Multi Agent Meta Reinforcement Learning on Neural Networks

Alexander Cai, Oliver Cheng

December 13, 2022

1 Introduction to Meta-Reinforcement Learning

In recent years there has been a growing amount of excitement about *meta-learning* in order to solve a wider set of problems. This is best used in reinforcement learning tasks. In a standard reinforcement learning task, we have a policy that governs how an agent transitions between states in an environment. There are rewards that the agent can receive, and the optimal policy is one that maximizes rewards. We can formalize this by defining a reinforcement learning regime as a Markov Decision Process (MDP) where S is the set of states, A is the set of actions, T is the transition distribution (how states transition given an action), and R is the reward, which is a function of the state. Let π_θ be the policy, which is parameterized by some θ. Thus for some "task" $\mathcal{T}$, the goal is thus to maximize the objective

$$\mathcal{J}_\mathcal{T}(\theta) = \mathbb{E}_{a_t \sim \pi_\theta(s_t), s_{t+1} \sim T(s_t, a_t)} \left(\sum_t \gamma^t R(s_t, a_{t-1}) \right).$$

The task $\mathcal{T}$ determines the rewards, and hence the policy that is optimized.

Reinforcement learning is used to solve many problems such as Go, Chess, Atari games, etc. where the environment is unknown and is structured in a way which an agent must discover an optimal strategy. This framework gives us a way to view human beings and the nervous system as a reinforcement learning program where we humans are the agent. (Un)Fortunately, this is a bit too basic of a way of modeling human beings and the brain, as reinforcement learning networks have not taken over humanity yet (citation needed). One of the reasons is that rewards and the task that an agent needs to optimize a policy for is constantly changing in the real world. Thus we introduce the idea of meta-learning, where rather than optimizing a policy, we optimize for a learning strategy on how to optimize a policy for a distribution of tasks. This idea is similar to "learning how to learn" and is conceptually closer to how the brain works (that one prefrontal cortex Deepind paper.).

2 Reinforcement Learning in the Brain

In the Theory of Neural Computation, we can see reinforcement-learning-esque regimes. In particular researchers have found that the dopamine system seems to follow the TD algorithm, a reinforcement learning algorithm. In particular, the error function of the TD algorithm is operates in the same way to how dopamine is used to associate stimuli with future rewards. In this case, the dopamine neuron is the agent and it can change its firing rate as an action.

3 Implementation of Meta-Learning MARL

4 Comparison to Known Learning Rules

We compare this meta-learning (learning of the best learning rule) to accepted learning rules in the literature. Stochastic gradient descent converges the quickest, but due to the weight transport problem, is not biologically plausible (citation needed). Biologically plausible alternatives involve learning from a global error signal, which has been observed (citation needed), and involve perturbation-type learning rules. We will look at node and weight perturbation in particular.