Project Proposal: Learning From Your Peers

Harvey Hu, Chirag Sharma

September 2022

1 Motivation

When learning to perform a complex task for which existing knowledge is limited, humans benefit significantly from having a peer group and they are often able to better navigate unfamiliar topics/environments through being influenced by their peers' thoughts and actions. More generally, communication between agents that are learning to perform the same task can lead to benefits due to knowledge/opinion sharing, which expands the learner's horizon beyond their own experiences and allows for shortcuts in learning. We seek to explore this peer influence in learning in the context of actor-critic methods in deep reinforcement learning (RL). In RL, having independent agents that are trained to optimize for the same task communicate with each other could potentially lead to faster and better results due to information transfer.

2 Hypothesis

We aim to show that communication between independent actor-critic (AC) networks that are optimizing for the same task, via connections between critic networks, increases the quality of the resulting policies and/or the speed of training.

3 Background

This project falls within the broader concepts of ensemble-based and multi-agent RL. Existing studies show that simple ensemble methods often yield better results for the ensemble than single agents [Weiring, M. A. & Hasselt, H. (2008)] but these typically involve simple aggregation of agent outputs without any stronger notion of communication. Some work in multi-agent RL has explored the effect of collaborative agents trained on the same task, but this has been restricted to communication between actors in a distributed AC scheme where the models are collectively trained to optimize the average of the values predicted by the critics [Pennesi, P. & Paschalidis, I. C. (2010)] and a distributed tabular Q-learning scheme, where each network's Q-value updates are influenced linearly by their deviation from the other networks' predicted Q-values [Kar, S. et al. (2012)]. However, as far as we know, our project will be the first to explore the setting where the training of critic networks in a distributed AC scheme includes passing in other critics' predictions as input variables.

4 Methods and Environments

Our general architecture consists of multiple vanilla AC networks, all of which are trained to optimize for the Humanoid task/environment in Open AI's Gym simulator. The actor networks are trained as usual using policy gradients, based on value functions predicted by the critic networks. Within this framework, we aim to explore two specific choices of communication between critics:

1. *Peer influence via value signals.* When training each critic network $V_\phi^\pi(\mathbf{s})$, we use the usual target (either batch target, online target, or GAE/n-step returns target). However, during training, we also add an additional encoder to the critic network that takes $\{V_{\phi'}^{\pi'}((\mathbf{s})) : (\pi', \phi') \neq (\pi, \phi)\}$ as inputs.

2. *Peer influence via compressed internal state.* We follow the same procedure as above, except that the encoder doesn't simply take the other predicted value functions as inputs, but rather takes as input embedding vectors corresponding to the internal states of each of the other critic networks $\{V_{\phi'}^{\pi'} : (\pi', \phi') \neq (\pi, \phi)\}$ – obtained via a decoder head attached to each critic network, which is trained via a reconstruction loss.

In both cases, the gradients while training the critic networks are not propagated through to the other networks – this should make the training problem easier and helps to maintain the independence of the networks. We will compare our results against those obtained by using standard ensembling techniques.