Semi-Centralised Multi-Agent Reinforcement Learning with Policy-Embedded Training

Taher Jafferjee[1,*], Juliusz Ziomek[1,*], Tianpei Yang[2], Zipeng Dai[1], Jianhong Wang[3], Matthew Taylor[2], Kun Shao[1], Jun Wang[4], David Mguni[†1]

[1]Huawei Noah's Ark Lab [2]University of Alberta
[3]Imperial College, London [4]University College, London

Abstract

Centralised training (CT) is the basis for many popular multi-agent reinforcement learning (MARL) methods because it allows agents to quickly learn high-performing policies. However, CT relies on agents learning from one-off observations of other agents' actions at a given state. Because MARL agents explore and update their policies during training, these observations often provide poor predictions about other agents' behaviour and the expected return for a given action. CT methods therefore suffer from high variance and error-prone estimates, harming learning. CT methods also suffer from explosive growth in complexity due to the reliance on global observations, unless strong factorisation restrictions are imposed (e.g., monotonic reward functions for QMIX). We address these challenges with a new *semi-centralised* MARL framework that performs *policy-embedded* training and decentralised execution. Our method, policy embedded reinforcement learning algorithm (PERLA), is an enhancement tool for Actor-Critic MARL algorithms that leverages a novel parameter sharing protocol and policy embedding method to maintain estimates that account for other agents' behaviour. Our theory proves PERLA dramatically reduces the variance in value estimates. Unlike various CT methods, PERLA, which seamlessly adopts MARL algorithms, scales easily with the number of agents without the need for restrictive factorisation assumptions. We demonstrate PERLA's superior empirical performance and efficient scaling in benchmark environments including *StarCraft Micromanagement II* and *Multi-agent Mujoco*.

1 Introduction

Multi-agent reinforcement learning (MARL) has emerged as a powerful tool to enable autonomous agents to tackle difficult tasks such as ride-sharing (Zhou et al. 2020) and swarm robotics (Mguni, Jennings, and de Cote 2018). Recently, various methodologies have produced significant performance boosts for MARL algorithms (Mguni et al. 2021b; Kuba et al. 2021). Nevertheless, an important impediment for MARL is the high variance of the critic and policy gradient estimators. Reducing this variance is a critical challenge since high variance estimates can significantly hinder training, leading to low sample efficiency and poor overall performance (Gu et al. 2016). This variance has multiple causes. First, MARL methods are applied to unknown environments whose reward signals are often noisy, especially as the sizes of the state and action spaces increases (Kuba et al. 2021). Second, unlike in single agent reinforcement learning (RL), MARL agents are faced with the challenge of distinguishing the aleatoric uncertainty due to environmental stochasticity from randomness due to agents' exploratory actions. These challenges can deeply undermine the performance of MARL methods, especially in centralised learning (CL) methods where agents rely on observations of others' actions while training.

CL serves as the foundation for many popular MARL methods such as MAPPO (Yu et al. 2021a), Q-DPP (Yang et al. 2020) QMIX (Rashid et al. 2018), SPOT-AC (Mguni et al. 2021a), and COMA (Foerster et al. 2018b). In CL, each agent has a (possibly shared) critic that makes use of all available information generated by the system, including the global state and the joint action (Peng et al. 2017). Accounting for actions of other agents is crucial to achieve good coordination. Despite the obvious advantages from using knowledge of the joint actions in the critic, this methodology has several weaknesses that exacerbate the MARL variance problem. The Q-functions are based on *one-off* observations of actions sampled from other agents' policies at a given state. This can produce VF updates based on improbable events and hence, inaccurate estimates of expected returns, leading to higher variance value function (VF) estimates. Also, since each agent's critic has an explicit dependence on the actions of others, the CL critic suffers from an explosive growth in complexity with the number of agents (Yang and Wang 2020). This results in CL methods needing large numbers of samples to complete training. Current methods for alleviating this rely on VF factorisations that require restrictive representational constraints. These constraints can cause poor exploration and performance failures when violated (Mahajan et al. 2019) (e.g., QMIX (Rashid et al. 2018) requires a monotonicity constraint that can produce suboptimal value approximation).

Because of these issues affecting CL, (decentralised) independent learning (IL) represents an attractive alternative (de Witt et al. 2020a). IL decomposes a MARL problem with N agents into N decentralised single-agent problems while the agents train solely with local observations. This avoids the explosive growth in complexity suffered by CL methods.

[*] Equal Contribution.

[†]Corresponding author: davidmguni@hotmail.com.