continuous in dynamic environments. In terms of average steps, our approach performs better than the autonomous RRL. However, the performance of the IRRL agent with probability $L = 0.5$ is close to that of the autonomous RRL agents. In terms of reward, we note that a cumulative reward of 1 is achieved for any probability L; however, values such as $L = 0.7$ have greater difficulty in the first learning episodes. This behavior is influenced by uninformative guidance, although the advice is correct concerning the space of actions that provide less information. Receiving much advice of this nature may not help in learning, even more, when the state is disturbed externally.

As future work, we intend to implement our approach in problems with larges domains, as well as other additional architectures, such as deep learning-based methods to carry out more complex tasks.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge to financing in part by the Coordenacão de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, the Brazilian agencies FACEPE and CNPq - Code 432818/2018-9, the partial support by Universidad Central de Chile under the research project CIP2018009, and the partial support by the Prysmian Group with funding code 001/2017.

REFERENCES

[1] Angel Ayala, Claudio Henríquez, and Francisco Cruz. 2019. Reinforcement learning using continuous states and interactive feedback. In *Proceedings of the 2nd International Conference on Applications of Intelligent Systems - APPIS '19*. ACM Press, New York, New York, USA, 1–5. https://doi.org/10.1145/3309772.3309801

[2] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13, 5 (1983), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

[3] Francisco Cruz, Sven Magg, Cornelius Weber, and Stefan Wermter. 2016. Training Agents With Interactive Reinforcement Learning and Contextual Affordances. *IEEE Transactions on Cognitive and Developmental Systems* 8, 4 (2016), 271–284. https://doi.org/10.1109/TCDS.2016.2543839

[4] Francisco Cruz, German I Parisi, and Stefan Wermter. 2018. Multi-modal feedback for affordance-driven interactive reinforcement learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Rio de Janeiro, Brazil, 5515–5122.

[5] Francisco Cruz, Peter Wüppen, Sven Magg, Alvin Fazrie, and Stefan Wermter. 2017. Agent-advising approaches in an interactive reinforcement learning scenario. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 209–214.

[6] Richard Dearden, Nir Friedman, and Stuart Russell. 1998. Bayesian Q-learning. In *15th AAAI*. AAAI, Wisconsin, 761–768.

[7] 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 (NIPS 2013)*. NIPS, Lake Tahoe, 2625–2633. https://papers.nips.cc/paper/5187-policy-shaping-integrating-human-feedback-with-reinforcement-learning

[8] A. Harry Klopf and L. C. Baird. 1993. *Reinforcement learning with high-dimensional, continuous actions*. Technical Report 513. WRIGHT LAB WRIGHT-PATTERSON AFB OH. 14 pages.

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

[10] Cristian Millán, Bruno Fernandes, and Fransisco Cruz. 2019. Human feedback in continuous actor-critic reinforcement learning. In *Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN, Bruges (Belgium), 661–666.

[11] Jun Morimoto and Kenji Doya. 2005. Robust Reinforcement Learning. *Neural Computation* 17, 2 (feb 2005), 335–359. https://doi.org/10.1162/0899766053011528

[12] Andrew Y. Ng, Harada Daishi, and Russell Stuart. 1999. Policy invariance under reward transformations theory and application to reward shaping. *ICML* 99 (1999), 278–287.

[13] PM Pilarski and RS Sutton. 2012. Between Instruction and Reward: Human-Prompted Switching. In *AAAI Fall Symposium: Robots Learning Interactively from Human Teachers*. AAAI, Arlington, Virginia, 46–52.

[14] Guillermo Puriel-Gil, Wen Yu, and Humberto Sossa. 2018. Reinforcement Learning Compensation based PD Control for Inverted Pendulum. In *2018 15th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. IEEE, Mexico City, Mexico, 1–6. https://doi.org/10.1109/ICEEE.2018.8533946

[15] Emrah Sisbot, Luis Marin, Rachid Alami, and Thierry Simeon. 2006. A mobile robot that performs human acceptable motions. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Beijing, 1811–1816. https://doi.org/10.1109/IROS.2006.282223

[16] Halit Bener Suay and Sonia Chernova. 2011. Effect of human guidance and state space size on Interactive Reinforcement Learning. In *RO-MAN 2011 - The 20th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Atlanta, 1–6. https://doi.org/10.1109/ROMAN.2011.6005223

[17] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT press, Cambridge, Massachusetts.

[18] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*. MIT Press Cambridge, Denver, CO, 1057–1063.

[19] Andrea L. Thomaz and Cynthia Breazeal. 2007. Asymmetric Interpretations of Positive and Negative Human Feedback for a Social Learning Agent. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Jeju, South Korea, 720–725. https://doi.org/10.1109/ROMAN.2007.4415180

[20] Hado Van Hasselt and Marco A. Wiering. 2007. Reinforcement Learning in Continuous Action Spaces. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE, Honolulu, HI, USA, 272–279. https://doi.org/10.1109/ADPRL.2007.368199