Approximating Large Cooperative Multi-Agent Reinforcement Learning (MARL) Problems via Mean-Field Control (MFC)

Vaneet Aggarwal, Purdue University

Joint Work with:

Washim U. Mondal

Mridul Agarwal

Satish V. Ukkusuri

Learning with Trials and Feedback

Figure: Learning in everyday life. Images are taken from the internet.

Multi-Agent Learning

Figure: Multi-player games, traffic signal control, autonomous driving. Images are taken from the internet.

- Connected local environments.
- Individual rewards.
- Action of one agent can impact
  - all local states.
  - the rewards of all agents.

Mathematical Formulation

- N agents.
- Individual state space $\mathcal{S} = \{1, 2, \cdots, S\}$.
- Individual action space $\mathcal{A} = \{1, 2, \cdots, A\}$.
- State and action of ith agent at time t: s_t^i, and a_t^i.
- Joint state and action at time t: $\mathbf{s}_t = \{s_t^i\}_{i \in \{1, \cdots, N\}}$, and $\mathbf{a}_t$.
- Reward of ith agent at time t: $r_i(\mathbf{s}_t, \mathbf{a}_t)$.
- State transition of ith agent: $s_{t+1}^i \sim P_i(\mathbf{s}_t, \mathbf{a}_t)$.

Mathematical Formulation

- Policy of ith agent: $a_t^i \sim \pi_t^i(\mathbf{s}_t)$
- Joint policy-sequence: $\boldsymbol{\pi} = \{\pi_t^i\}_{i \in \{1,\cdots,N\}, t \in \{0,1,\cdots\}}$
- In cooperative setup, the following is maximized:

$$v_N(\mathbf{s}_0, \boldsymbol{\pi}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(\mathbf{s}_t, \mathbf{a}_t)\right] \quad (1)$$

over all policy-sequence $\boldsymbol{\pi}$.

- Expectation is over all trajectory generated by $\boldsymbol{\pi}$ from $\mathbf{s}_0$.
- Joint state-space: $\mathcal{S}^N$. The goal is difficult in general.

Existing Approaches

Localisation of Policy:

- Each policy is dependent on local states i.e., $\pi_t^i(\mathbf{s}_t) = \pi_t^i(s_t^i)$

Training:

- Independent Q-Learning (IQL).
- Centralised training with decentralised execution (CTDE)
  - VDN [7], QMIX [5], WQMIX [4], QTRAN [6] etc.

Merit and Demerit:

- Works well empirically for moderately high number of agents.
- No optimality guarantee.

Mean-Field Control (MFC)

Basic Premise:

- One can accurately infer group behaviour by studying only a representative agent if the agents are
  - (A1) identical and exchangeable, and
  - (A2) infinite in number
- Consequence of (A1) in an N-agent system:
  - $r_i(\mathbf{s}_t, \mathbf{a}_t) = r(s_t^i, a_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)$
  - $P_i(\mathbf{s}_t, \mathbf{a}_t) = P(s_t^i, a_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)$
  - $\pi_t^i(\mathbf{s}_t) = \pi_t(s_t^i, \boldsymbol{\mu}_t^N)$ where

$$\boldsymbol{\mu}_t^N(s) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta(s_t^i = s), \quad \boldsymbol{\nu}_t^N(a) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta(a_t^i = a) \qquad (2)$$

Behaviour of an Infinite Agent System

- State and action of representative at time t: s_t, and a_t.
- Policy-sequence of representative: $\boldsymbol{\pi} = \{\pi_t\}_{t\in\{0,1,\cdots\}}$.
- State and action distributions at time t: $\boldsymbol{\mu}_t^\infty$, $\boldsymbol{\nu}_t^\infty$.
- Action Distribution Evolution:

$$\boldsymbol{\nu}_t^\infty \triangleq \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^\infty, \pi_t) = \sum_{s\in\mathcal{S}} \pi_t(s, \boldsymbol{\mu}_t^\infty)\boldsymbol{\mu}_t^\infty(s) \tag{3}$$

- State Distribution Evolution:

$$\begin{aligned}\boldsymbol{\mu}_{t+1}^\infty &\triangleq P^{\mathrm{MF}}(\boldsymbol{\mu}_t^\infty, \pi_t) \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} P(s, a, \boldsymbol{\mu}_t^\infty, \boldsymbol{\nu}_t^\infty)\pi_t(s, \boldsymbol{\mu}_t^\infty)(a)\boldsymbol{\mu}_t^\infty(s)\end{aligned} \tag{4}$$

Goal in MFC

- Expected reward of the representative at time t:

$$r^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\infty}, \pi_t) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a, \boldsymbol{\mu}_t^{\infty}, \boldsymbol{\nu}_t^{\infty}) \pi_t(s, \boldsymbol{\mu}_t^{\infty})(a) \boldsymbol{\mu}_t^{\infty}(s) \tag{5}$$

- Maximize over all $\boldsymbol{\pi}$ the following for initial distribution, $\boldsymbol{\mu}_0$.

$$v_{\infty}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) = \sum_{t=0}^{\infty} \gamma^t r^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\infty}, \pi_t) \tag{6}$$

Research Gap

- It is known [1] that for large N, and for all $\boldsymbol{\pi}$,

$$|v_N(\mathbf{s}_0, \boldsymbol{\pi}) - v_\infty(\boldsymbol{\mu}_0, \boldsymbol{\pi})| = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{7}$$

- How the error changes when
  - agents are heterogeneous? (JMLR 2022 [2])
  - non-exchangeable? (UAI 2022 [3])
  - additional constraints are present? (Submitted to NeurIPS)
- How to solve MFC sample-efficiently?
- Construction of local policy? (Submitted to TMLR)

Approximating Heterogeneous MARL

- K classes of agents $\{\mathcal{N}_1, \cdots, \mathcal{N}_K\}$
- Populations $N_1, \cdots, N_K$.
- $N_1 + \cdots + N_K = N$ and $\mathbf{N} \triangleq \{N_1, \cdots, N_K\}$.
- Agents within each class are identical and exchangeable.

Reward and state-transition depend on:

- Case 1: Joint state and action distributions over all classes.
- Case 2: State and action distributions of individual classes.
- Case 3: Marginalized state and action distributions.

Approximating Heterogeneous MARL: Case 1

For an agent i belonging to k-th class,

- $r_i(\mathbf{s}_t, \mathbf{a}_t) = r_k(s_t^i, a_t^i, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}})$
- $P_i(\mathbf{s}_t, \mathbf{a}_t) = P_k(s_t^i, a_t^i, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}})$

where $\boldsymbol{\mu}_t^{\mathbf{N}} = \{\boldsymbol{\mu}_t^{k,N_k}\}_{k\in\{1,\cdots,K\}}$, $\boldsymbol{\nu}_t^{\mathbf{N}} = \{\boldsymbol{\nu}_t^{k,N_k}\}_{k\in\{1,\cdots,K\}}$ and

$$\boldsymbol{\mu}_t^{k,N_k}(s) = \frac{1}{N}\sum_{i\in\mathcal{N}_k}\delta(s_t^i = s), \tag{8}$$

$$\boldsymbol{\nu}_t^{k,N_k}(a) = \frac{1}{N}\sum_{i\in\mathcal{N}_k}\delta(a_t^i = a) \tag{9}$$

Example: Ride sharing market where classes may be vehicle type, driver type etc.

Approximating Heterogeneous MARL: Results

The error between MARL and MFC is $\mathcal{O}(e)$ where

- $e = \left[\frac{1}{N}\sum_k \sqrt{N_k}\right][\sqrt{S}+\sqrt{A}]$ (Case 1)
- $e = \left[\sum_k \frac{1}{\sqrt{N_k}}\right][\sqrt{S}+\sqrt{A}]$ (Case 2)
- $e = \left[\frac{A}{N}\sum_k \sqrt{N_k} + \sum_k \frac{B}{\sqrt{N_k}}\right][\sqrt{S}+\sqrt{A}]$ for some constants A, B (Case 3)

We also develop an algorithm that approximately solves MFC and therefore also solves MARL with $\mathcal{O}(e)$ error and $\mathcal{O}(e^{-3})$ sample complexity.

Crux of the Proof for Case 1

Assumptions

- $|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)| \leq M$
- $|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) - r(x,u,\boldsymbol{\mu}_2,\boldsymbol{\nu}_2)| \leq L_R[|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1]$
- $|P(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) - P(x,u,\boldsymbol{\mu}_2,\boldsymbol{\nu}_2)|_1 \leq L_P[|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1]$
- $|\pi(x,\boldsymbol{\mu}_1) - \pi(x,\boldsymbol{\mu}_2)| \leq L_Q|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|$
- $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2$ are arbitrary joint distributions
- Bounded reward
- Lipschitz reward, transition, policy

Crux of the Proof for Case 1

Consequence of Assumption

- Lipschitz continuity extends to mean field system

- $|\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1, \pi) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_2, \pi)|_1 \leq (1 + L_Q)|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$ (Lemma 1)
- $|P^{\mathrm{MF}}(\boldsymbol{\mu}_1, \pi) - P^{\mathrm{MF}}(\boldsymbol{\mu}_2, \pi)|_1 \leq S_P|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$ (Lemma 2)
- $|r^{\mathrm{MF}}(\boldsymbol{\mu}_1, \pi) - r^{\mathrm{MF}}(\boldsymbol{\mu}_2, \pi)|_1 \leq S_R|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$ (Lemma 3)

Crux of the Proof for Case 1

Where does $\sqrt{N}$ factor come from?

Lemma 4

If $\{X_{m,n}\}_{m\in[M],n\in[N]}$ are random variables and $\{C_{m,n}\}_{m\in[M],n\in[N]}$ are constants such that

- If $\forall m \in [M]$, $\{X_{m,n}\}_{n\in[N]}$ are independent
- $0 \le X_{m,n} \le 1, \ \forall m, n$
- $\sum_{m\in[M]} \mathbb{E}[X_{m,n}] = 1, \ \forall n \in [N]$
- $|C_{m,n}| \le C, \ \forall m \in [M], \forall n \in [N]$, then

$$\sum_{m=1}^{M} \mathbb{E}\left|\sum_{n=1}^{N} C_{m,n}\Big(X_{m,n} - \mathbb{E}[X_{m,n}]\Big)\right| \le C\sqrt{MN} \qquad (10)$$

Consequence of Lemma 4

Lemma 5:

$$\mathbb{E}|\boldsymbol{\nu}_t^{\mathbf{N}} - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)|_1 \leq \frac{1}{N}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\sqrt{|\mathcal{U}|}$$

Lemma 6:

$$\mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}} - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)\right|_1 \leq C_P\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]\frac{1}{N}\left(\sum_{k\in[K]}\sqrt{N_k}\right)$$

Consequence of Lemma 4

Lemma 7:

$$\mathbb{E}\left|\frac{1}{N_{\text{pop}}}\sum_{k\in[K]}\sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}) - \sum_{k\in[K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)\right|$$

$$\leq C_R\sqrt{|\mathcal{U}|}\frac{1}{N}\left(\sum_{k\in[K]}\sqrt{N_k}\right)$$

What do these differences (Lemma 5, 6, 7) mean?

- Characterizing a one-step difference between MARL and MFC
- $\boldsymbol{\mu}_t^N \to \boldsymbol{\mu}_{t+1}^N$ (MARL update)
- $\boldsymbol{\mu}_t^N \to P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)$ (MFC update)

Multi-Step Difference

Via Recursion, $\mathbb{E}\left|\boldsymbol{\mu}^{\mathbf{N}}_{t+1} - \boldsymbol{\mu}_{t+1}\right|_1$ can be bounded.

- Our goal: the difference between MARL and MFC rewards
- It translates to γ-discounted sum of $\mathbb{E}\left|\boldsymbol{\mu}^{\mathbf{N}}_{t+1} - \boldsymbol{\mu}_{t+1}\right|_1$

Approximating MARL with Non-Uniform Interaction

Motivational Example: Traffic Signal Control.

- Nearby intersections interact stronger than far-away ones.

Model of Non-Uniform Interaction:

- N agents with identical reward and state transition functions.
- Interaction between agent i, j: $W(i,j)$.
- State and action distribution as seen by ith agent:

$$\boldsymbol{\mu}_t^{i,N}(s) = \sum_{j=1}^{N} W(i,j)\delta(s_t^j = s), \tag{11}$$

$$\boldsymbol{\nu}_t^{i,N}(a) = \sum_{j=1}^{N} W(i,j)\delta(a_t^j = a) \tag{12}$$

Approximating MARL with Non-Uniform Interaction

- Reward of ith agent: $r(s_t^i, a_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})$
- State transition of ith agent: $s_{t+1}^i \sim P(s_t^i, a_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})$

Main Result:

- MFC can still approximate MARL if
  - W is doubly-stochastic matrix (DSM)
  - reward functions are affine
- The approximation error is $\mathcal{O}(e)$ where $e = \frac{1}{\sqrt{N}}\left[\sqrt{S} + \sqrt{A}\right]$.
- Developed algorithm to obtain optimal policy with
  - $\mathcal{O}(\max\{e, \epsilon\})$ error, and
  - $\mathcal{O}(\epsilon^{-3})$ sample complexity for any $\epsilon > 0$.

Numerical Results

Consider a network of N firms operated by a single operator. All of the firms produce the same product but with varying quality (with Q levels).

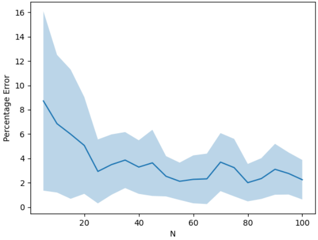
At each time, each firm decides whether to invest to improve the quality of its product. The quality improves as

$$x_{t+1}^i = \begin{cases} x_t^i + \left\lfloor \chi \left(1 - \frac{\bar{\boldsymbol{\mu}}_t^{i,N}}{Q}\right)(Q - x_t^i) \right\rfloor & \text{if } u_t^i = 1, \\ x_t^i & \text{otherwise} \end{cases}$$

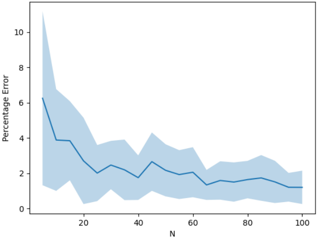
where χ is a uniform random variable between $[0,1]$, and $\bar{\boldsymbol{\mu}}_t^{i,N}$ is average product quality of its $K < N$ neighbouring firms. The total reward can be expressed as follows.

$$r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}) = \alpha_R x_t^i - \beta_R (\bar{\boldsymbol{\mu}}_t^{i,N})^\sigma - \lambda_R u_t^i$$

Numerical Results



(a) Affine Reward



(b) Nonlinear Reward

Figure: Percentage error between MARL and MFC as a function of N.

Approximating Constrained MARL

Premise:

- In addition to reward, each agent incurs cost $c(s_t^i, a_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)$
- Consider the reward and cost values:

$$V_N^r(\mathbf{s}_0, \boldsymbol{\pi}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t^i, a_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)\right], \qquad (13)$$

$$V_N^c(\mathbf{s}_0, \boldsymbol{\pi}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t^i, a_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)\right] \qquad (14)$$

Approximating Constrained MARL

$$
\begin{aligned}
&\max_{\boldsymbol{\pi}} \; V_N^r(\mathbf{s}_0, \boldsymbol{\pi}) \\
&\text{subject to: } V_N^c(\mathbf{s}_0, \boldsymbol{\pi}) \leq 0
\end{aligned}
\tag{15}
$$

Main Result:

- MFC approximation error $\mathcal{O}(e)$ where $e = \frac{1}{\sqrt{N}}[\sqrt{S} + \sqrt{A}]$.
- Zero constraint violation for large N.
- Devised Primal-Dual algorithm that computes the optimal policy with
  - $\mathcal{O}(e)$ error,
  - Zero constraint violation for large N
  - $\mathcal{O}(e^{-6})$ sample complexity.

Numerical Result

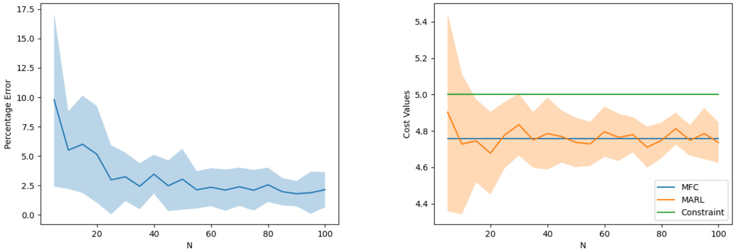


Figure: Percentage error in approximating the optimal objective value and constraint violation respectively as functions of N.

Constructing Near-Optimal Local Policy

Idea:

- Collecting network-wide information to compute $\boldsymbol{\mu}_t^N$, $\boldsymbol{\nu}_t^N$ is costly or impossible at each instant.
- $\boldsymbol{\mu}_t^\infty$, $\boldsymbol{\nu}_t^\infty$ can be obtained deterministically via mean-field updates if $\boldsymbol{\mu}_0$ is known.
- Can we use $\boldsymbol{\mu}_t^\infty$, $\boldsymbol{\nu}_t^\infty$ as proxy for $\boldsymbol{\mu}_t^N$, $\boldsymbol{\nu}_t^N$?
- It eliminates the cost of communication except at $t = 0$.

Constructing Near-Optimal Local Policy

- Let, $\boldsymbol{\pi}_N^*$ be the optimal N-agent policy sequence.
- $\boldsymbol{\pi}_\infty^* = \{\pi_{t,\infty}^*\}$ be optimal infinite agent policy-sequence.
- Define $\tilde{\boldsymbol{\pi}}_\infty^* = \{\tilde{\pi}_{t,\infty}^*\}$ such that,

$$\tilde{\pi}_{t,\infty}^*(s, \boldsymbol{\mu}) = \pi_{t,\infty}^*(s, \boldsymbol{\mu}_t^\infty), \quad \forall s, \forall \boldsymbol{\mu} \tag{16}$$

- We show that,

$$|v_N(\mathbf{s}_0, \boldsymbol{\pi}_N^*) - v_N(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\pi}}_\infty^*)| = \mathcal{O}(e), \quad e = \frac{1}{\sqrt{N}}[\sqrt{S} + \sqrt{A}]$$

- We develop an algorithm that computes $\tilde{\boldsymbol{\pi}}_\infty^*$ with $\mathcal{O}(\max\{e, \epsilon\})$ error and $\mathcal{O}(\epsilon^{-3})$ sample complexity.

Numerical Result

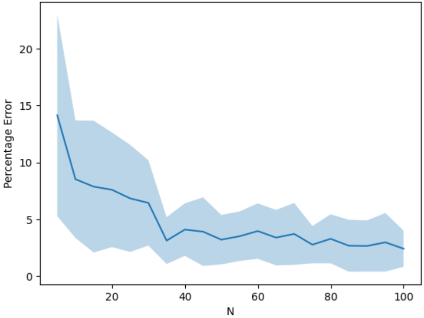


Figure: Percentage error of approximating the optimal policy via a local policy as a function of N.

References I

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu.
Mean-field controls with q-learning for cooperative marl: convergence and complexity analysis.
SIAM Journal on Mathematics of Data Science, 3(4):1168–1196, 2021.

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V. Ukkusuri.
On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC).
Journal of Machine Learning Research, 23(129):1–46, 2022.

Washim Uddin Mondal, Vaneet Aggarwal, and Satish V Ukkusuri.
Can mean field control (mfc) approximate cooperative multi agent reinforcement learning (marl) with non-uniform interaction?
Uncertainity in Artificial Intelligence (UAI), 2022.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson.
Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning.
Advances in neural information processing systems, 33:10199–10210, 2020.

References II

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson.
Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning.
In *International conference on machine learning*, pages 4295–4304. PMLR, 2018.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi.
Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning.
In *International conference on machine learning*, pages 5887–5896. PMLR, 2019.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al.
Value-decomposition networks for cooperative multi-agent learning based on team reward.
In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018.