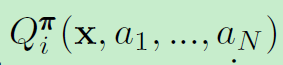
Problem

1. One issue is that each agent’s policy changes during training, resulting in a non-stationary environment and preventing the naïve application of experience replay (crucial for deep Q-learning to stabilize training).

Solution:

If we know the actions taken by all agents, the environment is stationary even as the policies change.

Centralized training. Centralized action-value function:



Each agent can have different reward structures.

1. The environment non-stationarity due to the agents’ changing policies is particularly true in competitive settings, where agents can derive a strong policy by overfitting to the behavior of their competitors. Such policies are undesirable as they are brittle and may fail when the competitors alter their strategies.

Solution:

Agents with Policy Ensembles.

Train a collection of K different sub-policies.

1. Policy gradient methods, on the other hand, usually exhibit very high variance when coordination of multiple agents is required.

The use of baselines, such as value function baselines typically used to ameliorate high variance, is problematic in multi-agent settings due to the non-stationarity issues mentioned previously.

4 Methods

4.1 Multi-Agent Actor Critic

multi-agent deep deterministic policy gradient (MADDPG)

Critic is augmented with extra information about the policies of other agents.

4.2 Inferring Policies of Other Agents

Online fashion?

4.3 Agents with Policy Ensembles

?

Network

Separate MLP RELU.

Why traditional RL methods fail?

Lack of a consistent gradient signal.

For example, if the speaker utters the correct symbol while the listener moves in the wrong direction, the speaker is penalized.

This problem is exacerbated as the number of time steps grows: we observed that traditional policy gradient methods can learn when the objective of the listener is simply to reconstruct the observation of the speaker in a single time step, or if the initial positions of agents and landmarks are fixed and evenly distributed.

This indicates that many of the multi-agent methods previously proposed for scenarios with short time horizons (e.g. [16]) may not generalize to more complex tasks.

5.3 Effect of Learning Polices of Other Agents

We evaluate the effectiveness of learning the policies of other agents in the cooperative communication environment, following the same hyperparameters as the previous experiments and setting lambda = 0.001 in Eq. 7. The results are shown in Figure 7. We observe that despite not fitting the policies of other agents perfectly (in particular, the approximate listener policy learned by the speaker has a fairly large KL divergence to the true policy), learning with approximated policies is able to achieve the same success rate as using the true policy, without a significant slowdown in convergence.

Downside

One downside to our approach is that the input space of Q grows linearly (depending on what information is contained in x) with the number of agents N. This could be remedied in practice by, for example, having a modular Q function that only considers agents in a certain neighborhood of a given agent. We leave this investigation to future work.