Two-Level Actor-Critic Using Multiple Teachers

Extended Abstract

Su Zhang
Washington State University
Pullman, United States
su.zhang2@wsu.edu

Srijita Das
University of Alberta
Edmonton, Canada
srijita1@ualberta.ca

Sriram Ganapathi Subramanian
Vector Institute, Toronto, Canada
University of Waterloo, Waterloo, Canada
sriram.subramanian@vectorinstitute.ai

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

KEYWORDS

Reinforcement Learning, Teaching RL Agents, Simulated Robotics

ACM Reference Format:
Su Zhang, Srijita Das, Sriram Ganapathi Subramanian, and Matthew E. Taylor. 2023. Two-Level Actor-Critic Using Multiple Teachers: 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

Reinforcement learning (RL) has been successful in a variety of domains ranging from solving difficult games like Go [10] and drug discovery [4]. Most of these domains are characterized by high-dimensional states and continuous action spaces. However, sample inefficiency is a major challenge when applying these algorithms to real-world tasks such as robotics and healthcare care [6]. To address improved sample efficiency, rather than forcing agents to learn from scratch, domain knowledge from humans or existing agents can be leveraged in various ways [3].

Many existing approaches can leverage advice in RL from a single, near-optimal teacher [1, 5, 11]. In this work, we consider settings where a student can receive *action advice from multiple teachers*. This setting can be particularly appropriate when different teachers have different skills (e.g., teachers may perform well in different parts of the state space or perform different sub-tasks). In addition, we also consider that teachers may be suboptimal, or even random. This allows us to leverage teachers that perform well while not being hurt (much) by teachers that perform poorly. This paper focuses on the question, "When should the student listen to which teacher?" to effectively use the best teacher's policy for a given state, or to decide not to listen to any teacher.[1]

Our work is inspired by Two-Level Q-Learning (TLQL) [8] but is different in two ways: (1) it can be applied to a variety of tasks having both continuous and discrete action spaces, and (2) is more resistant to teachers of different qualities ranging from fully optimal to partially suboptimal teachers. In addition, by using two actor-critic networks, our approach can be easily incorporated into existing actor-critic algorithms.

[1]This work assumes teachers do not learn and are not antagonistic.

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.

2 TWO-LEVEL ACTOR-CRITIC USING MULTIPLE TEACHERS

We introduce the *Two-Level Actor-Critic* (TL-AC) method to learn from multiple teachers. TL-AC extends the actor-critic algorithm[9] to a two-level network structure, with a single critic for both levels, enabling the agent to leverage teachers with different expertise and advice quality.

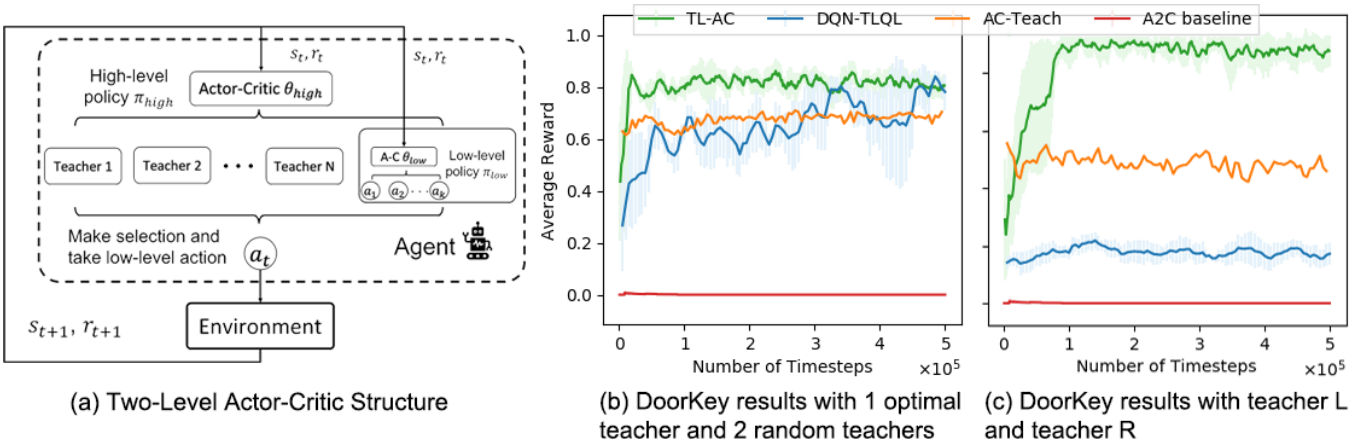
2.1 Problem Statements

Given: A set of (sub)optimal teachers as denoted by $E_0 = \{e_1, e_2, \cdots, e_N\}$ where $|E_0| = N$, with a set of corresponding policies denoted by $\pi_E = \{\pi_{e_1}, \pi_{e_2}, \cdots, \pi_{e_N}\}$.
Objective: Train the learning agent using action advice from the set of multiple teacher policies π_E when useful to make the agent learn a good policy with fewer environmental interactions.
Assumptions: We consider a single-agent learning problem with multiple teachers. The teachers are fixed during the agent training process.

The multiple teachers used in our approach can be pre-trained agents, classifiers trained from demonstrations, human teachers, etc. At each time-step t, each individual teacher e_i provides action advice (optimal or sub-optimal) based on the state vector s_t. The student agent will choose an action to execute, which can be either its own policy or the policy of any one of the several teachers in the set E. The goal of the learning agent is to maximize the return, while determining which teachers to listen to when.

2.2 Algorithm and Methodology

Figure 1(a) shows the TL-AC structure.
Low-level Policy: Select Action The low-level policy network is the first actor-critic network, parameterized with θ_{low}. The low-level policy $\pi_{low}(s, a) \doteq \pi(a|s; \theta_{low})$ maps the states to a probability distribution over actions. The advantage of taking action a_t at state s_t is $A(s_t, a_t) := r(s_t, a_t) + \gamma V_{\pi_{low}}(s_{t+1}) - V_{\pi_{low}}(s_t)$. The objective is to maximize the agent's value over all states and find



(a) Two-Level Actor-Critic Structure

(b) DoorKey results with 1 optimal teacher and 2 random teachers

(c) DoorKey results with teacher L and teacher R

Figure 1: (a) Two-Level Actor-Critic Structure and (b), (c) show the capability of TL-AC incorporating multiple teachers with different qualities or expertise in the DoorKey environment.

the low-level optimal policy with the loss function[2]

$$\mathcal{L}(\theta_{low}) = \log \pi_{low}(a|s;\theta_{low})A(s,a).$$

High-level Policy: Select Teacher and Take Advice The high-level policy network is parameterized with θ_{high}. The high-level policy $\pi_{high}(s,e) \doteq \pi(e|s;\theta_{high})$ maps the states to a probability distribution over teachers. The high-level network estimates the expected return of selecting each teacher e. The reward for a teacher is the environmental reward received when executing the chosen teacher's action advice. Hence, we have $r(s_t,e_t) = r(s_t,e_t,a_t) = r(s_t,a_t)$. The discounted value $V_{\pi_{high}}(s)$ at state s under the high-level policy π_{high} is equal to the $V_{\pi_{low}(s)}$. The advantage of choosing the teacher e_t in the state s_t at the high level can be written as

$$\begin{aligned} A(s_t,e_t) &:= r(s_t,e_t) + \gamma V_{\pi_{high}}(s_{t+1}) - V_{\pi_{high}}(s_t) \\ &= r(s_t,a_t) + \gamma V_{\pi_{low}}(s_{t+1}) - V_{\pi_{low}}(s_t) = A(s_t,a_t) \end{aligned}$$

This shows that the networks at both levels use the same critic for estimating the value function of the states. The objective is to maximize the values of all states and to find a (near) optimal high-level policy with the loss function $\mathcal{L}(\theta_{high}) = \log \pi_{high}(e|s;\theta_{high})A(s,a)$. Zhang et al. [12] showed a similar reduction in an option learning framework.

3 EXPERIMENTS

3.1 Experimental Settings

We use the A2C algorithm without any advice as the baseline. We also benchmark against the DQN variant Two-level Q-learning (DQN-TLQL) [8] and AC-Teach [7], both of which can learn from multiple teachers.

Here we present experimental results on the MiniGrid DoorKey [2]) task. This is a discrete grid room environment; the rooms are separated by a wall and the agent needs to use the key to open the door and enter the other half to get to the target grid. The state is a 3-tuple vector and there are 6 discrete actions. A (sparse) positive reward is given when the goal is reached, otherwise it is 0.

[2]The subscript 't' is dropped from notation for convenience because the loss-function is defined with respect to a single mini-batch.

Teacher Set Details: To construct the teacher set, we first obtain a near-optimal policy, then construct *partial teachers*, who could only provide good advice on partial state space. We have *teacher L* to give advice on how to pick up the key and open the door when at the left room, and *teacher R* to give advice on navigating to the target square when at the right room.

3.2 Results and Discussion

We plot the learning curve to show the *mean training reward* as well as the standard deviation. Figure 1(b) shows the results of providing one optimal full teacher and two random full teachers in the teacher set. TL-AC can effectively learn from multiple teachers of different quality as compared to other baselines in this discrete task, and is robust to the effect of noisy teachers' advice.

In Figure 1(c), when allowing both teacher L and teacher R to provide advice to the agent, TL-AC could learn a near-optimal policy within 1×10^5 steps while AC-Teach and DQN-TLQL could not benefit much from these settings. This indicates TL-AC could efficiently incorporate the advice from multiple teachers with different areas of expertise. Additional details, analyses, and ablation studies with different teacher sets and domains are excluded from this extended abstract.

To summarize, advantages of using TL-AC include: (1) being lightweight and having a simple structure, (2) easily switching between policies at every time-step to incorporate the best teacher's advice, (3) simple incorporation into any actor-critic algorithm, and (4) working with both full and partial teachers. Future work includes conducting experiments with human teachers and investigating the possibility of incorporating uncertainty and confidence-related schemes into this framework.

ACKNOWLEDGMENTS

Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; Compute Canada; Huawei; Mitacs; and NSERC.

REFERENCES

[1] Adam Bignold, Francisco Cruz, Matthew E Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021. A conceptual framework for externally-influenced agents: An assisted reinforcement learning review. *Journal of Ambient Intelligence and Humanized Computing* (2021), 1–24.

[2] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. 2018. *Minimalistic Gridworld Environment for Gymnasium.* https://github.com/Farama-Foundation/Minigrid

[3] Felipe Leno Da Silva and Anna Helena Reali Costa. 2019. A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research* (2019).

[4] Sai Krishna Gottipati, Yashaswi Pathak, Boris Sattarov, Sahir, Rohan Nuttall, Mohammad Amini, Matthew E. Taylor, and Sarath Chandar. 2021. Towered Actor Critic For Handling Multiple Action Types In Reinforcement Learning For Drug Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 142–150.

[5] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. 2018. Deep Q-learning from demonstrations. In *AAAI*.

[6] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. 2021. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research* (2021).

[7] Andrey Kurenkov, Ajay Mandlekar, Roberto Martin-Martin, Silvio Savarese, and Animesh Garg. 2019. AC-Teach: A bayesian actor-critic method for policy learning with an ensemble of suboptimal teachers. *CoRL* (2019).

[8] Mao Li, Yi Wei, and Daniel Kudenko. 2019. Two-level Q-learning: learning from conflict demonstrations. *The Knowledge Engineering Review* (2019).

[9] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[10] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* (2016).

[11] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[12] Shangtong Zhang and Shimon Whiteson. 2019. DAC: The double actor-critic architecture for learning options. *arXiv preprint arXiv:1904.12691* (2019).