Boosting Robustness in Preference-Based Reinforcement Learning with Dynamic Sparsity

Extended Abstract

Calarina Muslimani
University of Alberta
Canada
musliman@ualberta.ca

Bram Grooten
Eindhoven University of Technology
Netherlands
b.j.grooten@tue.nl

Deepak R.S. Mamillapalli
University of Alberta
Canada
mamillap@ualberta.ca

Mykola Pechenizkiy
Eindhoven University of Technology
Netherlands
m.pechenizkiy@tue.nl

Decebal C. Mocanu
University of Luxembourg
Luxembourg
decebal.mocanu@uni.lu

Matthew E. Taylor
University of Alberta
Alberta Machine Intelligence Institute
Canada
matthew.e.taylor@ualberta.ca

ABSTRACT

To integrate into human-centered environments, autonomous agents must learn from and adapt to humans in their native settings. Preference-based reinforcement learning (PbRL) can enable this by learning reward functions from human preferences. However, humans live in a world full of diverse information, most of which is irrelevant to completing any particular task. It then becomes essential that agents learn to focus on the subset of task-relevant state features. To that end, this work proposes R2N (Robust-to-Noise), the first PbRL algorithm that leverages principles of dynamic sparse training to learn robust reward models that can focus on task-relevant features. In experiments with a simulated teacher, we demonstrate that R2N can adapt the sparse connectivity of its neural networks to focus on task-relevant features, enabling R2N to significantly outperform several sparse training and PbRL algorithms across simulated robotic environments.

KEYWORDS

Reinforcement learning; preference learning; sparse training

ACM Reference Format:
Calarina Muslimani, Bram Grooten, Deepak R.S. Mamillapalli, Mykola Pechenizkiy, Decebal C. Mocanu, and Matthew E. Taylor. 2025. Boosting Robustness in Preference-Based Reinforcement Learning with Dynamic Sparsity: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Recent advances in reinforcement learning (RL) are bringing us closer to a future in which RL agents aid humans in their daily lives [3, 5, 12]. Preference-based RL (PbRL) is a promising paradigm that allows RL agents to leverage human preferences to adapt their behavior to better align with human intentions [2, 7, 10]. However, to effectively integrate agents into human-centered environments, autonomous agents should be able to learn from humans in their natural settings. Unfortunately, human environments are inherently noisy. For example, suppose a household robot is tasked with learning to clean a toy room and a human provides the robot with preferences on how the room should be cleaned. In this scenario, the robot might receive distracting information, such as the sounds of children playing. Only a subset of the robot's perceptions is relevant to the task, and identifying this subset can boost performance. To that end, we present R2N, a novel *robust-to-noise* PbRL algorithm that leverages principles of dynamic sparse training (DST) to learn robust reward models in *extremely noisy environments*. R2N continually adjusts the network topology of both the reward model and RL agent networks to focus on task-relevant features.

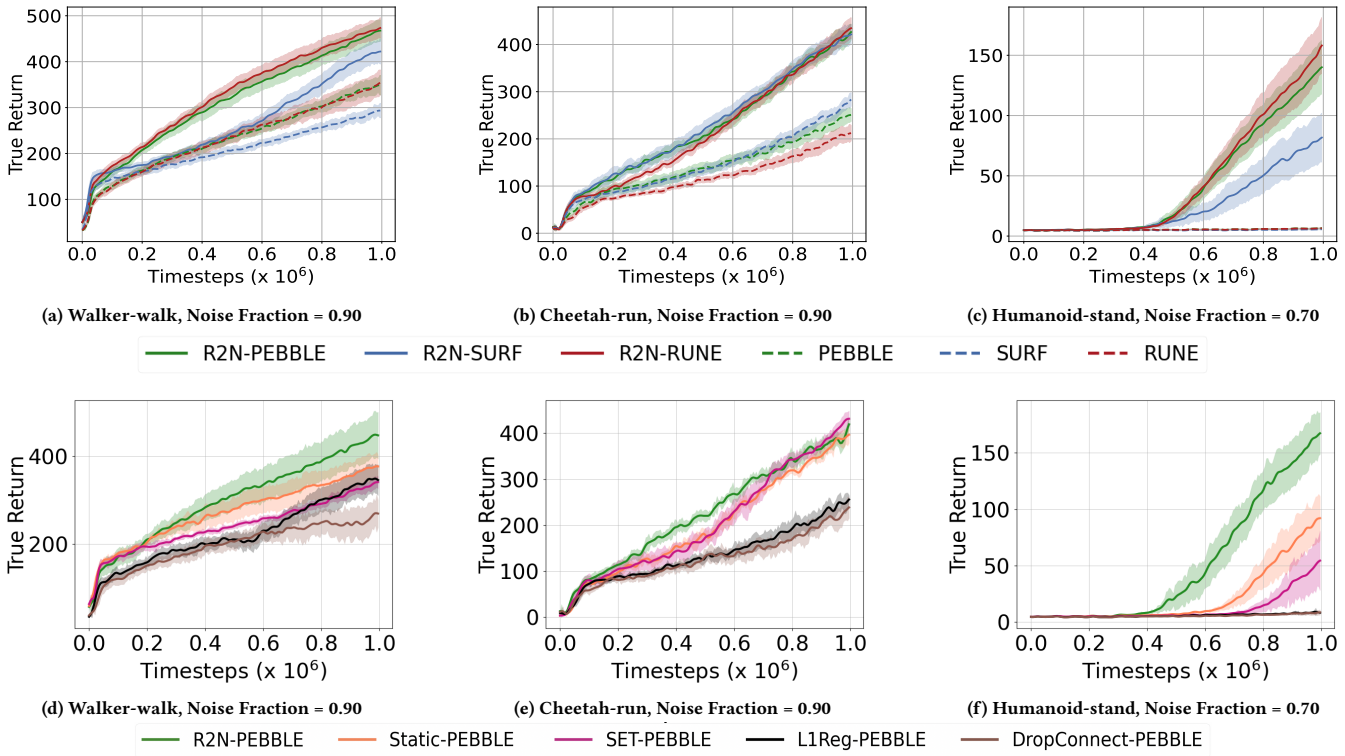
Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 BACKGROUND

This work assumes an MDP\R setting, where access to the environmental reward function is not provided. The goal is to learn a good policy while simultaneously estimating a reward function, $\hat{r}_\theta$, from human preferences. PbRL considers trajectory segments σ, where each segment consists of a sequence of state-action pairs. The teacher compares two segments, σ^0 and σ^1, assigning $y = 1$ if σ^1 is preferred, $y = 0$ if σ^0 is preferred, and $y = 0.5$ if both are equally preferred. As feedback is provided, it is stored as tuples (σ^0, σ^1, y) in a dataset. Then, we follow the Bradley-Terry model [1] to define a preference predictor P_θ using the reward function estimator, $\hat{r}_\theta$. Intuitively, if segment σ^i is preferred over segment σ^j, then the cumulative predicted reward (under $\hat{r}_\theta$) for σ^i should be greater than for σ^j. To train the reward function, we can use supervised learning where the teacher provides the labels y. We can then update $\hat{r}_\theta$ by minimizing the binary cross-entropy objective. The learned reward function, $\hat{r}_\theta$, is then used in place of the environmental reward function in the typical RL interaction loop.

3 ROBUST-TO-NOISE PBRL

The goal of R2N is to learn reward functions from feedback in environments with many task-irrelevant features. To achieve this, R2N applies DST techniques to PbRL algorithms to enable the learned reward model to focus on relevant features. R2N consists



(a) Walker-walk, Noise Fraction = 0.90 (b) Cheetah-run, Noise Fraction = 0.90 (c) Humanoid-stand, Noise Fraction = 0.70

(d) Walker-walk, Noise Fraction = 0.90 (e) Cheetah-run, Noise Fraction = 0.90 (f) Humanoid-stand, Noise Fraction = 0.70

Figure 1: Learning curves comparing R2N against various PbRL algorithms (top row) and sparse training methods (bottom row).

of two primary steps. First, at initialization, R2N randomly prunes the input layer of the reward model to a pre-defined sparsity level s^R. This is important, as prior works have shown that sparse neural networks can outperform their dense counterparts in both the supervised learning and RL settings [4, 9, 14]. Second, after every ΔT^R weight updates in the training loop, we prune the weakest active connections in the reward model's input layer. The strength of a connection is defined by the absolute value of its weight. After dropping a fraction $d_f^R \in (0, 1)$ of active connections, R2N regrows an equal number in new locations, maintaining a consistent sparsity level throughout training. To choose which inactive connections to grow, we use RigL [4], which activates connections with the highest gradient magnitude. We also apply the DST procedure, SET [9], to the input layers of the actor and critic networks in the RL agent, following a noise-filtering algorithm for standard RL [6].

4 EXPERIMENTS

We consider the Extremely Noisy Environment (ENE) [6], where noise is treated as a distracting feature in the environment. Specifically, the state space is expanded such that a fraction $n_f \in [0, 1)$ of the total state space consists of noise features. The PbRL algorithms must identify the most relevant features to (1) learn a robust reward function and (2) learn an adequate policy. We evaluate R2N in three DMControl environments [15]: Walker-walk, Cheetah-run, and Humanoid-stand. To assess its effectiveness, we compare R2N against four sparse training baselines: SET [9], Static Sparse Training, L1 Regularization [11], and DropConnect [16], each integrated into the reward learning module of the PbRL algorithm PEBBLE [7]. To further examine R2N's applicability across diverse PbRL algorithms, we integrate it with two additional approaches: SURF [13] and RUNE [8]. For PbRL baselines, we use a simulated teacher that provides preferences between trajectory segments based on the ground truth reward function. We train all algorithms for 1 million timesteps and evaluate performance every 5000 timesteps. Evaluation is based on the average offline performance over ten episodes using the ground truth reward function. Results are averaged over 14 or 5 seeds (Figure 1—top and bottom, respectively), with shaded regions representing the standard error.

In all environments, R2N-PEBBLE is the only algorithm that *consistently* achieves superior performance. Furthermore, in Figure 1 - top, the addition of R2N significantly improves both the learning efficiency and final return of the base PbRL algorithm. In Figure 1 - bottom, R2N-PEBBLE significantly outperforms L1-Regularization and DropConnect in learning efficiency. While Static-PEBBLE and SET-PEBBLE are more competitive, R2N maintains a performance advantage in Humanoid-stand and Cheetah-run.

ACKNOWLEDGMENTS

Part of this work has taken place in the Intelligent Robot Learning Lab at the University of Alberta, which is supported in part by research grants from Alberta Innovates; Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; Digital Research Alliance of Canada; Mitacs; and the National Science and Engineering Research Council.

REFERENCES

[1] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* (1952).

[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *The Conference on Neural Information Processing Systems*.

[3] Yogesh K. Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, Vassilis Galanos, P. Vigneswara Ilavarasan, Marijn Janssen, Paul Jones, Arpan Kumar Kar, Hatice Kizgin, Bianca Kronemann, Banita Lal, Biagio Lucini, Rony Medaglia, Kenneth Le Meunier-FitzHugh, Leslie Caroline Le Meunier-FitzHugh, Santosh Misra, Emmanuel Mogaji, Sujeet Kumar Sharma, Jang Bahadur Singh, Vishnupriya Raghavan, Ramakrishnan Raman, Nripendra P. Rana, Spyridon Samothrakis, Jak Spencer, Kuttimani Tamilmani, Annie Tubadji, Paul Walton, and Michael D. Williams. 2021. Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. *International Journal of Information Management* (2021).

[4] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. 2020. Rigging the Lottery: Making All Tickets Winners. In *The International Conference on Machine Learning*.

[5] Google Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models.

[6] Bram Grooten, Ghada Sokar, Shibhansh Dohare, Elena Mocanu, Matthew E. Taylor, Mykola Pechenizkiy, and Decebal Constantin Mocanu. 2023. Automatic Noise Filtering with Dynamic Sparse Training in Deep Reinforcement Learning. In *The International Conference on Autonomous Agents and Multiagent Systems*.

[7] Kimin Lee, Laura M Smith, and Pieter Abbeel. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In *The International Conference on Machine Learning*.

[8] Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. 2022. Reward Uncertainty for Exploration in Preference-based Reinforcement Learning. In *The International Conference on Learning Representations*.

[9] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. 2018. Scalable Training of Artificial Neural Networks with Adaptive Sparse Connectivity inspired by Network Science. *Nature Communications* (2018).

[10] Calarina Muslimani and Matthew E Taylor. 2024. Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning (extended abstract). In *The International Conference on Autonomous Agents and Multiagent Systems*.

[11] Andrew Y. Ng. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *The International Conference on Machine Learning*.

[12] OpenAI. 2023. GPT-4 Technical Report.

[13] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2022. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. In *The International Conference on Learning Representations*.

[14] Ghada Sokar, Elena Mocanu, Decebal Constantin Mocanu, Mykola Pechenizkiy, and Peter Stone. 2022. Dynamic Sparse Training for Deep Reinforcement Learning. In *The International Joint Conference on Artificial Intelligence*.

[15] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. 2018. Deepmind Control Suite.

[16] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of Neural Networks using DropConnect. In *The International Conference on Machine Learning*.