Training for Implicit Norms in Deep Reinforcement Learning Agents through Adversarial Multi-Objective Reward Optimization

Peschl, Markus

DOI
10.1145/3461702.3462473

Publication date
2021

Document Version
Final published version

Published in
AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society

Training for Implicit Norms in Deep Reinforcement Learning Agents through Adversarial Multi-Objective Reward Optimization

Markus Peschl
Department of Intelligent Systems
Delft University of Technology
Delft, The Netherlands
m.peschl@student.tudelft.nl

ABSTRACT

We propose a deep reinforcement learning algorithm that employs an adversarial training strategy for adhering to implicit human norms alongside optimizing for a narrow goal objective. Previous methods which incorporate human values into reinforcement learning algorithms either scale poorly or assume hand-crafted state features. Our algorithm drops these assumptions and is able to automatically infer norms from human demonstrations, which allows for integrating it into existing agents in the form of multi-objective optimization. We benchmark our approach in a search-and-rescue grid world and show that, conditioned on respecting human norms, our agent maintains optimal performance with respect to the predefined goal.

CCS CONCEPTS

• **Computing methodologies** → **Inverse reinforcement learning**; **Adversarial learning**.

ACM Reference Format:
Markus Peschl. 2021. Training for Implicit Norms in Deep Reinforcement Learning Agents through Adversarial Multi-Objective Reward Optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3461702.3462473

1 INTRODUCTION

Recent progress in reinforcement learning (RL) has vastly increased the feasibility and applicability of artificially intelligent agents to real world problems by employing deep neural networks that approximately solve high-dimensional control tasks. However, most state of the art algorithms are inherently black-box models that optimize for a specific manually engineered reward function. This can lead to unforeseen societal impacts which need to be accounted for before training and deploying the system [7].

As a consequence, there is a need for taking human values into account when training RL systems, which has previously been tackled by various approaches including inverse RL [4], reward shaping [8], partially observable Markov decision processes [1], safe RL [5] and multi-agent social choice [2]. However, most of these approaches do not scale to continuous or complex discrete environments due to assuming a tabular representation of the state space or hand-crafted features. While Saunders et al. [5] do not require any assumptions about the state space by employing deep RL, their approach only considers immediate negative consequences that can be avoided by blocking a single action.

To tackle the issue of scalability, we propose a hybrid architecture that enables deep reinforcement learning agents to optimize for a predefined goal while adhering to implicit norms learned from human behavior. Furthermore, by employing a bottom-up approach which builds on the adversarial inverse reinforcement learning framework [3], our agent is able to adapt to temporally complex constraints which can not directly be achieved through action blocking.

2 METHODS

We cast the problem of optimizing for a predefined goal while adhering to human norms into a multi-objective Markov decision process (MOMDP) with linear preferences, which is given by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \boldsymbol{\lambda} \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the sets of possible states and actions respectively, $\mathcal{P}(s'|s,a)$ denotes the state transition probability function and $\mathbf{r}(s,a) \in \mathbb{R}^d$ is a vector-valued reward function. Finally, we consider preference vectors $\boldsymbol{\lambda} \in \mathbb{R}^d$ to determine preferences among competing objectives. The RL goal then consists of finding a policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ that maps states to a probability distribution over actions which maximizes the expected cumulative reward $\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \boldsymbol{\lambda}^T \mathbf{r}(s_t, a_t)\right]$, where $0 < \gamma \leq 1$ is a temporal discount factor.

We assume that some components of $\mathbf{r}(s,a) \in \mathbb{R}^d$ are known in advance, these correspond to the primary goal that the agent ought to maximize, whereas norms represent the other components. To learn norms, we make use of human demonstrations which we assume to be norm satisfying and infer corresponding reward signals with adversarial inverse reinforcement learning (AIRL) [3]. AIRL trains a discriminator of the form

$$D_\theta(s,a) = \frac{\exp f_\theta(s,a)}{\exp f_\theta(s,a) + \pi(a|s)}, \tag{1}$$

which outputs the probability of the state-action pair (s,a) coming from the dataset $\mathcal{D} = \{\tau_i\}_{i=1}^n$ of human demonstrations rather than from an agent following the policy π. Simultaneously, the policy π is updated to match the demonstrations in $\mathcal{D}$ more closely.

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462473

Assuming that we want the agent to optimize a primary goal given by rewards $r_0(s, a)$, we then use a modified update rule for the generator π for maximizing

$$\mathbb{E}_\pi \left[\sum_{t=0}^{T} \lambda(f_\theta(s_t, a_t) - \log \pi(a_t|s_t)) + (1-\lambda) r_0(s_t, a_t) \right]. \quad (2)$$

Mathematically speaking, this results in π optimizing for r_0 while regularizing by a Kullback-Leibler divergence term $KL(\pi(\tau)||p_\theta(\tau))$, where $p_\theta(\tau) \propto \mathcal{P}(s_0) \prod_{t=0}^{T} \mathcal{P}(s_{t+1}|s_t, a_t) e^{\gamma^t f_\theta(s_t, a_t)}$ is the maximum entropy inverse RL probability distribution induced by f_θ [9].

3 EXPERIMENTS

We train a deep reinforcement learning agent in a (stochastic) grid world, with state inputs being a numerical matrix representation encoding the states of each cell. Both the policy and the discriminator employ a three-layer convolutional neural network, followed by linear output layers respectively. Furthermore, the policy is trained via proximal policy optimization (PPO) [6] due to its ease of implementation.

Testing is done in the *burning warehouse* environment (Figure 1), where the primary goal for the agent is to spend most of its time in a specific tile G. In general, G can be understood as any primary goal of interest, such as calling a fire department, extinguishing fire or other relevant tasks. Besides G, there are workers in the building which move in any of the four directions at each time step. We assume that these workers are lost and they need to be picked up by the robot in order to successfully escape the building. The agent can achieve this by moving onto their respective positions before the episode ends after $T = 100$ time steps.

To enable learning norm-satisfying behavior, we supply the agent with $n = 10$ human demonstrations which primarily move towards lost workers and neglect the goal tile G. We then train multi-objective AIRL for different values of λ, with $1e6$ total environment steps each. For example, for $\lambda = 0.1$ the agent is able to optimize for the primary goal while adhering to the demonstrated norms (Figure 2). These preliminary results demonstrate that our agent learns to combine the norms inferred from human demonstrations (saving people) with the reward-driven behavior (spending time in G).

4 DISCUSSION & FUTURE WORK

One important aspect of our approach is that it does not require human demonstrations to optimize for the primary goal in any form. This way, we can utilize the strengths of traditional RL when a reward function is available, while ensuring that certain constraints, which might not be easily expressed in a reward function, are met. While in our example, the goal of saving humans could in theory be expressed with a reward function, our experiment showed that even when this is not the case, we can optimize for human constraints by having access to a small set of demonstrations. Furthermore, it directly extends the settings of Noothigattu et al. [4] and Wu, Lin [8] to deep RL, dropping the assumption of a handcrafted feature space. Nonetheless, a variety of open questions remain to be addressed in future work. Firstly, we assumed demonstrated norms to be consistent with each other. However, in real world scenarios this might not be the case, which would lead to inaccurate reward representations learned by AIRL. Secondly, the adaptation of multi-objective RL algorithms to allow for Pareto-efficient policy search with online reward learning will be of interest for scaling to more complex domains. Specializing AIRL to efficiently allow for learning multiple conflicting norms will therefore be the focus of following work.

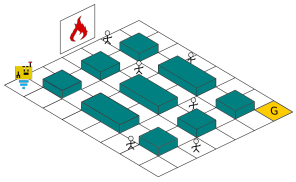

Figure 1: The burning warehouse toy environment.

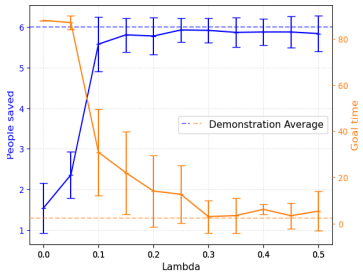

Figure 2: People saved (blue) and goal time (orange) for different values of λ. Error bars indicate a 2σ confidence interval over 100 distinct episodes.

ACKNOWLEDGMENTS

To Luciano Cavalcante Siebert and Arkady Zgonnikov for providing critical feedback and insightful supervision of this research.

REFERENCES

[1] David Abel, James MacGlashan, and Michael Littman. 2016. Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society.*

[2] Adrien Ecoffet and Joel Lehman. 2020. Reinforcement Learning Under Moral Uncertainty. arXiv:2006.04734

[3] Justin Fu, Katie Luo, and Sergey Levine. 2017. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning. arXiv:1710.11248

[4] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. 2019. Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development* 63, 4-5 (2019). https://doi.org/10.1147/JRD.2019.2940428

[5] William Saunders, Andreas Stuhlmüller, Girish Sastry, and Owain Evans. 2018. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the International Joint Conference AAMAS*, Vol. 3. 2067–2069. arXiv:1707.05173

[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347

[7] Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. 2021. The Societal Implications of Deep Reinforcement Learning. *Journal of Artificial Intelligence Research* 70 (March 2021). https://doi.org/10.1613/jair.1.12360

[8] Yueh Hua Wu and Shou De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018.* 1687–1694.

[9] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 3. 1433–1438.