Overview

- Studies problem of decentralized, cooperative multi-agent learning without explicit communication
- Independent agent updates induce a nonstationary environment for other agents
- Proposes DM2, a decentralized MARL algorithm that performs distribution matching against expert demonstrations to facilitate coordination
- Theoretical analysis shows that…
  - Individual distribution matching against coordinated expert demonstrations improves a lower bound on a joint imitation learning objective, leading to convergence
  - Expert policies are a Nash equilibrium for mixed task and distribution matching reward
- Experimental validation on StarCraft shows that the combined imitation and task reward improve on a fully decentralized baseline

DM2

- Each agent independently learns from a mixed reward that consists of the environment task reward and a distribution matching reward from GAIL [1]

\[
\pi_k \rightarrow \ldots \rightarrow \pi_k \\
\hat{p}_k \rightarrow \ldots \rightarrow \hat{p}_k \\
\hat{r}_{i,gail} + c \cdot \hat{r}_{env} \\
\]

- Agent and expert policies trained with IPPO [2]
- Expert demonstrations (state-only trajectories) are compatible (sampled concurrently from co-trained experts)

DM2 allows a team of RL agents to learn a cooperative task by independently imitating corresponding demonstrations from an expert team.

Experimental Results

- StarCraft II benchmark [3]
- Expert demonstrations sampled from trained IPPO or QMIX policies
- CTDE Baselines: QMIX [4], RMAPPO [2]
- Distribution Matching Baseline: DM2 w/SIL [5]


Which Demonstrations Work?

Legend

- Concurrent
- Nonconcurrent

<table>
<thead>
<tr>
<th>Demonstration Type</th>
<th>DM2</th>
<th>Not DM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concurrent</td>
<td>✔</td>
<td>X</td>
</tr>
<tr>
<td>Nonconcurrent</td>
<td>✔</td>
<td>X</td>
</tr>
</tbody>
</table>