Problem

Centralised approach fail:

One agent learns a useful policy, but a second agent is discouraged from learning because its exploration would hinder the first agent and lead to worse team reward.

Design individual reward functions:

Reward shaping is difficult and only a small class of shaped reward functions are guaranteed to preserve optimality w.r.t. the true objective.

1.1 Related works

Discuss approaches how to deal with global objective vs local objective.

1. Method

Each agent only receives the joint reward.

We learn ~Qi by backpropagating gradients from the Q-learning rule using the joint reward through the summation, i.e. ~Qi is learned implicitly rather than from any reward specific to agent i, and we do not impose constraints that the ~Qi are action-value functions for any specific reward.

Advantage:

Don’t have to specify ri (reward shaping is difficult and may bring error).

4.1

Since observations are from a local perspective, we do not benefit from convolutional networks, but use a fully connected linear layer to process the observations.