DM²: Decentralized Multi-Agent Reinforcement Learning via Distribution Matching

Caroline Wang*¹
Ishan Durugkar*¹
Elad Liebman*²
Peter Stone¹, ³

¹ The University of Texas at Austin
² SparkCognition Research,
³ Sony AI

AAAI 2023

*Equal contribution
Motivation:

- Multi-agent reinforcement learning (MARL) is challenging — agents learning simultaneously makes the environment nonstationary

- Strategies:
  - Fully centralized learning
Motivation:

- Multi-agent reinforcement learning (MARL) is challenging — agents learning simultaneously makes the environment nonstationary.

- Strategies:
  - Fully centralized learning
  - Centralized training, decentralized execution (CTDE) \[^1\]

\[^1\] Sunehag et al., Value Decomposition Networks for Cooperative Multiagent learning, AAMAS 2018.
Motivation:

- Multi-agent reinforcement learning (MARL) is challenging — agents learning simultaneously makes the environment nonstationary

- Strategies:
  - Fully centralized learning
  - Centralized training, decentralized execution (CTDE) \(^1\)
  - Decentralized learning + communication \(^2\)

\(^1\) Sunehag et al., Value Decomposition Networks for Cooperative Multiagent learning, AAMAS 2018.
\(^2\) Jaques et al., Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning, ICML 2019.
Motivation:

- Multi-agent reinforcement learning (MARL) is challenging — agents learning simultaneously makes the environment nonstationary

- Strategies:
  - Fully centralized learning
  - Centralized training, decentralized execution (CTDE) \(^1\)
  - Decentralized learning + communication \(^2\)

\(^1\) Sunehag et al., Value Decomposition Networks for Cooperative Multiagent learning, AAMAS 2018.
\(^2\) Jaques et al., Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning, ICML 2019.
How can we foster team cooperation in the decentralized learning scenario w/o explicit communication?

Fully decentralized learning: no shared model components or communication between agents during training or execution

- Search-and-rescue robotics
- Autonomous driving
- Scalability
- Parallelism
DM$^2$: a MARL algorithm that enables cooperation in the decentralized setting w/o explicit communication
Contributions

- Propose DM\(^2\), a decentralized MARL algorithm based on independent **distribution matching to encourage coordination**
- Theoretical analysis shows
  - Conditions under which DM\(^2\) **converges**
  - **Expert policies are a Nash equilibrium** for mixed task and distribution matching reward
- **Empirical validation** in StarCraft II tasks
Background: Stochastic Games

- Stochastic game\textsuperscript{[1]} $\langle K, S, A, \rho_0, T, R, \gamma \rangle$
  - Number of agents $K$
  - State space $S$
  - Action space $A \equiv A^K$
  - Initial state distribution $\rho_0 : \Delta(S)$
  - Transition function $T : S \times A_0 \times \cdots \times A_{K-1} \mapsto \Delta(S)$
  - Reward function $R_i : S \times A_0 \times \cdots \times A_{K-1} \mapsto \mathbb{R}$
  - Discount factor $\gamma$
- Per-agent policy $\pi_i : S \mapsto \Delta(A_i)$

\textsuperscript{[1]} Littman, Markov Games as a Framework for Multi-agent Reinforcement Learning, ICML 1994.
Background: Distribution Matching

- Approach to imitation learning (IL) \([1, 2]\)
- The **per agent** state-action visitation distribution

\[
\rho_{\pi_i, \pi_i^-}(s, a_i) := (1 - \gamma)\pi_i(a_i | s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | \pi_i, \pi_i^-)
\]

...should match the **per expert** state-action visitation distribution \(\rho_{\pi_{E_i}, \pi_{E_i^-}}(s, a_i)\)

Background: Distribution Matching

- Approach to imitation learning (IL) \([1, 2]\)

Background: Distribution Matching

- Approach to imitation learning (IL) [1, 2]

1. Individual distribution matching leads to agent policies converging to compatible expert policies

2. Expert policies also constitute a Nash equilibrium under a mixed task and distribution matching reward
DM$^2$: Decentralized MARL via Distribution Matching

Algorithm 1: DM$^2$ (Decentralized MARL via distribution matching)

- **Input**: Number of agents $K$, expert demonstrations $D_0, \ldots, D_K$, environment $\text{env}$, number of epochs $N$, number of time-steps per epoch $M$, reward mixture coefficient $c$

1. for $k = 0, \ldots, K - 1$
   2. Initialize discriminator parameters $\phi_k$;
   3. Initialize policy parameters $\theta_k$;
   4. end

5. for $n = 0, 1, \ldots, N - 1$
6. Gather $m = 1, \ldots, M$ steps of data $(s^m, a^m, r^m_{\text{env}})$ from $\text{env}$;
7. for $k = 0, \ldots, K - 1$
8. Sample $M$ states from demonstration $D_k$;
9. Update discriminator $D^k_{\phi}$;
10. Get GAIL reward $r^m_{k, \text{GAIL}} = -\log D^k_{\phi}(s^m)$
   for $m = 1, \ldots, M$;
11. Set agent reward $r^m_{k, \text{mix}} = r^m_{\text{env}} + r^m_{k, \text{GAIL}} \cdot c$;
12. Update agent policy $\pi^k_{\theta}$ with data $(s^m, a^m, r^m_{k, \text{mix}})$ for $m = 1, \ldots, M$;
13. end
14. end

- **Output**: $K$ agent policies $\pi^k_{\theta}$
Experimental Setting

• StarCraft II Multi-Agent Challenge\[1\] tasks
  - 5m vs 6m (5v6)
  - 3s vs 4z (3sv4z)
• Baselines w/environment reward alone
  - IPPO (decentralized)
  - QMIX\[2\] (CTDE)
  - R-MAPPO\[3\] (CTDE)
• Distribution Matching Baseline: DM\(^2\) w/SIL \[4\]

Experimental Setting

• MARL algorithm: Independent PPO (IPPO)\textsuperscript{[1]}
• Demonstrations from K experts
  – State-only demonstrations sampled from saved IPPO and QMIX checkpoints
• Per-agent reward function:

\[ r_{i,mix} = r_{env} + r_{i,GAIL} \times c \]

\textsuperscript{[1]} Yu et al., The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games, ArXiv 2021.
1. Sample efficiency of $\text{DM}^2$ vs baselines
2. Coordination of expert demonstrations

Demonstrations could be **concurrently** sampled from **jointly trained** expert policies

<table>
<thead>
<tr>
<th></th>
<th>concurrent</th>
<th>nonconcurrent</th>
</tr>
</thead>
<tbody>
<tr>
<td>joint</td>
<td>DM²</td>
<td>ablation</td>
</tr>
<tr>
<td>not joint</td>
<td>ablation</td>
<td>ablation</td>
</tr>
</tbody>
</table>
2. Coordination of expert demonstrations

Legend:
- **joint**
  - DM²
  - ablation (1)
- **not joint**
  - ablation (2)
  - ablation (3)
DM²: Decentralized Multi-Agent Reinforcement Learning via Distribution Matching

\[ \pi E_1 \cdots \pi E_k \]

\[ \rho E_i \]

\[ \rho_i \]

\[ r_{i, gail} + c \cdot r_{env} \]

Caroline Wang
caroline.l.wang@utexas.edu

Ishan Durugkar
ishand@cs.utexas.edu

Elad Liebman
eliebman@sparkcognition.com

Peter Stone
pstone@cs.utexas.edu