TA-Explore: Teacher-Assisted Exploration for Facilitating Fast Reinforcement Learning

Extended Abstract

Ali Beikmohammadi
Department of Computer and Systems Sciences
Stockholm University
Stockholm, Sweden
beikmohammadi@dsv.su.se

Sindri Magnússon
Department of Computer and Systems Sciences
Stockholm University
Stockholm, Sweden
sindri.magnusson@dsv.su.se

ABSTRACT

Reinforcement Learning (RL) is crucial for data-driven decision-making but suffers from sample inefficiency. This poses a risk to system safety and can be costly in real-world environments with physical interactions. This paper proposes a human-inspired framework to improve the sample efficiency of RL algorithms, which gradually provides the learning agent with simpler but similar tasks that progress toward the main task. The proposed method does not require pre-training and can be applied to any goal, environment, and RL algorithm, including value-based and policy-based methods, as well as tabular and deep-RL methods. The framework is evaluated on a Random Walk and optimal control problem with constraint, showing good performance in improving the sample efficiency of RL-learning algorithms.

KEYWORDS

Deep RL; Policy Optimization; PPO; Sample Efficiency; Exploration

ACM Reference Format:
Ali Beikmohammadi and Sindri Magnússon. 2023. TA-Explore: Teacher-Assisted Exploration for Facilitating Fast Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

RL needs many samples to learn good decision policies, hindering its potential in many areas [26, 28, 32, 34]. Agents need expensive interactions with the environment to learn, especially in real-world settings such as cyber-physical, medical, and robotic systems, where slow learning can be dangerous [2, 7, 31]. Efficient exploration and learning are crucial for RL to overcome this challenge.

RL exploration has been well studied, with many efforts focused on high-probability state visitation for successful, sample-efficient learning. Two fundamental questions arise in this setting: what should agents look for in the absence of rewards, and when should they stop exploring and start acting greedily? Research in this area can be divided into two categories: single-environment approaches (e.g., *visitation counts* [6, 25], *optimism* [3, 4, 14], *curiosity* [12, 20, 23, 24], and *reward shaping* [8, 10, 16]) and multi-environment approaches (e.g., *transfer learning* [1, 15, 19, 30, 33], *continual learning* [13], *meta-learning* [9], and *curriculum learning* [17, 18]). These approaches are discussed and contrasted with our contributions in the full version of this paper [5].

Human learning and exploration in new environments differ fundamentally from RL agents. Humans approach difficult and nested tasks by breaking them down into step-by-step learning of short-term, aligned auxiliary goals [11]. Similarly, before tackling the main goal, RL agents can benefit from solving simpler, related tasks that gradually increase in difficulty and progress towards it. This paper aims to provide a human-inspired framework for Teacher-Assisted exploration in RL, facilitating learning by gradually providing simpler auxiliary goals that converge to the main goal, independent of the algorithm used. Auxiliary goals are achieved by defining an assistant reward (not many and not following from a particular distribution) , the target reward, and an annealing function to sequence the associated Markov decision processes (MDPs). The approach eliminates the need for defining many separate MDPs. The agent shifts between the auxiliary goals, learning each of them for one iteration (without having ϵ-accuracy with δ probability assumption) , and uses that knowledge when learning the main task. The effectiveness of the approach is demonstrated through experiments with tabular algorithm and deep RL algorithm on a real-world problem, which indicates that it speeds up learning without increasing computational complexity. The experiments' code is publicly available at https://github.com/AliBeikmohammadi/TA-Explore.

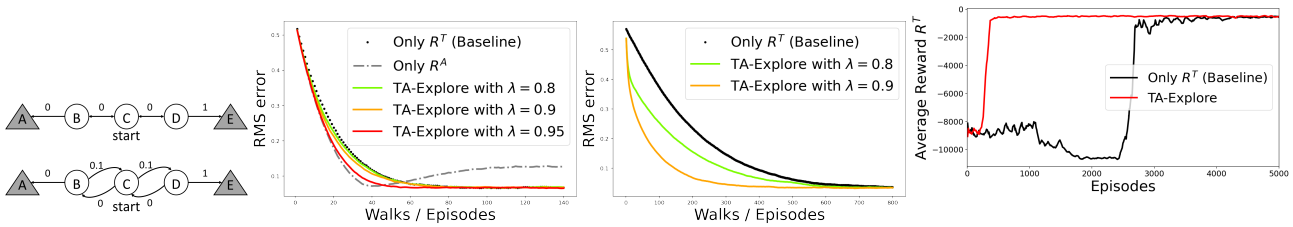
2 OUR FRAMEWORK: TA-EXPLORE

Formally, a MDP is characterized by a 5-tuple (S, A, P, R^T, γ) where S denotes the set of states; A denotes the set of actions; $P : S \times A \to \Delta(S)$ denotes the transition probability from state $s \in S$ to state $s' \in S$ when the action $a \in A$ is taken; $R^T : S \times A \times S \to \mathbb{R}$ is the immediate reward received by the agent after transitioning from (s, a) to s'; $\gamma \in [0, 1)$ is the discount factor . The agent makes decision by following a parameterized policy $\pi : S \times \Theta \to \Delta(A)$. In particular, we have $a_t \sim \pi(\cdot|s_t; \theta)$ where $\theta \in \Theta$ is an adjustable parameter. The goal of the agent is to find the policy $\pi(\cdot|s_t; \theta)$ by tuning the parameter θ that optimizes the cumulative reward

$$\text{Main Goal: } M(\theta) := \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R^T(s_t, a_t, s_{t+1}) \middle| a_t \sim \pi(.|s_t, \theta), s_0 \sim \mu \right].$$

Just like with human learning, we may consider some auxiliary goal $A(\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R^A(s_t, a_t, s_{t+1}) \middle| a_t \sim \pi(.|s_t, \theta), s_0 \sim \mu \right]$, where R^A is an assistant reward. Ideally, we should choose R^A in a way that a) it results in a simple RL problem that can be solved

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.



(a) (b) 5 states (c) 33 states (d)

Figure 1: (a) Random Walk example [27], where (top) describes how to receive R^T and (bottom) illustrates how to acquire R^A; (b) and (c) the main goal learning curves in Random Walk for different λ values and the different number of states; (d) average performance on optimal temperate control with constraints (i.e, data center cooling task [21]).

fast, and b) solving it is a side-step towards the main goal. The goal of the agent is not to learn the auxiliary goal, but rather to use it to facilitate learning. Thus it is important that the agent can gradually put more effort into the main goal $M(\theta)$. To that end, let $\beta(e)$ denote the parameter that controls how quickly the agents progress towards the main goal, where $e \in \mathbb{N}$ is the episode index. Then, during episode e, the agent uses the immediate reward $R_e(s_t, a_t, s_{t+1}) = \beta(e)R^A(s_t, a_t, s_{t+1}) + (1-\beta(e))R^T(s_t, a_t, s_{t+1})$.

Finding a suitable $\beta(e)$ is clearly an important part of making the learning progress smooth, which also determines the number of MDPs. In general, $\beta(e)$ should be decreasing in e and should converge to 0 as e increases, to ensure convergence to the main goal. How fast $\beta(e)$ decreases depends on the task, and hence a specific and unique formula cannot be stated for it. In general, how well the main goal and auxiliary goal align can be a good clue for how to select $\beta(e)$. Specifically, if they are well aligned, then $\beta(e)$ can be decreased at a slower rate, e.g., linear rate, by setting $\beta(e) = [(E - e)\beta(0)/E]_+$. Here $\beta(e)$ starts to gradually decrease from $\beta(0)$ and reaches zero at episode E. Conversely, if the alignment between these two goals is low then it is better to decrease $\beta(e)$ at a faster rate. For example, in that case we might have $\beta(e)$ decrease exponentially fast, i.e., $\beta(e) = \beta(0)\lambda^e$, where $\lambda \in (0, 1)$. For a more detailed explanation of the method, please see the full version of this paper [5].

3 EXPERIMENTS

First, we demonstrate the efficiency and simplicity of `TA-Explore` by examining a simple Random Walk example (Figure 1a) [27]. The agent's main goal is to learn the value of each state through experience since there are no actions involved. However, this process is slow because the agent only receives feedback in the form of a non-zero reward when it reaches the right terminal state.

However, we might facilitate learning by providing the agent with simpler tasks that provide immediate feedback more frequently. For example, as shown in Figure 1a (bottom), we consider the assistant reward R^A that provides the immediate reward 0.1 every time the agent goes to the right except when it reaches the terminal state on the right then we give 1 as the immediate reward. Also, a simple but intelligent idea for determining the $\beta(e)$ function could be $\beta(e) = \beta(0)\lambda^e$, where $\beta(0) = 1$, and $\lambda \in \{0.8, 0.9, 0.95\}$. The agent could first start to learn goal A but focus very quickly on learning goal M. We use TD(0) method [29] with a constant step size of 0.1.

So in this example, transfer learning is taking place through value transfer. This means that the value functions obtained for each task are used as the initial values of the next task. We report the average results of 100 times tests. As shown in Figure 1b, if we learn only A, it is observed that by around episode 38, the agent is surprisingly learning M as well. But then, as the agent tries to learn A more, the error increases. But the speed of learning A is faster than learning M. Hence if the agent starts learning A first, but before over-fitting on it, starts learning M, it can get the most out of prior information. Then, it learns the main goal M much faster. As the number of states increases to 33 in Figure 1c, the area enclosed between the baseline and `TA-Explore` increases more, which means the superiority of our proposed method becomes more noticeable.

Next, we illustrate the potential of `TA-Explore` by applying it to an optimal control problem with constraints - specifically, the task of data center cooling [21]. In RL, constraints are commonly included in the reward (i.e., as the main goal M), requiring the agent to learn both constraint satisfaction and reward optimization. However, constraint satisfaction is easier as the agent has more action choices leading to it. Thus, we propose utilizing a negative assistant reward R^A to make the agent learn to satisfy the constraint as an auxiliary goal A in our framework. We define the target reward R^T and assistant reward R^A as follows in our framework:

$$R^T = \begin{cases} -10\|a\|^2 \\ -10\|a\|^2 - 100 \end{cases} \qquad R^A = \begin{cases} 0 & \text{if constraint is satisfied} \\ -100 & \text{otherwise} \end{cases}$$

The R^A is part of the target reward R^T, allowing for high alignment between the two rewards, resulting in a $\beta(e)$ function of $\beta(e) = [(E - e)\beta(0)/E]_+$. To train the agent, we select PPO, a deep-RL approach, as the backbone [22]. Weights obtained in each episode for each task are used as the initial weighting of the neural network for the next task, enabling transfer learning through policy transfer.

Figure 1d shows that `TA-Explore` outperforms the baseline PPO method, demonstrating faster convergence. This is achieved by learning the assistant reward R^A in the initial episodes, which prevents confusion for the agent when faced with complex reward signals that could be due to constraint violations or being far from the main goal. For further details and discussion about the experiments conducted, as well as additional experiments, please refer to the full version of the paper [5].

ACKNOWLEDGMENTS

This work was partially supported by the Swedish Research Council through grant agreement no. 2020-03607 and in part by Digital Futures, the C3.ai Digital Transformation Institute, and Sweden's Innovation Agency (Vinnova). The computations were enabled by resources in project SNIC 2022/22-942 provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

REFERENCES

[1] David Abel, Yuu Jinnai, Sophie Yue Guo, George Konidaris, and Michael Littman. 2018. Policy and value transfer in lifelong reinforcement learning. In *International Conference on Machine Learning*. PMLR, 20–29.

[2] Greg Anderson, Abhinav Verma, Isil Dillig, and Swarat Chaudhuri. 2020. Neurosymbolic Reinforcement Learning with Formally Verified Exploration. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6172–6183. https://proceedings.neurips.cc/paper/2020/file/448d5eda79895153938a8431919f4c9f-Paper.pdf

[3] Peter Auer, Thomas Jaksch, and Ronald Ortner. 2009. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), Vol. 21. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf

[4] Peter Auer and Ronald Ortner. 2007. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. In *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press. https://proceedings.neurips.cc/paper/2006/file/c1b70d965ca504aa751ddb62ad69c63f-Paper.pdf

[5] Ali Beikmohammadi and Sindri Magnússon. 2023. Human-Inspired Framework to Accelerate Reinforcement Learning. *Preprint* (2023).

[6] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/afda332245e2af431fb7b672a68b659d-Paper.pdf

[7] Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. 2019. End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 3387–3395. https://doi.org/10.1609/aaai.v33i01.33013387

[8] Michael Dann, Fabio Zambetta, and John Thangarajah. 2019. Deriving Subgoals Autonomously to Accelerate Learning in Sparse Reward Domains. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 881–889. https://doi.org/10.1609/aaai.v33i01.3301881

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[10] Zhao-Yang Fu, De-Chuan Zhan, Xin-Chun Li, and Yi-Xing Lu. 2019. Automatic Successive Reinforcement Learning with Multiple Auxiliary Rewards. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2336–2342. https://doi.org/10.24963/ijcai.2019/324

[11] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. 2013. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences* 17, 11 (2013), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

[12] Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. VIME: Variational Information Maximizing Exploration. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/abd815286ba1007abfbb8415b83ae2cf-Paper.pdf

[13] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526. https://doi.org/10.1073/pnas.1611835114 arXiv:https://www.pnas.org/content/114/13/3521.full.pdf

[14] Tze Leung Lai, Herbert Robbins, et al. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[15] Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L Littman. 2021. Lipschitz lifelong reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8270–8278.

[16] Yiming Liu and Zheng Hu. 2020. The Guiding Role of Reward Based on Phased Goal in Reinforcement Learning. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing* (Shenzhen, China) *(ICMLC 2020)*. Association for Computing Machinery, New York, NY, USA, 535–541. https://doi.org/10.1145/3383972.3384039

[17] Sha Luo, Hamidreza Kasaei, and Lambert Schomaker. 2020. Accelerating reinforcement learning for reaching using continuous curriculum learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[18] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *arXiv preprint arXiv:2003.04960* (2020).

[19] Simone Parisi, Victoria Dean, Deepak Pathak, and Abhinav Gupta. 2021. Interesting Object, Curious Agent: Learning Task-Agnostic Exploration. *Advances in Neural Information Processing Systems* 34 (2021).

[20] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2778–2787. https://proceedings.mlr.press/v70/pathak17a.html

[21] Benjamin Recht. 2019. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), 253–279.

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[23] Matthias Schultheis, Boris Belousov, Hany Abdulsamad, and Jan Peters. 2020. Receding Horizon Curiosity. In *Proceedings of the Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 100)*, Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (Eds.). PMLR, 1278–1288. https://proceedings.mlr.press/v100/schultheis20a.html

[24] Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. arXiv:1507.00814 [cs.AI]

[25] Alexander L. Strehl and Michael L. Littman. 2008. An analysis of model-based Interval Estimation for Markov Decision Processes. *J. Comput. System Sci.* 74, 8 (2008), 1309–1331. https://doi.org/10.1016/j.jcss.2007.08.009 Learning Theory 2005.

[26] Flood Sung, Li Zhang, Tao Xiang, Timothy Hospedales, and Yongxin Yang. 2017. Learning to Learn: Meta-Critic Networks for Sample Efficient Learning. arXiv:1706.09529 [cs.LG]

[27] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*.

[28] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).

[29] Gerald Tesauro et al. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[30] Andrea Tirinzoni, Riccardo Poiani, and Marcello Restelli. 2020. Sequential transfer in reinforcement learning with a generative model. In *International Conference on Machine Learning*. PMLR, 9481–9492.

[31] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. 2020. Safe Reinforcement Learning via Curriculum Induction. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12151–12162. https://proceedings.neurips.cc/paper/2020/file/8df6a65941e4c9da40a4fb899de65c55-Paper.pdf

[32] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q -Learning. *Mach. Learn.* 8, 3–4 (may 1992), 279–292. https://doi.org/10.1007/BF00992698

[33] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.

[34] Yang Yu. 2018. Towards Sample Efficient Reinforcement Learning.. In *IJCAI*. 5739–5743.