Multi-Source Policy Aggregation in Heterogeneous and Private Environmental Dynamics

Mohammadamin Barekatain*
Technical University of Munich
Munich, Germany
m.barekatain@tum.de

Ryo Yonetani
OMRON SINIC X
Tokyo, Japan
ryo.yonetani@sinicx.com

Masashi Hamaya
OMRON SINIC X
Tokyo, Japan
masashi.hamaya@sinicx.com

1. Introduction

We envision a future scenario where robotic agents working in diverse and private environments help a new agent in an unknown environment to learn its policy efficiently. For instance, imagine various types of pick-and-place robotic agents working in a factory. While the agents are involved in the same task, dynamics of the environment in which the task is performed is different based on each robot's kinematics (*e.g.*, degree of freedom, link length, and joint orientations) and dynamics (*e.g.*, joint damping, armature, and friction) [1]. Moreover, no prior knowledge about the dynamics of environments as well as the specification of agents policies can be available for a new agent due to the confidentiality of products and processes in the factory. Other relevant scenarios include autonomous vehicles on private land and home assistants interacting with people privately.

The problem setting shown above makes it hard to adopt many existing approaches for efficient learning of a target agent's policy. For instance, meta-learning approaches typically require an agent to be trained on a diverse task distribution [4], which is not possible here due to the privacy of environments. Also, existing transfer learning approaches that focus on the transfer of policies between dynamics, require prior information about the environments [1] or policy configuration (*e.g.*, actor network weights and agent's value function [2]) which are both unavailable.

In such scenarios, we argue that the target agent can get information from other private agents through their policies (hereafter *source* policies) that act as a black-box function mapping states to actions. Specifically, we propose a new sample efficient approach named *MULTI-source POLicy AggRegation (MULTIPOLAR)*. Much like a multipolar neuron that can integrate information coming from other neurons, our MULTIPOLAR aggregates the actions produced by the source policies to serve a robust baseline action. It also learns an additional policy to predict a 'residual' around the baseline actions to mitigate the unseen dynamics of the target agent's environment. As a result, MULTIPOLAR achieves training sample efficiency since the aggregation of source actions provides a strong inductive bias.

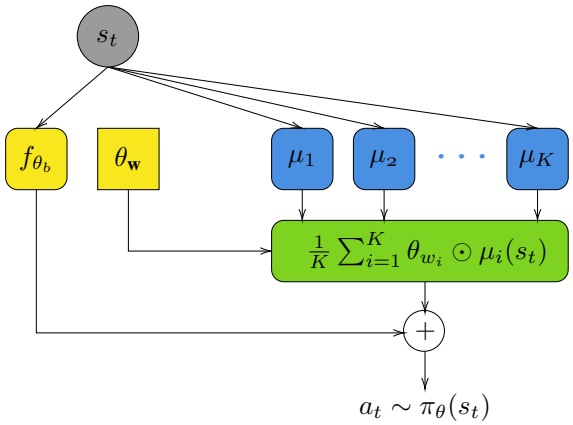

Figure 1: The proposed MULTIPOLAR policy network.

As a preliminary experiment, we evaluate MULTIPOLAR with two public simulated environments with continuous and discrete action spaces: Roboschool Hopper[1] and OpenAI Acrobot[2]. Our experimental results demonstrate that MULTIPOLAR allows a new agent to learn its policy significantly faster on average compared to when it is trained from scratch.

2. Proposed Method

Preliminaries We formulate our policy aggregation problem under the standard Reinforcement Learning (RL) framework, where an agent interacts with its environment modeled by a Markov Decision Process (MDP). An MDP is represented by a 6-tuple $(\rho_0, \gamma, \mathbb{S}, \mathbb{A}, R, T)$ where ρ_0 is the initial state distribution and $\gamma \in (0, 1]$ is the discount

*Work done as an intern at OMRON SINIC X

[1]https://github.com/openai/roboschool
[2]https://gym.openai.com/envs/Acrobot-v1