Detecting Influence Structures in Multi-Agent Reinforcement Learning Systems

Fabian Raoul Pieroth, 1 **Katherine Fitch,** 2 **Lenz Belzner,** 3

[1]Decision Sciences & Systems Department
Technical University Munich fabian.pieroth@tum.de
[2]LARALAB GmbH
Munich, Germany
[3]Technische Hochschule Ingolstadt
Ingolstadt, Germany

Abstract

We consider the problem of quantifying the amount of influence one agent can exert on another in the setting of multi-agent reinforcement learning (MARL). As a step towards a unified approach to express agents' interdependencies, we introduce the total and state influence measurement functions. Both of these are valid for all common MARL systems, such as the discounted reward setting. Additionally, we propose novel quantities, called the total impact measurement (TIM) and state impact measurement (SIM), that characterize one agent's influence on another by the maximum impact it can have on the other agents' expected returns and represent instances of impact measurement functions in the average reward setting. Furthermore, we provide approximation algorithms for TIM and SIM with simultaneously learning approximations of agents' expected returns, error bounds, stability analyses under changes of the policies, and convergence guarantees. The approximation algorithm relies only on observing other agents' actions and is, other than that, fully decentralized. Our work appears to be the first study of determining influence structures in the multi-agent average reward setting with provable convergence guarantees.

1 Introduction

The knowledge of mutual influence among a general system consisting of several entities, subsequently called agents, is beneficial to learn good strategies. The present work is regarding the influence among agents in the area of multi-agent reinforcement learning (MARL). Here, a shared environment is affected by the joint action of multiple agents. For each state of the environment, each agent chooses an action from its action space. The resulting joint action determines the transition to the following state. Each agent receives a reward for each transition, which is allowed to be different for every agent. Here, we consider the problem of giving a unified representation and an interpretable and measurable quantification of influence among agents.

Existing work addresses specific use cases and objectives of influence structures in MARL systems, such as reducing the number of agents that need to collaborate (Guestrin, Lagoudakis, and Parr 2002), guiding exploration to states with high influence (Wang et al. 2020), or determining which agents need to communicate (Jaques et al. 2018). They focus on analyzing their method's effect on the system's objective without explicitly addressing the influence measurement's common theoretical aspects. Furthermore, the mentioned methods to measure influence among agents are exclusively focusing on the discounted reward setting (Sutton and Barto 2018). As such, there is a lack of research related to influence in the average reward setting (Puterman 1994), which is typically used for ongoing applications, e.g., load management in energy networks (Callaway and Hiskens 2011), formation control of vehicles (Fax and Murray 2004), or repeated auctions (Hoen et al. 2005).

The existing approaches mentioned above seek to resolve specific problems, such as a reduction of the joint action space by using a proxy of agents' influence on one another. While our method can be used for these applications as well, the main goal of our work is to address the fundamental question of how to reliably detect the inherent influence structure of an environment given a specific policy.

The main contributions of our work are the following. We introduce a unified approach to express a multi-agent system's inherent influence structure, regardless of the reward setting and overall objective. We then build upon this foundation by introducing the *total impact measurement* and *state impact measurement*. These measurements quantify the overall and state-dependent influence structure, respectively, in the multi-agent average reward setting. In addition, we provide decentralized algorithms with stability analysis and convergence guarantees along with complementary empirical evaluations.

To the best of our knowledge, our work is the first study of determining influence structures in the multi-agent average reward setting with provable convergence guarantees.

2 Related Work

One popular representation of agents' dependencies is a coordination graph (Guestrin, Lagoudakis, and Parr 2002), which is used to determine which agents' actions are relevant for the individual state-action functions. Several works try to detect the influence that the agents can exert on one another, e.g., (Kok et al. 2005). In contrast to our method, they require storing all estimations of the state-action values for the whole time horizon. Furthermore, they do not provide any theoretical analysis of their approximation method's quality. Another approach estimates the maximum expected utility one agent can potentially receive when coordinating with subgroups (Zhang and Lesser 2013). Unlike our method, they rely on an approximation of the state transition probabilities of the underlying Markov decision process and only provide empirical evaluations for their method.

Wang et al. (2020) introduce the *Value of Interaction* to guide exploration to relevant states. Their formulation is similar to our proposed formulation of dependencies among agents. However, they rely on empirical estimation of the state transition probabilities, which is not the case for our work. Furthermore, their formulation is restricted to a specific state, whereas TIM, as proposed in this work, is formulated for the overall influence of one agent on another.

Recently, researchers use the variance in state-action functions to construct context-aware coordination graphs (Wang et al. 2021). Contrary to our work, they do not provide any error bounds of their approximation quality and their formulation is again restricted to specific states only.

Instead of examining the influence between agents via their ability to alter the expected long-term return, Jaques et al. (2018) define causal influence by the changes of one agent's actions in the policy of another. However, their approach either demands that the probability of another agent's action, given a counterfactual action, is known or estimated. Our approach does not rely on this information, as we only require observing the other agents' actions.

3 Background

This section introduces the multi-agent Markov decision process (MDP) in the infinite horizon average reward setting. It is the natural extension of the single-agent case introduced by Puterman (1994). In the section's second part, we present some results from stochastic approximation (Borkar 2008), which we need for the proofs of our main results.

Multi-Agent MDP

We consider a system of N agents operating in a shared environment with discrete time steps $t \in \mathbb{N}$. The set of agents is denoted by $\mathcal{N}$. The environment can be described by a multi-agent MDP, specified in the following definition.

Definition 3.1 (Multi-agent Markov decision process)**.** A multi-agent MDP is defined by a tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{r^i\}_{i\in\mathcal{N}})$, where $\mathcal{N} = \{1, \dots, N\}$ denotes the set of agents, $\mathcal{S}$ is a finite state space which is shared by all agents, $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}^i$ is the joint action space, where $\mathcal{A}^i$ denotes the set of actions of agent i. Additionally, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the MDP's state transition probability. There exist functions $R^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $R^i(s, a) = \mathbb{E}[r^i_{t+1}|s_t = s, a_t = a]$, which are denoted as the individual reward functions. Furthermore, the states and the joint actions are observable by every agent.

For every time step, each agent chooses its action according to its policy π^i, which is a probability distribution over $\mathcal{A}^i$. Therefore, we assume that the individual policies are conditionally independent given the state, i.e., the joint policy is given by $\pi(s, a) = \prod_{i\in\mathcal{N}} \pi^i(s, a^i)$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For a subset of agents $B^j = \{b^j_1, \dots, b^j_k\} \subset \mathcal{N}$ we denote $a^{Bj} = (a^{b^j_1}, \dots, a^{b^j_k})$, and $-B^j = \mathcal{N} \setminus B^j$. We are concerned with the average reward setting. The individual expected time-average reward of agent $i \in \mathcal{N}$ is

$$J^i(\pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[r^i_{t+1}\right]. \tag{1}$$

To quantify the effects of a specific state and joint action, we define the relative individual state-action function for agent $i \in \mathcal{N}$, state $s \in \mathcal{S}$, and joint action $a \in \mathcal{A}$ as

$$Q^i_\pi(s, a) := \sum_{t\geq 0} \mathbb{E}\left[r^i_{t+1} - J^i(\pi)|s_0 = s, a_0 = a\right]. \tag{2}$$

Consider states $s, s' \in \mathcal{S}$. The probability of transitioning from state s to s' given a joint policy π can be denoted by

$$P_\pi(s'|s) = \sum_{a\in\mathcal{A}} \pi(s, a) \cdot P\left(s'|s, a\right). \tag{3}$$

This induces a Markov chain over the states $\{s_t\}_{t\geq 0}$ with transition matrix $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$. We make the following assumption on this Markov chain and the joint policy.

Assumption 3.2. The policies satisfy $\pi_i(s, a^i) > 0$ for every $i \in \mathcal{N}, s \in \mathcal{S}$ and $a^i \in \mathcal{A}^i$. Moreover, for every joint policy π the induced Markov chain over the states $\{s_t\}_{t\geq 0}$ is ergodic, i.e., it is irreducible and aperiodic.

By Theorem 4.1 on page 119 in the book of Seneta (2006), there exists a unique stationary distribution for any ergodic Markov chain. We denote the stationary distribution of the Markov chain over the states by d_π. Given some states $s, s' \in \mathcal{S}$ and joint actions $a, a' \in \mathcal{A}$, the probability to transition from (s, a) to (s', a') can be expressed by

$$P^{\mathcal{A}}_\pi(s', a'|s, a) = P(s'|s, a) \cdot \pi(s', a'). \tag{4}$$

This induces a Markov chain over the states and actions $\{(s_t, a_t)\}_{t\geq 0}$ with transition matrix $P^{\mathcal{A}}_\pi \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times|\mathcal{S}|\cdot|\mathcal{A}|}$. Note that this Markov chain is ergodic (Zhang et al. 2018) and its stationary distribution is given by $d^{\mathcal{A}}_\pi(s, a) = d_\pi(s) \cdot \pi(s, a)$, for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. One property that we use throughout this paper is a simplified representation of averages over an ergodic Markov chain (Zhang et al. 2018). For example, one can represent the individual long-term return defined in Equation (1) by

$$J^j(\pi) = \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \pi(s, a) \cdot R^j(s, a). \tag{5}$$

Stochastic Iteration Approximation

Our main results use the following statements. We state here a special case of Corollary 8 and Theorem 9 on pages 74-75 of Borkar (2008). These special cases have been formulated by Zhang et al. (2018).

Consider an n-dimensional stochastic approximation iteration for $\gamma_t > 0$

$$x_{t+1} = x_t + \gamma_t \left[h\left(x_t, Y_t\right) + M_{t+1} + \beta_{t+1}\right], t \geq 0 \tag{6}$$

where $\{Y_t\}_{t\geq 0}$ is a Markov chain on a finite set A.

Assumption 3.3. We make the following assumptions:

(a) $h : \mathbb{R}^n \times A \to \mathbb{R}^n$ is Lipschitz in its first argument;

(b) $\{Y_t\}_{t\geq 0}$ is an irreducible Markov chain with stationary distribution π;

(c) The stepsize sequence $\{\gamma_t\}_{t\geq 0}$ satisfies $\sum_{t\geq 0} \gamma_t = \infty$ and $\sum_{t\geq 0} \gamma_t^2 < \infty$;

(d) $\{M_t\}_{t\geq 0}$ is a martingale difference sequence, satisfying for some $K > 0$ and $t \geq 0$

$$\mathbb{E}\left(\|M_{t+1}\|^2 \mid x_\tau, M_\tau, Y_\tau; \tau \leq t\right) \leq K \cdot \left(1 + \|x_t\|^2\right);$$

(e) The sequence $\{\beta_t\}_{t\geq 0}$ is a bounded random sequence with $\beta_t \to 0$ almost surely as $t \to \infty$.

If Assumption 3.3 holds, then the asymptotic behavior of the iteration in Equation (6) is related to the behavior of the solution to the ordinary differential equation (ODE)

$$\dot{x} = \bar{h}(x) = \sum_{i \in A} \pi(i) h(x, i) \tag{7}$$

Furthermore, suppose the ODE in Equation (7) has a unique globally asymptotically stable equilibrium x^*, then we have the following two theorems connecting this solution to the original iteration algorithm (6).

Theorem 3.4 ((Zhang et al. 2018))**.** Under Assumption 3.3, if $\sup_{t\geq 0} \|x_t\| < \infty$ a.s., we have $x_t \to x^*$

Theorem 3.5 ((Zhang et al. 2018))**.** Under Assumption 3.3, suppose that $\lim_{c\to\infty} \frac{\bar{h}(cx)}{c} = h_\infty(x)$ exists uniformly on compact sets for some $h_\infty \in C(\mathbb{R}^n)$. If the ODE $\dot{y} = h_\infty(y)$ has the origin as the unique globally asymptotically stable equilibrium, then $\sup_{t\geq 0} ||x_t|| < \infty$ almost surely.

4 Influence Representations

The present work aims to specify and detect influence structures among agents in a multi-agent system. For this purpose, we first specify dependent and independent agents, following the definition of Guestrin, Venkataraman, and Koller (2002). Afterward, we introduce a novel representation framework of agents' influence structures, which is valid for all common reward settings and MDP formulations.

Dependencies and Independencies in Multi-Agent Systems

Given a state $s \in \mathcal{S}$, one agent's actions are relevant for another, if these actions either directly influence the reward of the other agent, or affect the state for the other agent and, therefore, influence the reward indirectly. Both effects are captured in the individual state-action functions. Let $B^j \subset \mathcal{N}$ be a subset of agents and $j \in \mathcal{N}$, then agent j is exclusively dependent on the agents in B^j in $s \in \mathcal{S}$ if

$$Q^j_\pi(s, a^{B^j}, a^{-B^j}) = Q^j_\pi(s, a^{B^j}) \text{ for all } a \in \mathcal{A}. \tag{8}$$

If this holds for all $s \in \mathcal{S}$, then agent j acts completely independent in the MDP from agents in B^{-j}.

Influence Measurement Functions

A binary representation of the dependency group B^j is given by so-called coordination graphs (Guestrin, Venkataraman, and Koller 2002). However, strict independence as defined above often does not hold, which leads to large B^j's or even $B^j = \mathcal{N}$. Several approaches demonstrated that one can approximate the individual state-action functions quite well by assuming some agents to be independent of others (Sunehag et al. 2018; Böhmer, Kurin, and Whiteson 2019; Zhang and Lesser 2013). That means $Q^j_\pi(s,a) \approx Q^j_\pi(s, a^{\overline{B}^j})$ for $\overline{B}^j \subsetneq B^j$. That indicates not every agent in B^j has equal influence on agent j's individual state-action function. Therefore, one needs a representation that allows a more fine-grained distinction of influence to express these differences.

There is no single quantity to express influence in a multi-agent system, as it depends on the specific use case. However, the study of different kinds of influence structures offers great value as a descriptive inherent property. Therefore, we propose a general framework to express influence structures in the form of abstract functions that are only bound by the independence criterion from Equation (8). We introduce an expression of state-dependent influence structures with the so-called state influence measurement function.

Definition 4.1 (State influence measurement function)**.** Let Ω be an arbitrary set, $\mathcal{N}$ a set of N agents with joint policy $\pi = \prod_{j\in\mathcal{N}} \pi^j$, and individual state-action functions $Q^1_\pi, \ldots, Q^N_\pi$. Furthermore, let $\Psi^{\mathcal{S}} : \mathcal{S} \times \Omega \to [0,\infty)^{N\times N}$ be a matrix-valued function such that for any $s \in \mathcal{S}$ and $\omega \in \Omega$ an entry $\Psi^{\mathcal{S}}_{i,j}(s,\omega) > 0$ if and only if there exist actions $a^{-i} \in \mathcal{A}^{-i}, a^i, \bar{a}^i \in \mathcal{A}^i$ such that $Q^j(s, a^{-i}, a^i) \neq Q^j(s, a^{-i}, \bar{a}^i)$. Then, the function $\Psi^{\mathcal{S}}$ is called a state influence measurement function of the system of agents $\mathcal{N}$.

A total influence measurement function is suitable to express global influence structures.

Definition 4.2 (Total influence measurement function)**.** Let Ω be an arbitrary set, $\mathcal{N}$ a set of N agents with joint policy $\pi = \prod_{j\in\mathcal{N}} \pi^j$, and individual state-action functions $Q^1_\pi, \ldots, Q^N_\pi$. Furthermore, let $\Psi : \Omega \to [0,\infty)^{N\times N}$ be a matrix-valued function such that for any $\omega \in \Omega$ an entry $\Psi_{i,j}(\omega) > 0$ if and only if there exist a state $s \in \mathcal{S}$ and actions $a^{-i} \in \mathcal{A}^{-i}, a^i, \bar{a}^i \in \mathcal{A}^i$ such that $Q^j(s, a^{-i}, a^i) \neq Q^j(s, a^{-i}, \bar{a}^i)$. Then, the function Ψ is called a total influence measurement function of the system of agents $\mathcal{N}$.

Note that the definitions of state and total influence measurement functions are valid for any setting with a well-defined state-action function. Therefore, it holds for the average reward setting, which we focus on in our later analyses, but also for the discounted reward setting (Sutton and Barto 2018). Furthermore, it holds for setups with infinite state and action spaces. The set Ω offers a parametrization of an influence measurement, for example, in the form of a prior that holds expert knowledge about the environment.

The value of an influence measurement function's knowledge is contingent on its semantic meaning. Nonetheless, there are specific interpretations that are valid for any influence measurement function. For a total influence measurement function, one can assume that for every agent j there exists at least one agent $i \in \mathcal{N}$ such that the individual state-action function Q^j_π is dependent on the actions of agent i. Otherwise, no action in any state in the system could influence the reward of agent j in any way. Note that $i = j$ is allowed here. That means that the matrix $\Psi(\omega)$ has a positive entry in any row and column. Therefore, one can always get either a row- or column-stochastic matrix $\overline{\Psi}(\omega)$ from $\Psi(\omega)$ by respectively normalizing the rows or columns.

For a column stochastic $\overline{\Psi}(\omega)$, one can interpret column j as a probability distribution of the influence the agents in $\mathcal{N}$ can have on agent j's state-action function. In this case, one can deduce a ranking depending on Ψ, e.g., to determine which agents should be in the coordination group B^j.

The entries in row i in a row-stochastic matrix $\overline{\Psi}(\omega)$ can, on the other hand, be interpreted as a probability distribution of agent i's influence on the system of agents according to Ψ. This can be used, for example, in a cooperative setting, where the objective is to maximize the long-term return of the whole system. An entry $\Psi_{i,j}(\omega)$ describes the influence agent i has on agent j according to Ψ. If this entry is large compared to the other ones in the row, then agent i should pay attention to its effects on agent j's expected reward when taking its actions.

The same deductions are valid for a state influence measurement function $\Psi^{\mathcal{S}}$, although the assumption of a positive entry in every row and column does not necessarily hold.

5 Influence Measurement Functions in the Average Reward Setting

In this section, we propose novel quantities to measure influence among agents, as the maximum impact an agent can have on the individual state-action function of another. We show that the proposed quantities are instances of a state and total influence measurement function respectively, and give approximation algorithms with convergence proofs.

The Total Impact Measurement

The core of the proposed measurements consists of the so-called impact sample, which quantifies the maximum impact one agent can have on the expected return of another given a specific state and joint action.

Definition 5.1 (Impact sample)**.** Let $\pi = \prod_{i \in \mathcal{N}} \pi^i$ be a joint policy of a set of agents $\mathcal{N}$, which are acting in a multi-agent MDP. For a state $s \in \mathcal{S}$ and joint action $a \in \mathcal{A}$, we define the impact sample of agent i on agent j as

$$U^{i \to j}_\pi(s,a) := \max_{a^i \in \mathcal{A}^i} Q^j_\pi(s, a^{-i}, a^i) - \min_{a^i \in \mathcal{A}^i} Q^j_\pi(s, a^{-i}, a^i). \tag{9}$$

Here Q^j_π denotes the individual state-action function for agent j and $a^{-j} = (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_N)$ the joint action except the action of agent j.

The impact sample for agent j on agent i, given a specific $s \in \mathcal{S}$ and a joint action $a \in \mathcal{A}$, indicates how much agent j can influence the expected long-term return of agent i. Averaging this over all possible states and joint actions yields the total impact measurement.

Definition 5.2 (Total impact measurement)**.** Let $\pi = \prod_{i \in \mathcal{N}} \pi^i$ be a joint policy and $\{(s_t, a_t)\}_{t \geq 0}$ the induced Markov chain over the states and actions in a multi-agent MDP. The total impact measurement (TIM) of agent i on agent j, for $i, j \in \mathcal{N}$, is defined as

$$TI^{i \to j}(\pi) := \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[U^{i \to j}_\pi(s_t, a_t)\right]. \tag{10}$$

Note that under Assumption 3.2, there exists a stationary distribution over the states and actions $d^{\mathcal{A}}_\pi = d_\pi \cdot \pi$, where d_π is the stationary distribution over the states. Then one can represent TIM by

$$TI^{i \to j}(\pi) = \sum_{s \in \mathcal{S}} d_\pi(s) \sum_{a \in \mathcal{A}} \pi(s,a) \cdot U^{i \to j}_\pi(s,a). \tag{11}$$

As the stationary distribution $d^{\mathcal{A}}_\pi$ is strictly positive and the impact samples $U^{i \to j}_\pi$ are greater or equal to zero, we see that $TI^{i \to j}(\pi) = 0$ if and only if $U^{i \to j}_\pi(s,a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. When we observe Equation (9), we see that $TI^{i \to j}(\pi) = 0$ if and only if $Q^j(s, a^{-i}, a^i) = Q^j(s, a^{-i}, \bar{a}^i)$ for all $s \in \mathcal{S}, a^{-i} \in \mathcal{A}^{-i}$ and $a^i, \bar{a}^i \in \mathcal{A}^i$. Therefore, the constant matrix-valued function $TI_\pi : \Omega \to [0, \infty)^{N \times N}$, with entries given by $(TI_\pi)_{i,j}(\omega) = TI^{i \to j}(\pi)$, is a total influence measurement function by Definition 4.2.

That means, if we can estimate TIM reliably, we obtain an unbiased total influence measurement function. Its semantic meaning is determined by the impact sample, i.e., it represents the maximum impact of an agent on the expected long-term return of another. In general, one does not know the individual state-action functions, but only some approximations of them. We denote an approximation of an individual state-action function by $\overline{Q}^j_\pi$ and a resulting formulation of an approximated TIM using Equation (11) by $\overline{TI}^{i \to j}_\pi$. The following theorem gives an error bound between the approximated TIM and the true TIM, depending on the individual state-action functions' approximation error.

Theorem 5.3. The error of the approximated TIM to the true TIM of agent i on agent j satisfies

$$\left| TI^{i \to j}(\pi) - \overline{TI}^{i \to j}(\pi) \right| \leq 2 \cdot \left\| Q^j_\pi - \overline{Q}^j_\pi \right\|_\infty.$$

Proof. Let $i, j \in \mathcal{N}$, then we see

$$\begin{aligned}
\left| TI^{i \to j}(\pi) - \overline{TI}^{i \to j}(\pi) \right| &\leq \sum_{s \in \mathcal{S}} d_\pi(s) \sum_{a \in \mathcal{A}} \pi(s,a) \cdot \\
&\left(\left| \max_{a^i \in \mathcal{A}^i} Q^j_\pi(s, a^{-i}, a^i) - \max_{a^i \in \mathcal{A}^i} \overline{Q}^j_\pi(s, a^{-i}, a^i) \right| \right. \\
&\left. + \left| \min_{a^i \in \mathcal{A}^i} Q^i_\pi(s, a^{-i}, a^i) - \min_{a^i \in \mathcal{A}^i} \overline{Q}^j_\pi(s, a^{-i}, a^i) \right| \right) \\
&\leq \sum_{s \in \mathcal{S}} d_\pi(s) \sum_{a \in \mathcal{A}} \pi(s,a) \left(2 \cdot \left\| Q^j_\pi - \overline{Q}^j_\pi \right\|_\infty \right) \\
&= 2 \cdot \left\| Q^j_\pi - \overline{Q}^j_\pi \right\|_\infty.
\end{aligned}$$

□

This bound shows that if we can determine $\overline{TI}^{i\to j}(\pi)$, we get a good approximation of TIM provided that the approximation error of $\overline{Q}^j_\pi$ is small. For an approximation function, we consider parametrized function classes. Denote with $Q^j_\pi : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^{k_j} \to \mathbb{R}$ the individual state-action function of agent j, parametrized by $\eta^j \in \mathbb{R}^{k_j}$ for $k_j \in \mathbb{N}$. We denote the parametrized impact samples and TIM by $U^{i\to j}_\pi(s, a, \eta^j)$ and $TI^{i\to j}(\pi, \eta^j)$ respectively.

Our proposed approximation algorithm of TIM works together with a simultaneously learning state-action function approximation algorithm, which provides an iteration sequence $\{\eta^j_t\}_{t\geq 0}$. For our later results, we state two mild assumptions on the iteration algorithm creating $\{\eta^j_t\}_{t\geq 0}$ and the parametrized individual state-action functions.

Assumption 5.4. The parametrized state-action function $Q^j(s, a, \eta)$ is continuous in $\eta \in \mathbb{R}^{k_j}$, for every $j \in \mathcal{N}$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$.

Assumption 5.5. Let $j \in \mathcal{N}$. We assume that the iteration sequence $\{\eta^j_t\}_{t\geq 0}$ is almost surely bounded, i.e., there exists a $K > 0$ such that $\sup_{t\geq 0} \left\| \eta^j_t \right\| < K < \infty$ almost surely. Additionally, there exists an $\eta^{j,*} \in \mathbb{R}^{k_j}$ such that $\eta^j_t \to \eta^{j,*}$ almost surely.

The above assumption essentially demands that the iteration algorithm, to approximate the individual state-action function, converges at some point. The considered iteration algorithm of TIM with parametrized individual state-action functions is given by

$$\nu^{i\to j}_{t+1} = (1 - \alpha_t)\nu^{i\to j}_t + \alpha_t \cdot U^{i\to j}_\pi(s_t, a_t, \eta^j_t), \qquad (12)$$

where $\{\alpha_t\}_{t\geq 0}$ is a stepsize sequence satisfying part (c) of Assumption 3.3. With this, we state our main result.

Theorem 5.6. Under Assumptions 3.2, 5.4, and 5.5, the iteration defined in Equation (12) has the following property

$$\nu^{i\to j}_{t+1} \to TI^{i\to j}(\pi, \eta^{j,*}_\pi) \text{ almost surely.} \qquad (13)$$

Proof. We define

$$\begin{aligned} h(\nu^{i\to j}_t, (s_t, a_t)) &:= U^{i\to j}_\pi(s_t, a_t, \eta^{j,*}_\pi) - \nu^{i\to j}_t, \\ M_{t+1} &:= 0, \\ \beta_{t+1} &:= U^{i\to j}_\pi(s_t, a_t, \eta^j_t) - U^{i\to j}_\pi(s_t, a_t, \eta^{j,*}_\pi), \end{aligned}$$

where we can see that the iteration algorithm

$$\nu^{i\to j}_{t+1} = \nu^{i\to j}_t + \alpha_t \cdot \left(h(\nu^{i\to j}_t, (s_t, a_t)) + M_{t+1} + \beta_{t+1} \right)$$

is equal to the iteration algorithm defined in Equation (12). To show the convergence of this iteration algorithm, we consider in the first step a slightly different algorithm. For this, observe that by Assumption 5.5, the sequence $\{\eta^j_t\}_{t\geq 0}$ is almost surely bounded. That means there exists $K > 0$ such that $P(\sup_{t\geq 0} \left\| \eta^j_t \right\| < K) = 1$. Define the error term $\tilde{\beta}_t := \mathbb{I}_{\{\sup_{t\geq 0} \|\eta^j_t\| < K\}} \cdot \beta_t$, where $\mathbb{I}_A$ denotes the indicator function on a set A. Define the following iteration algorithm with the restricted error sequence $\tilde{\beta}_t$

$$\tilde{\nu}^{i\to j}_{t+1} := \tilde{\nu}^{i\to j}_t + \alpha_{\nu,t} \cdot \left(h(\tilde{\nu}^{i\to j}_t, (s_t, a_t)) + \tilde{\beta}_{t+1} \right). \qquad (14)$$

First, we check that parts (a) to (e) from Assumption 3.3 hold. The function $h : \mathbb{R}^{k_j} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is Lipschitz continuous in its first argument, i.e.,

$$|h(\nu, (s, a)) - h(\nu', (s, a))| = |\nu - \nu'| \text{ for } s \in \mathcal{S}, a \in \mathcal{A},$$

and part (a) holds. By Assumption 3.2, the Markov chain $\{(s_t, a_t)\}_{t\geq 0}$ is ergodic, which means that it satisfies part (b). Furthermore, the stepsizes $\{\alpha_t\}_{t\geq 0}$ satisfy part (c). As the sequence $\{M_t\}_{t\geq 0}$ is the zero sequence, it is trivially a martingale difference sequence with a conditionally bounded norm, and satisfies part (d). Finally, for part (e), it remains to show that $\{\tilde{\beta}_t\}_{t\geq 0}$ is a bounded random sequence that converges to zero almost surely. For this, note that η^j_t is uniformly bounded on the set $\{\sup_{t\geq 0} \left\| \eta^j_t \right\| < K\}$. By Assumption 5.4, we get that the parametrized impact samples $U^{i\to j}_\pi(s, a, \eta^j)$ are continuous in η^j. In particular, as it is a continuous function on the compact set $\{\eta^j \in \mathbb{R}^{k_j} : \left\| \eta^j \right\| < K\}$, we get that it is bounded. Therefore, together with the convergence of $\eta^j_t \to \eta^{j,*}_\pi$, we get that $\{\tilde{\beta}_t\}_{t\geq 0}$ is a bounded random sequence that converges to zero. Therefore, Assumption 3.3 is satisfied for the iteration algorithm from Equation (14). Next, consider the ODE given by

$$\begin{aligned} \dot{\nu}^{i\to j} &= \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \pi(s, a) \cdot h\left(\nu^{i\to j}, (s, a)\right) \\ &= -\nu^{i\to j} + \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \pi(s, a) \cdot U^{i\to j}_\pi(s, a, \eta^{j,*}) \end{aligned}$$

and define the right-hand side as $f(\nu^{i\to j})$. We can see that $\nu^{i\to j} = \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \pi(s, a) U^{i\to j}_\pi(s, a, \eta^{j,*})$ is an equilibrium solution to the ODE above, and as f is Lipschitz continuous, we get by the theorem of Picard-Lindelöf, see page 89 in the book of Adkins and Davidson (2012), that this solution is unique. Define the function $f_c(\nu^{i\to j}) = c^{-1} \cdot f(c\nu^{i\to j})$. Then $\lim_{c\to\infty} f_c(\nu^{i\to j}) = -\nu^{i\to j} =: f_\infty(\nu^{i\to j})$ exists and the ODE $\dot{\nu}^{i\to j} = f_\infty(\nu^{i\to j})$ has the origin as unique asymptotically stable equilibrium. Therefore, we get by Theorem 3.5 that $\sup_{t\geq 0} ||\nu^{i\to j}_t|| < \infty$ almost surely. Then, we can use Theorem 3.4 to conclude that

$$\tilde{\nu}^{i\to j}_t \to TI^{i\to j}(\pi, \eta^{j,*}_\pi) \text{ a.s.} \qquad (15)$$

To extend this result to the original iteration sequence, observe that $\left\{ \sup_{t\geq 0} \left\| \eta^j_t \right\| \geq K \right\}$ is a null-set. □

The State Impact Measurement

TIM averages the maximum impact one agent can have on the individual state-action function of another over all possible transitions. However, given a specific state, some agents might have a significant impact on the individual state-action functions of others, even though their average influence is small. Therefore, one would like to quantify state-dependent influence structures among the agents. Therefore, we introduce the state impact measurement, which constitutes a state influence measurement function by Definition 4.1.

Definition 5.7 (State impact measurement)**.** Let π be a joint policy of the N agents over the joint action space $\mathcal{A}$. Take the state $s \in \mathcal{S}$ and denote the Markov chain over the actions in state s by $\{a^s_{t^s}\}_{t^s \geq 0}$. The state impact measurement (SIM) of agent i on agent j, for $i, j \in \mathcal{N}$ is defined as

$$\mathit{SI}^{i \to j}(s, \pi) := \lim_{T^s \to \infty} \frac{1}{T^s} \sum_{t^s=0}^{T^s-1} \mathbb{E}\left[U^{i\to j}_{\pi}(s, a^s_{t^s})\right], \qquad (16)$$

where $U^{i \to j}_{\pi}(s, a)$ denotes the impact sample from Definition 5.1.

Note that SIM only considers the Markov chain over the actions $\{a^s_{t^s}\}_{t^s \geq 0}$ given a specific state $s \in \mathcal{S}$. Hence, one ignores the state transition probabilities of the underlying MDP and only considers the distribution over the joint actions for a state s. As we only consider the actions for a given state s, $\pi(s, \cdot)$ is the stationary distribution of the Markov chain $\{a^s_{t^s}\}_{t^s \geq 0}$. Therefore, one can represent SIM by

$$\mathit{SI}^{i \to j}(s, \pi) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot U^{i \to j}_{\pi}(s, a). \qquad (17)$$

Under Assumption 3.2, one can record the instances of $\{a_t\}_{t \geq 0}$ for each state s in a tabular fashion, which allows sampling from $\{a^s_{t^s}\}_{t^s \geq 0}$. With this insight, one can observe that the theoretical results from above carry over with only slight variations in the proofs. Therefore, we state the following two results without proof.

First, we give an error bound similar to the statement from Theorem 5.3. We denote the approximated SIM by $\overline{\mathit{SI}}^{i \to j}$ using the approximated individual state-action function $\overline{Q}^{j}_{\pi}$.

Theorem 5.8. Let s be in $\mathcal{S}$. The error of the approximated SIM in s to the true one of agent i on agent j satisfies

$$\left|\mathit{SI}^{i \to j}(s, \pi) - \overline{\mathit{SI}}^{i \to j}(s, \pi)\right| \leq 2 \cdot \left\|Q^j_{\pi}(s, \cdot) - \overline{Q}^j_{\pi}(s, \cdot)\right\|_{\infty}.$$

We denote the parametrized SIM by $\mathit{SI}^{i \to j}(\cdot, \pi, \eta^j)$ for $i, j \in \mathcal{N}$ and $\eta^j \in \mathbb{R}^{k_j}$. The approximation algorithm is

$$\nu^{i \to j}_{t^s+1}(s) = (1 - \alpha_{t^s})\nu^{i \to j}_{t^s}(s) + \alpha_{t^s} \cdot U^{i \to j}_{\pi}(s, a^s_{t^s}, \eta^j_t), \quad (18)$$

where $\{\alpha_{t^s}\}_{t^s \geq 0}$ denotes a stepsize sequence satisfying part (c) of Assumption 3.3.

Theorem 5.9. Under Assumptions 3.2, 5.4, and 5.5, the iteration defined in Equation (18) has the following convergence property for every $s \in \mathcal{S}$

$$\nu^{i \to j}_{t^s+1}(s) \to \mathit{SI}^{i \to j}(s, \pi, \eta^{j,*}_{\pi}) \text{ almost surely.} \qquad (19)$$

Continuity in Policy Changes

The preceding analyses treated the joint policy π as fixed. In the following, we relax this restriction and show that TIM and SIM are continuous in changes of the policy π, which is crucial for applications as one can expect the approximation algorithm's behavior to be highly unstable otherwise.

We consider parameterized functions to track changes in the policies. Let $\theta^j \in \mathbb{R}^{m_j}$ for $m_j \in \mathbb{N}$ and $\pi^j_{\theta^j}$ be the policy of agent j. Denote with $\theta = [(\theta^1)^T, \ldots, (\theta^N)^T]^T \in \mathbb{R}^m := \prod_{j \in \mathcal{N}} \mathbb{R}^{m_j}$ the joint policy parameters, and denote the parametrized joint policy by $\pi_\theta = \prod_{j \in \mathcal{N}} \pi^j_{\theta^j}$. Note that when we require Assumption 3.2 to hold, that it is assumed that the parametrized policies have a positive probability for every state and action. Furthermore, we assume the following for the parametrized policies:

Assumption 5.10. The function $\pi^j_{\theta^j}(s, a^j)$ is continuously differentiable in $\theta^j \in \mathbb{R}^{m_j}$ for $s \in \mathcal{S}, a^j \in \mathcal{A}^j$, and $i \in \mathcal{N}$.

To prove the continuity of TIM and SIM in θ, one needs to establish the continuity of the stationary distribution d_θ, the joint policy π_θ, and the impact samples $U^{i \to j}_{\theta}$. We omit the proof due to space restrictions.

Theorem 5.11. Let $\Theta \subset \mathbb{R}^m$ be a compact set, and let π_θ be the joint policy. Under Assumptions 3.2 and 5.10, the total impact measurement $\mathit{TI}^{i \to j}(\pi_\theta)$ and state impact measurement $\mathit{SI}^{i \to j}(s, \pi_\theta)$ are continuous in $\theta \in \Theta$ for every $i, j \in \mathcal{N}$ and $s \in \mathcal{S}$.

6 Empirical Results

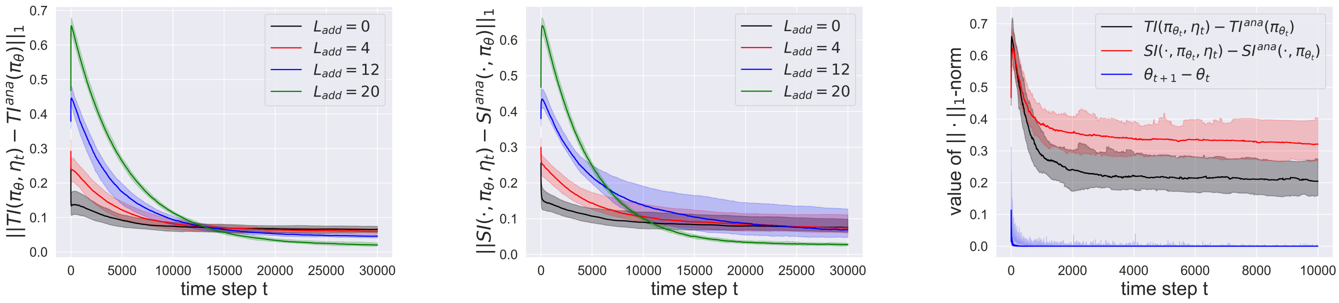
The applied techniques to guarantee convergence of our proposed algorithms do not provide statements about specifics of the convergence behavior (Borkar 2008). We present empirical results to test our approximation algorithms in practice in this section.

Setup and Methodology

The used environment is a randomly generated multi-agent MDP in the form described by Zhang et al. (2018). We chose randomly generated environments since they represent a problem-independent way to evaluate our methods. Furthermore, one can control their inherent structure, such that we can, e.g., analytically approximate TIM and SIM.

We generate random environments with $|\mathcal{N}| = 5$ agents, $|\mathcal{S}| = 5$ states, and binary action spaces $\mathcal{A}^j = \{0, 1\}$. The randomly generated embeddings of the state-action pairs (s, a^j) are denoted by $\phi(s, a^j) \in \mathbb{R}^{m_j}$. We sample the entries of the policy parameters $\theta^j \in \mathbb{R}^{m_j}$ uniformly from $[-\frac{1}{\sqrt{m_j}}, \frac{1}{\sqrt{m_j}}]$. To determine the probability distribution of the policies $\pi^j_{\theta^j}$, we use the Boltzmann policies (Sutton and Barto 2018). Note that this setup satisfies Assumption 3.2 and the policies satisfy Assumption 5.10.

We want to test TIM's and SIM's approximation algorithms with different influence structures among the agents. Therefore, when we conduct experiments where some agents are independent of others, we adjust the environment in the following way. We set the rows of the transition probability matrix all equal to the first row, i.e., $P(s'|\cdot, \cdot) = P(s^0|\cdot, \cdot)$ for all $s' \in \mathcal{S}$. This prevents the agents to influence one another over long-term effects, depending on the transitions to other states. To achieve that



(a) TIM approximation error with static π_θ and varying dependency structures

(b) SIM approximation error with static π_θ and varying dependency structures

(c) TIM and SIM approximation error with learning π_{θ_t}

Figure 1: Empirical performance of the approximation algorithms of TIM and SIM. Each scenario was conducted for 50 different seeds. The bold line represents the median and the shaded areas the 95%-quantiles.

agent j is independent of the immediate effects on the reward of agent i's actions, we set the entries for a state $s \in \mathcal{S}$, and actions $a^{-i} = (a^1, \ldots, a^{i-1}, a^{i+1}, \ldots a^N) \in \mathcal{A}^{-i}$ in the reward matrix to $R^j(s, a^{-i}, a^i) = R^j(s, a^{-i}, \overline{a}^i)$ for all $a^i, \overline{a}^i \in \mathcal{A}^i$. If we change the environment in this way, we say that we enforced some influence structure on the agents.

In general, this randomly generated multi-agent MDP allows us to approximate TIM and SIM analytically. With this, we can track the performance of the approximation algorithms precisely. The approximated TIM and SIM matrices using this state-action function approximation are denoted by $\mathit{TI}^{\text{ana}}(\pi_\theta)$ and $\mathit{SI}^{\text{ana}}(s, \pi_\theta)$.

As learning algorithm for the state-action functions $Q^j_\theta(\cdot, \cdot, \eta^j)$, we use the tabular SARSA algorithm for the average-reward setting (Sutton and Barto 2018). The learning rates α and β are set to $\alpha = \beta = 0.036$. The entries of the state-action table are initially set to one. Note that this learning algorithm satisfies Assumptions 5.4 and 5.5. The TIM and SIM approximation matrices using the learning state-action function for the algorithm from Equations (12) and (18) are denoted by $\mathit{TI}(\pi_\theta, \eta_t)$ and $\mathit{SI}(s, \pi_\theta, \eta_t)$.

For both approximation algorithms, we initialize the approximation of TIM and SIM for all $i, j \in \mathcal{N}$ to $\frac{1}{|\mathcal{N}|} = 1/5$ and set the learning rates to $\alpha^{\text{TIM}}_t = \frac{0.471}{t^{0.726}}$ and $\alpha^{\text{SIM}}_t = \frac{0.74}{t^{0.539}}$.

In the first experiment, we consider a static policy π_θ for different dependency structures among the agents. We assume that each agent can at least influence its state-action function. To determine the overall dependency structures among the agents, we randomly sample a number of additional dependencies L_{add} from the remaining ones.

The second experiment measures the above errors with changing policy parameters θ_t without enforcing any influence structure. As the policies' learning algorithm, we use Algorithm 1 of Zhang et al. (2018), which is a multi-agent actor-critic algorithm for a fully cooperative setup.

Results

The first experiment's results can be seen in Figures 1a and 1b. They show the approximation errors of $\mathit{TI}(\pi_\theta, \eta_t)$ to $\mathit{TI}^{\text{ana}}(\pi_\theta)$ and $\mathit{SI}(s, \pi_\theta, \eta_t)$ to $\mathit{SI}^{\text{ana}}(s, \pi_\theta)$ for different values of L_{add}. In all scenarios, the error is monotonically decreasing in t. One observes that the initial approximation error increases with an increasing number of dependencies among the agents. However, the final approximation error has the reversed order. This results from the fact that detecting that two agents are independent, the impact samples need to be zero. However, a non-zero approximation error in the individual state-action functions leads to the impact samples being bounded away from zero, which leads to an overestimation of the TIM and SIM approximations.

Figure 1c shows the second experiment's results for a learning policy. The blue line shows the $\|\cdot\|_1$-norm of the differences in the policy parameters from t to $t+1$. One can observe that the approximation error of TIM and SIM decrease consistently, but slower and with a higher error than in the experiment with a static policy. Nonetheless, this experiment demonstrates the validity of using the approximation algorithm in the context of the policies' learning algorithm.

7 Final Remarks

The present work investigates influence structures in MARL systems. We introduce influence measurement functions as a unified descriptive framework for influence structures in all common setups. The total and state impact measurements are presented, which constitute instances of influence measurement functions in the average reward setting. Thorough theoretical analyses of their stability and corresponding approximation algorithms' convergence and error bounds are given. The empirical experiments indicate that the proposed algorithms are promising for practical applications.

There are various directions of future work. A more extensive study of TIM's and SIM's expressive prowess in environments with a semantic meaning could be conducted. Another valuable contribution would be extending the theoretical results of SIM's approximation with function approximation, to make it tractable for large state spaces. Lastly, one could introduce and study instances of influence measurement functions in other settings, such as the discounted reward setting, or with infinite state and action spaces.

References

Adkins, W. A.; and Davidson, M. G. 2012. *Ordinary Differential Equations*. Undergraduate Texts in Mathematics. New York: Springer New York, first edition. ISBN 978-1-4614-3617-1.

Böhmer, W.; Kurin, V.; and Whiteson, S. 2019. Deep Coordination Graphs. *arXiv e-prints*, arXiv:1910.00091.

Borkar, V. S. 2008. *Stochastic Approximation*, volume 48 of *Texts and Readings in Mathematics*. Gurgaon: Hindustan Book Agency, first edition. ISBN 978-81-85931-85-2.

Callaway, D. S.; and Hiskens, I. A. 2011. Achieving Controllability of Electric Loads. *Proceedings of the IEEE*, 99(1): 184–199.

Fax, J.; and Murray, R. 2004. Information Flow and Cooperative Control of Vehicle Formations. *IEEE Transactions on Automatic Control*, 49(9): 1465–1476.

Guestrin, C.; Lagoudakis, M. G.; and Parr, R. 2002. Coordinated Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 227–234. Sydney: Morgan Kaufmann Publishers Inc. ISBN 1-55860-873-7.

Guestrin, C.; Venkataraman, S.; and Koller, D. 2002. Context-Specific Multiagent Coordination and Planning with Factored MDPs. In *Eighteenth National Conference on Artificial Intelligence*, 253–259. Edmonton, Alberta: American Association for Artificial Intelligence. ISBN 0-262-51129-0.

Hoen, P. J.; Tuyls, K.; Panait, L.; Luke, S.; and Poutré, J. A. L. 2005. An Overview of Cooperative and Competitive Multiagent Learning. In Tuyls, K.; Hoen, P. J.; Verbeeck, K.; and Sen, S., eds., *Learning and Adaption in Multi-Agent Systems*, volume 3898 of *Lecture Notes in Computer Science*, 1–46. Utrecht: Springer.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P. A.; Strouse, D.; Leibo, J. Z.; and de Freitas, N. 2018. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. *arXiv e-prints*, arXiv:1810.08647.

Kok, J. R.; 't Hoen, P. J.; Bakker, B.; and Vlassis, N. 2005. Utile Coordination: Learning Interdependencies among Cooperative Agents. In *Proceedings of the Symposium on Computational Intelligence and Games*, 29–36. Colchester, Essex: IEEE.

Puterman, M. L. 1994. *Markov Decision Processes*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN 978-0-470-31688-7.

Seneta, E. 2006. *Non-Negative Matrices and Markov Chains*. Springer Series in Statistics. New York: Springer Science & Business Media, second edition. ISBN 978-0-387-29765-1.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; and Graepel, T. 2018. Value-Decomposition Networks for Cooperative Multi-Agent Learning Based on Team Reward. In *Seventeenth International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, 2085–2087. Stockholm: International Foundation for Autonomous Agents and Multiagent Systems.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: A Bradford Book, second edition. ISBN 0-262-03924-9.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2020. Influence-Based Multi-Agent Exploration. In *Eighth International Conference on Learning Representations*. Addis Ababa: OpenReview.net.

Wang, T.; Zeng, L.; Dong, W.; Yang, Q.; Yu, Y.; and Zhang, C. 2021. Context-Aware Sparse Deep Coordination Graphs. *arXiv:2106.02886 [cs]*.

Zhang, C.; and Lesser, V. 2013. Coordinating Multi-Agent Reinforcement Learning with Limited Communication. In *Twelfth International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '13, 1101–1108. St. Paul, Minnesota: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-1993-5.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. In Dy, J. G.; and Krause, A., eds., *Thirty-Fifth International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5867–5876. Stockholm: PMLR.