Context-Aware Bayesian Network Actor-Critic Methods for Cooperative Multi-Agent Reinforcement Learning

Dingyang Chen [1] **Qi Zhang** [1]

Abstract

Executing actions in a correlated manner is a common strategy for human coordination that often leads to better cooperation, which is also potentially beneficial for cooperative multi-agent reinforcement learning (MARL). However, the recent success of MARL relies heavily on the convenient paradigm of purely decentralized execution, where there is no action correlation among agents for scalability considerations. In this work, we introduce a Bayesian network to inaugurate correlations between agents' action selections in their joint policy. Theoretically, we establish a theoretical justification for why action dependencies are beneficial by deriving the multi-agent policy gradient formula under such a Bayesian network joint policy and proving its global convergence to Nash equilibria under tabular softmax policy parameterization in cooperative Markov games. Further, by equipping existing MARL algorithms with a recent method of differentiable directed acyclic graphs (DAGs), we develop practical algorithms to learn the context-aware Bayesian network policies in scenarios with partial observability and various difficulty. We also dynamically decrease the sparsity of the learned DAG throughout the training process, which leads to weakly or even purely independent policies for decentralized execution. Empirical results on a range of MARL benchmarks show the benefits of our approach.

1. Introduction

Cooperative multi-agent reinforcement learning (MARL) methods equip a group of autonomous agents with the capability of planning and learning to maximize their joint utility, or reward signals in the reinforcement learning (RL) literature, which provides a promising paradigm for a range of real-world applications, such as traffic control (Chu et al., 2019), coordination of multi-robot systems (Corke et al., 2005), and power grid management (Callaway & Hiskens, 2010). As a key distinction from the single-agent setting, multi-agent joint action spaces grow exponentially with the number of agents, which imposes significant scalability issues. As a convenient and commonly adopted solution, most existing cooperative MARL methods only consider *product policies*, i.e., each agent selects its local action independently given the state or its observations. Restricting to product policies, however, does come at a cost for cooperative tasks: consider an example where cars wait at a crossroads, it would be hard for the cars to coordinate their movements without knowing others' intentions, potentially resulting in a crash or congestion. Intuitively, optimizing over the smaller joint policy space of all product policies can lead to suboptimal joint policies compared to optimizing over the entire set of joint policies that also includes *correlated policies* where the local actions of all agents are sampled together in a potentially correlated manner.

The research question then arises naturally: how can we introduce correlations for cooperative multi-agent joint policies, while taming the scalability issues? Noting that a joint policy is joint distributions (over agents' local actions), a straightforward yet underexplored solution idea is to use a Bayesian network (BN) that represents conditional dependencies between agents' local actions via a directed acyclic graph (DAG), where a desirable DAG topology structure captures important dependencies that exist among hopefully a set of sparsely connected agents. As our first contribution, we formalize this solution idea of BN joint policies in the cooperative Markov game framework (Boutilier, 1999; Peshkin et al., 2001), derive its associated BN policy gradient formula, and then prove the global convergence of its gradient ascent to Nash equilibria under the tabular policy parameterization.

As our second contribution, we then adapt existing multi-agent actor-critic methods such as MAPPO (Yu et al., 2021) to incorporate BN joint policies. For practicality and efficiency, our algorithm features the following two key design

[1]Artificial Intelligence Institute, University of South Carolina, SC, USA. Correspondence to: Dingyang Chen <dingyang@email.sc.edu>, Qi Zhang <qz5@cse.sc.edu>.

choices: (i) Our DAG topology of the BN joint policy is learnable to be context-aware based on the environment state or the agents' joint observations, leveraging a recently developed technique for differentiable DAG learning. (ii) To execute a BN joint policy, the agents need to communicate their intended actions to their children in the BN, unless the BN's DAG topology reduces to product policies, and the corresponding communication overhead is directly determined by the DAG's denseness/sparseness. To encourage sparse communication during execution, we develop a learning strategy that dynamically increases the sparsity of the learned DAG, where full sparsity (i.e., product policies) can be achieved at the last stage of the training process and therefore the learned joint policy can be executed in a purely decentralized manner, making our algorithm compatible with the centralized training, decentralized execution (CTDE) paradigm (Lowe et al., 2017). Empirically results show the benefits of our algorithm equipped with these two design choices.

The rest of this paper is structured as follows: Section 2 reviews closely related work; Section 3 introduces preliminaries of cooperative Markov games and solution concepts therein; Section 4 formulates our novel notion of Bayesian network joint policy, followed by the theoretical results in Section 5.1; Section 6 describes our practical algorithm, followed by the empirical results in Section 7; Section 8 concludes the paper.

2. Related Work

Convergence of policy gradient in cooperative MGs. Cooperative MGs are an important subclass of Markov games, where each agent has the same reward function. Recent work has established the convergence guarantee of policy gradient in Cooperative MGs to Nash policies under tabular setting with direct parameterization (Leonardos et al., 2021), and with softmax parameterization (Zhang et al., 2022; Chen et al., 2022).

Policy correlations in MARL. Some prior work has noticed the limitation of purely decentralized execution and made some efforts to introduce correlations among policies. Value-based method (Rashid et al., 2018) following the CTDE-based training paradigm has been combined with coordination graph (Böhmer et al., 2020) for introducing pairwise correlation. However, the optimization process requires Max-Sum (Rogers et al., 2009) which is computationally intensive when the coordination graph is dense. (Wang et al., 2022) proposes a rule-based pruning method to generate a sparse coordination graph that can speed up the Max-Sum algorithm without harming the performance. However, the extension from the pairwise correlation to more complicated ones is not trivial. There are also some

policy-based algorithms augmented with correlated execution. (Ruan et al., 2022) combines MAPPO (Yu et al., 2021) with a graph generator outputting Bayesian Network that determines action dependencies. The optimization of the graph generator is achieved by maximizing the cumulative rewards constrained to the depth and dagness of the output graph. However, there is no theoretical justification for why using Bayesian networks is reasonable, and the output graph is not guaranteed to be a DAG, which requires some non-differentiable rule-based pruning and can harm performance. Moreover, the existing methods do not generate fully decentralized policies at the end of the training, which increases the execution time in the deployment of the model.

Differentiable DAG learning. The goal is to learn such an adjacency matrix that can help the actors better coordinate. However, the generation of DAG requires non-differentiable operations due to its discreteness and acyclicity, which prevents end-to-end training. Fortunately, recent work (Charpentier et al., 2022) proposes a simple fully differentiable DAG learning algorithm. Every DAG can be decomposed into the multiplication of a permutation matrix determining topological ordering and upper triangular matrix (edge matrix) determining DAG structure. We can use neural networks to learn the logits for permutation and edge matrices, and use Gumbel-Softmax (Jang et al., 2016) and Gumbel-Sinkhorn (Mena et al., 2018) to differentiably transform them to the corresponding discrete ones.

3. Preliminaries

Cooperative Markov game. We consider a cooperative Markov game (MG) $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, r, \mu \rangle$ with N agents indexed by $i \in \mathcal{N} = \{1, ..., N\}$, state space $\mathcal{S}$, action space $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, (team) reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ shared by all agents $i \in \mathcal{N}$, and initial state distribution $\mu \in \Delta(\mathcal{S})$, where we use $\Delta(\mathcal{X})$ to denote the set of probability distributions over $\mathcal{X}$. For ease of exposition, we assume full observability, i.e., each agent observes the global state $s \in \mathcal{S}$, until Section 6 where we introduce our practical algorithm that incorporates partial observability. Under full observability, we consider general *joint policy*, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which maps from the state space to distributions over the joint action space. As the size of action space $\mathcal{A}$ grows exponentially with N, the commonly used joint policy subclass is the *product policy*, $\pi = (\pi^1, \cdots, \pi^N) : \mathcal{S} \to \times_{i \in \mathcal{N}} \Delta(\mathcal{A}^i)$, which is factored as the product of local policies $\pi^i : \mathcal{S} \to \Delta(\mathcal{A}^i)$, $\pi(a|s) = \prod_{i \in \mathcal{N}} \pi^i(a^i|s)$, each mapping the state space only to the action space of an individual agent. Define the discounted return from time step t as $G_t = \sum_{l=0}^{\infty} \gamma^l r_{t+l}$, where $r_t := r(s_t, a_t)$ is the team reward at time step t. Joint policy π induces a value function defined as $V_\pi(s_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim \pi}[G_t|s_t]$, and action-value func-
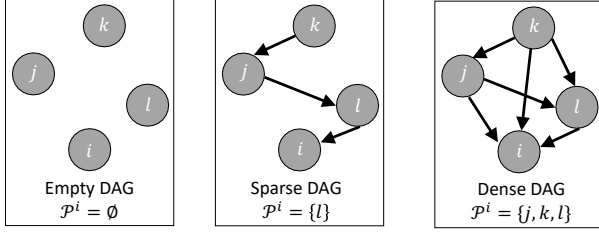
Figure 1. Illustration of various DAG topologies.

tion $Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t+1:\infty} \sim \pi}[G_t | s_t, a_t]$. Following policy π, the cumulative team reward, i.e., the value function, starting from $s_0 \sim \mu$ is denoted as $V_\pi(\mu) := \mathbb{E}_{s_0 \sim \mu}[V_\pi(s_0)]$. The (unnormalized) *discounted state visitation measure* by following policy π after starting at $s_0 \sim \mu$ is defined as

$$d_\mu^\pi(s) := \mathbb{E}_{s_0 \sim \mu}\left[\sum_{t=0}^\infty \gamma^t \Pr^\pi(s_t = s | s_0)\right]$$

where $\Pr^\pi(s_t = s | s_0)$ is the probability that $s_t = s$ when starting at state s_0 and following π subsequently.

As all agents share a team reward, cooperative MARL considers the same objective as single-agent RL of optimizing the joint policy from experience to maximize its value, i.e., $\max_\pi V_\pi(\mu)$. For product policies, we will also consider the weaker solution concept of the Nash policy, as formally defined below.

Definition 3.1 (Nash policy). Product policy $\pi = (\pi^1, \cdots, \pi^N) = (\pi^i, \pi^{-i})$ is a Nash policy if

$$\forall i \in \mathcal{N}, \forall \bar{\pi}^i \in \Delta(\mathcal{A}^i), V_{\bar{\pi}^i, \pi^{-i}}(\mu) \le V_\pi(\mu)$$

where π^{-i} is the local policies of the agents excluding i.

For a Nash policy, each agent i maximizes the value function given fixed local policies of other agents.

4. Bayesian Network Joint Policy

Most existing cooperative MARL methods consider only product policies to optimize, rather than the more general set of general joint policies. This is mainly because product policies can conveniently deal with the scalability issue of the joint action space. Another justification is that restricting to product policies incurs no optimality gap, since it is well-known that there is always an optimal joint policy that is deterministic and therefore a product policy. However, the existence of an optimal product policy does not guarantee that we can search it out easily. In fact, existing theoretical and empirical results (Leonardos et al., 2021; Zhang et al., 2022; Chen et al., 2022), including ours in this paper, have shown that restrictively searching the product policies via

gradient ascent can only find local optima such as Nash policies, even in the noiseless tabular setting.

As the key notion in this work, we now formally introduce a class of joint policies that is more general than product policies by introducing action dependencies captured by a Bayesian network (BN). We specify a BN by a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with vertex set $\mathcal{N}$ and directed edge set $\mathcal{E} \subseteq \{(i, j) : i, j \in \mathcal{N}, i \ne j\}$. Denote the parents of agent i as $\mathcal{P}^i := \{j : (i, j) \in \mathcal{E}\}$, and the corresponding parent actions as $a^{\mathcal{P}^i} \in \times_{j \in \mathcal{P}^i} \mathcal{A}^j$, as illustrated in Figure 1. Under full observability and with BN $\mathcal{G}$, we consider a *BN (joint) policy*, $(\pi, \mathcal{G}) = (\pi^1, \cdots, \pi^N, \mathcal{G}) : \mathcal{S} \to \Delta(\mathcal{A})$. Similar to product policies, BN policies are also factored as the product of local policies given the state and action dependencies determined by $\mathcal{G}$, i.e., $\pi^i : \mathcal{S} \times (\times_{j \in \mathcal{P}^i} \mathcal{A}^j) \to \Delta(\mathcal{A}^i)$, and thus joint action $a = (a^1, \cdots, a^N)$ is sampled as $\pi(a | s) = \prod_{i \in \mathcal{N}} \pi^i(a^i | s, a^{\mathcal{P}^i})$.

We make two remarks on BN policies: (i) With the factorization of a BN policy into its local policies, the concept of Nash policy in Definition 3.1 is also applicable to BN policies. (ii) BN policies naturally interpolate product policies and general joint policies, including them as two extremes: BN policies reduce to product policies when DAG $\mathcal{G}$ is an empty graph (Figure 1 (left)) and can model general joint policies when $\mathcal{G}$ is dense (Figure 1 (right)).

5. Convergence of the Tabular Softmax BN Policy Gradient in Cooperative MGs

In this section, we consider optimizing BN policies through policy gradient ascent under the tabular softmax parameterization. Under the same assumptions, we are able to extend existing convergence results from products policies to BN policies, asserting that optimizing BN policies through gradient ascent can indeed find global optima (rather than Nash) when the BN's DAG is dense.

Formally, the local policies in the BN policy are parameterized in the tabular softmax manner from the global state and parent actions, i.e., we have, for each agent i, its policy parameter

$$\theta^i = \left\{ \theta^i_{s, a^{\mathcal{P}^i}, a^i} \in \mathbb{R} : s \in \mathcal{S}, a^{\mathcal{P}^i} \in \times_{j \in \mathcal{P}^i} \mathcal{A}^j, a^i \in \mathcal{A}^i \right\}$$

and induced softmax local policy

$$\pi^i_{\theta^i}\left(a^i | s, a^{\mathcal{P}^i}\right) \propto \exp\left(\theta^i_{s, a^{\mathcal{P}^i}, a^i}\right) \quad (1)$$

with the BN policy parameterized as $\pi_\theta = (\pi^1_{\theta^1}, \cdots, \pi^N_{\theta^N})$.

In Lemma 5.1, we derive the policy gradient form for the BN policy as parameterized in Equation (1), which will used to establish our convergence results in this section.

It will be also convenient to introduce a few shorthands before stating Lemma 5.1. Consider a subset $\mathcal{M} \subseteq \mathcal{N}$ of all agents and its complement $-\mathcal{M}$, such that a joint action can be decomposed as $a = (a^{\mathcal{M}}, a^{-\mathcal{M}})$. Let

$$\pi^{\mathcal{M}}(a^{\mathcal{M}}|s, a^{-\mathcal{M}}) := \frac{\pi(a^{\mathcal{M}}, a^{-\mathcal{M}}|s)}{\sum_{\bar{a}^{\mathcal{M}}} \pi(\bar{a}^{\mathcal{M}}, a^{-\mathcal{M}}|s)}$$

be the conditional for $a^{\mathcal{M}}$ under π. Let

$$Q_\pi(s, a^{\mathcal{M}}) := \mathbb{E}_{a^{-\mathcal{M}} \sim \pi^{-\mathcal{M}}(\cdot|s, a^{\mathcal{M}})} \left[Q_\pi(s, a^{\mathcal{M}}, a^{-\mathcal{M}}) \right].$$

Let $\mathcal{P}_+^i := \mathcal{P}^i \cup \{i\}$ denote the set of agent i and its parents. We will also abbreviate $V_{\pi_\theta}, Q_{\pi_\theta}$ as V_θ, Q_θ, respectively.

Lemma 5.1 (Tabular softmax BN policy gradient form, proof in Appendix A.4). *For the tabular softmax BN policy parameterized as in Equation (1), we have:*

$$\frac{\partial V_\theta(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i},a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i}) A_\theta^i(s, a^{\mathcal{P}^i}, a^i)$$

where $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) := d_\mu^{\pi_\theta}(s) \sum_{a^{-\mathcal{P}^i}} \pi_\theta(a^{-\mathcal{P}^i}, a^{\mathcal{P}^i}|s)$, $A_\theta^i(s, a^{\mathcal{P}^i}, a^i) := Q_\theta(s, a^{\mathcal{P}_+^i}) - Q_\theta(s, a^{\mathcal{P}^i})$.

The policy gradient form in Lemma 5.1 generalizes its counterpart for single-agent policies (Agarwal et al., 2021) and for multi-agent product policies (Zhang et al., 2022; Chen et al., 2022) under the tabular softmax policy parameterization, which enables us to extend the convergence results to the BN joint policies.

Below we state the assumptions that have been used (Zhang et al., 2022; Chen et al., 2022), to generate the convergence results for product policies, i.e., $\mathcal{G} = (\mathcal{N}, \emptyset)$.

Assumption 5.2. For any π and any state s of the Markov game, $d_\mu^\pi(s) > 0$.

Assumption 5.3 (Reward function is bounded). The reward function r is bounded in the range $[r_{\min}, r_{\max}]$, such that the value function V is bounded as $V_{\min} \leq V_\pi(s) \leq V_{\max} \; \forall s, \pi$.

Assumption 5.4. Following the policy gradient dynamics (2), the policy of every agent i converges asymptotically, i.e., $\pi_{\theta_t^i}^i \to \pi_{\theta_*^i}^i$ as $t \to \infty$, $\forall i$.

Assumption 1 and assumption 2 are standard assumptions used in (Agarwal et al., 2021; Zhang et al., 2022; Chen et al., 2022), which ensures the sufficient coverage of all states and the boundness of the reward function, respectively. Assumption 3 is a stronger assumption used in (Zhang et al., 2022; Chen et al., 2022). A sufficient condition for assumption 5.4 by (Fox et al., 2022) is that the fixed point of the equation in Lemma 5.1 are isolated. The purpose of assumption 5.4 is to establish the convergence of $A_\theta^i(s, a^{\mathcal{P}_+^i})$ if $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i})$ is positive. Otherwise, it can be the case that

both $\pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i})$ and $A_\theta^i(s, a^{\mathcal{P}_+^i})$ are divergent when the gradient converges to zero.

We next present our convergence results for the standard policy gradient dynamics in Sections 5.1, where Assumptions 5.2 - 5.4 hold, with proofs in the appendix A.

5.1. Asymptotic Convergence of the Tabular Softmax BN Policy Gradient Dynamics

In Theorem 5.5, we establish, under the tabular softmax BN policy parameterization, the asymptotic convergence to a Nash policy in a MPG of the standard policy gradient dynamics:

$$\theta_{t+1}^i = \theta_t^i + \eta \nabla_{\theta^i} V_{\theta_t}(\mu) \tag{2}$$

where η is the fixed stepsize and the update is performed by every agent $i \in \mathcal{N}$.

For each agent i, parent actions $a^{\mathcal{P}^i}$, and local action a^i, Equation (2) becomes

$$\theta_{s,a^{\mathcal{P}^i},a^i}^{i,t+1} = \theta_{s,a^{\mathcal{P}^i},a^i}^{i,t} + \eta \nabla_{\theta_{s,a^{\mathcal{P}^i},a^i}^i} V_{\theta_t}^i(\mu) \tag{3}$$

Theorem 5.5 (Asymptotic convergence of BN policy gradient, proof in Appendix A.17). *Under Assumptions 5.2 - 5.4, suppose every agent i follows the policy gradient dynamics (2), which results in the update dynamics (3) for each each agent i, parent actions $a^{\mathcal{P}^i}$, and local action a^i, with $\eta \leq \frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, then the converged BN policy $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is a Nash policy.*

The main trick of our proof for Theorem 5.5 is to view the parent actions $a^{\mathcal{P}^i}$ as part of the state, i.e., $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i})$ becomes the new state visitation measure for the augmented state $(s, a^{\mathcal{P}^i})$. After this transformation, the update dynamics in 5.1 is resemble to the ones for the product policy, i.e., $\mathcal{G} := (\mathcal{N}, \emptyset)$, and thus straightforwardly generalize their results for the product joint policy to the BN policy. However, the problem with this formulation of new state $(s, a^{\mathcal{P}^i})$ is that $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) = d_\mu^{\pi_\theta}(s) \sum_{a^{-\mathcal{P}^i}} \pi_\theta(a^{-\mathcal{P}^i}, a^{\mathcal{P}^i}|s)$ can be zero even if the state visitation measure $d_\mu^{\pi_\theta}(s)$ is strictly positive. This is the main reason we cannot establish results stronger than the ones obtained in (Zhang et al., 2022), even for the fully connected Bayesian network with $N(N-1)/2$ edges which intuitively behave similar to the single-agent setting (Agarwal et al., 2021) and should therefore result in the optimal policy than only a Nash policy.

Assumption 5.6. Any augmented state $(s, a^{\mathcal{P}^i})$ has positive visitation measure, i.e., $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) > 0$.

Definition 5.7 (Fully-correlated BN policy). A BN policy $(\pi, (\mathcal{N}, \mathcal{E}))$ is fully-correlated if $|\mathcal{E}| = N(N-1)/2$, the maximum number of edges in a DAG.
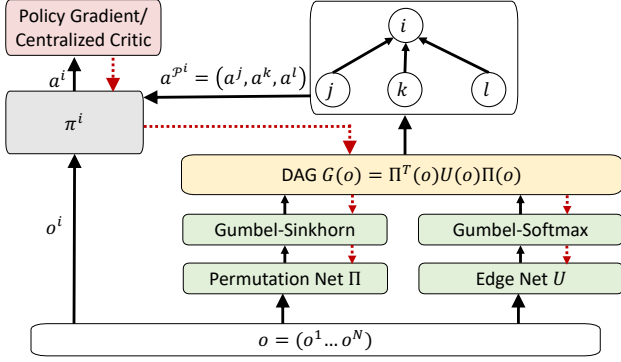
Figure 2. Our architecture for BN joint policy which includes each agent i's policy π^i and a differentiable DAG learner. DAG $G(o) = \Pi^T(o)U(o)\Pi(o)$ is generated by sending the joint local observation o to Permutation Net Π and Edge Net U. Based on $G(o)$, agent i requests actions $a^{\mathcal{P}^i} = (a^j, a^k, a^l)$ from its parents (j, k, l), which, together with local observation o^i, are taken as input into agent i's local policy π^i to output a^i. During training, the gradient (shown in the red dotted lines) flows from π^i to G, then to Π and U.

Corollary 5.8 (Asymptotic convergence of BN policy gradient to optimal fully-correlated BN joint policy, proof in Appendix A.18). *Under Assumptions 5.2 - 5.4 and additional Assumption 5.6 that assumes positive visitation measure for any augmented state, suppose every agent $i \in \mathcal{N}$ follows the policy gradient dynamics* (2), *which results in the update dynamics* (3) *for each each agent i, parent actions $a^{\mathcal{P}^i}$, and local action a^i, with $\eta \leq \frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, then the converged fully-correlated BN policy $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is an optimal policy.*

6. Practical Algorithm

The convergence guarantee in Theorem 5.5 relies on global observability and the availability of the oracle value function, which is hard to apply in more complicated scenarios. In this section, we relax those assumptions and propose an end-to-end training framework which can augment any multi-agent actor-critic methods with a differentiable Bayesian network determining action dependencies among agents' local policies. Figure 2 presents an overview of the our proposed neural architecture, consisting of the differentiable Bayesian network and the actor-critic networks as its main components that we describe below.

6.1. Differentiable Bayesian Network

The graph model G takes the joint partial observation $o = \{o^i\}_{i\in\mathcal{N}}$ as input, and outputs a DAG $G(o)$ represented by an adjacency matrix, i.e., $G(o)[j, i] = 1$ if and only if $j \in \mathcal{P}^i$. By using the same decomposition of DAG into the multiplication of permutation matrix and upper triangular

matrix (Charpentier et al., 2022) described in section 2, G consists of two sub-modules Permutation Net Π and Edge Net U which both takes the joint partial observation o as input and output the logits l_Π and l_U for the permutation matrix and upper triangular matrix, respectively. We use the reparameterization trick Straight-Through Gumbel-Softmax (Jang et al., 2016) and Gumbel-Sinkhorn (Mena et al., 2018) to differentiably transform l_Π and l_U into the corresponding permutation matrix $\Pi(o)$ and upper triangular matrix $U(o)$. The resulting DAG $G(o) = \Pi^T(o)U(o)\Pi(o)$, where $\Pi(o)$ determines the topological ordering of the agents and $U(o)$ determines the structure of the outputted DAG.

6.2. Actor and Critic Networks

Our communication network is compatible with any multi-agent actor-critic architecture. Our experiments mainly explore discrete actors which sample actions conditioning on local observation o^i and parent actions $a^{\mathcal{P}^i}$, i.e., $a^i \sim \pi^i(\cdot|o^i, a^{\mathcal{P}^i})$. The critic takes the joint local observation or the environment provided global state as input. Both actor and critics are implemented by deep neural networks with details in the appendix 2.

6.3. Training

Critic Q is trained to minimize TD error $\mathcal{L}_{\text{TD}} = \mathbb{E}_{o_t,a_t,r_t,o_{t+1}}[(Q(o_t, a_t) - y_t)^2]$, where $o_t := (o_t^1, ..., o_t^N)$, $a_t := (a_t^1, ..., a_t^N)$, and $y_t := r_t + \gamma Q(o_{t+1}, a_{t+1})$ is the TD target. Actor π^i can be updated by any multi-agent policy gradient algorithm, such as MAPPO $\mathcal{L}_{\text{actor}}^i = \mathbb{E}_{o_t,a_t}[\log\pi^i(a^i|o^i, a^{\mathcal{P}^i})A(o_t, a_t)]$. Due to the differentiability enabled by Gumbel-Softmax and Gumbel-Sinkhorn, the gradient can flow from π^i to DAG G, then to its sub-modules Permutation Net Π and Edge Net U. The DAG Density of G is defined as $\rho(G) := \frac{2}{|N(N-1)|}\sum_{i,j\in\mathcal{N}} G[j, i]$, and is regularized by the term $\alpha|\rho(G) - \eta|$. This places a restriction on the sparsity of the learned DAG by rate η.

7. Experiments

Theorem 5.5 only guarantees that the policy gradient ascent converges to Nash, but does not guarantee the solution quality of the convergent. Our experiments, in the tabular softmax Bayesian setting, aim to see how well different (fixed) DAG topologies of the BN policy perform empirically and the reasons behind it. Then, in the sample-based setting, we want to see 1) How well our algorithm proposed in Section 6 performed against baselines and ablations? 2) What are the potential meaning of the DAG learned by the context-aware differentiable DAG learner?

7.1. Environments

Our environments include (1) Coordination Game, a small-size domain where we can afford computing exact policy gradient under tabular parameterization, (2) Aloha, a domain where action correlations are intuitively helpful, and (3) StarCraft II Micromanagement (SMAC), a common cooperative MARL benchmark that is more complicated.

Coordination Game. We use the version in (Chen et al., 2022) with $N = 2, 3, 5$ agents. The state space and action space are $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^N, \mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, respectively, where $\forall i \leq N, \mathcal{S}^i \in \{0, 1\}, \mathcal{A}^i \in \{0, 1\}$. It is a cooperative setting with the same reward for all the agents, which favors more agents in the same local state. The transition function for each agent i's local state only depends on the local action: $P(s^i = 0|a^i = 0) = 1 - \epsilon, P(s^i = 0|a^i = 1) = \epsilon$, where $\epsilon = 0.1$. The performance of the learned joint policy is measured by *price of anarchy* (POA) (Roughgarden, 2015), $\frac{V_\pi(\mu)}{\max_\pi V_\pi(\mu)}$, which is bounded in the range $[0, 1]$. The convergence rate is captured by *Nash-Gap*, defined as $\text{Nash-gap}(\pi) := \max_i \left(\max_{\bar{\pi}^i} V_{\bar{\pi}^i, \pi^{-i}}(\mu) - V_\pi(\mu) \right)$, where Nash policy has a Nash-Gap of zero.

Aloha. We use the version in (Wang et al., 2022) with 10 agents (islands). 10 islands are stored in a 2×5 array, each of which has a backlog of messages to send. At each timestep, agents can either choose to send or not send. The goal is to send as many messages as possible without colliding with the ones sent by the neighboring islands. At each timestep, with a probability of 0.6, a new message can be generated for each agent. For each successfully sent message without collision, all agents receive a 0.1 reward, and a -10 reward if with collision.

StarCraft II Micromanagement (SMAC). SMAC (Samvelyan et al., 2019) has become one of the most popular MARL benchmarks. We choose the *Super Hard* scenarios 6h_vs_8z and MMM2 to evaluate our proposed algorithm, which has 6 agents and 10 agents, respectively.

7.2. Baselines

As baselines to compare against our context-aware DAG topology learning to bring in correlations between local policies, we consider the following DAGs that are fixed during training (i.e., no context-awareness). The **Fully-correlated** baseline has DAG $(\mathcal{N}, \{(j, i)|i > j\})$, which have the maximum number of $(N(N-1)/2)$ edges for any DAG. **Uncorrelated** has DAG $(\mathcal{N}, \emptyset)$, i.e., product policy. **Line-correlated** has DAG $(\mathcal{N}, \{(j, i)|i = j + 1\})$. The DAGs of all baselines have a topological ordering of $(1, 2, \cdots, N)$, i.e., Π defined in Section 6.1 is fixed as the identity matrix.
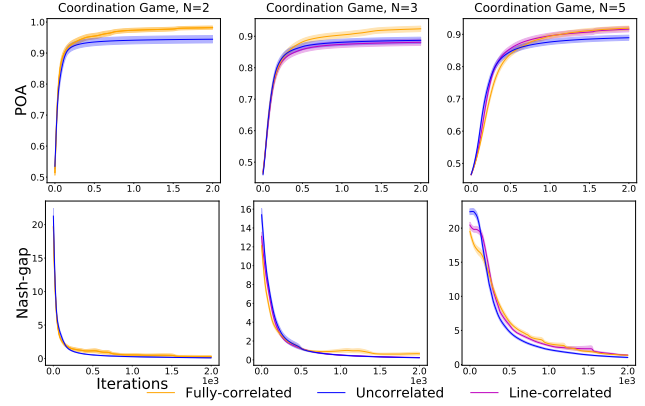

Figure 3. POA (top) and Nash-gap (bottom) under the tabular softmax BN policy gradient dynamics with various BN DAG topologies (means and standard errors over 50 random seeds). Note that with $N = 2$ agents, Line-correlated and Fully-correlated are the same and thus have overlapping curves.

7.3. Results of Fixed DAG Topologies with Tabular Exact Policy Gradients

Figure 3 presents the POA and the Nash-gap of the algorithms under the tabular softmax parameterization with different DAG topologies. The results demonstrate that the Nash-gap indeed decreases and converges close to zero as proved in Theorem 5.5. Fully-correlated consistently outperforms Line-correlated and Uncorrelated but does not converge to an optimal policy with POA of 1, because Assumption 5.6 is violated, i.e., some $(s, a^{\mathcal{P}^i})$ has visitation probability converges to zero. On the other hand, the convergence rate of Fully-correlated is the slowest and one possible reason is that it has the most number of parameters. Line-correlated has a similar performance to Fully-correlated in scenarios with $N = 2, 5$ agents, but it has poor performance in the scenario with 3 agents. This illustrates the fact that fixed DAG topology is not desirable in all scenarios and can degenerate to the performance of Uncorrelated.

7.4. Results of Context-Aware DAG Topology Learning with Multi-Agent Actor-Critic Methods

In this section, we run experiments to compare our context-aware DAG topology against the baselines in Coordination Game, where we assume global observability, and in Aloha and SMAC, where we assume partial observability. We relax the requirement of only sharing local action to also include local observations when finding beneficial, based on the context-aware DAG. Specifically, based on the context-aware DAG, the experiments in Aloha share both local actions and observations, whereas the ones in SMAC only share local actions. We implement the algorithms based on MAPPO without recurrency, i.e., the model only incorporates information from the current timestep instead of from
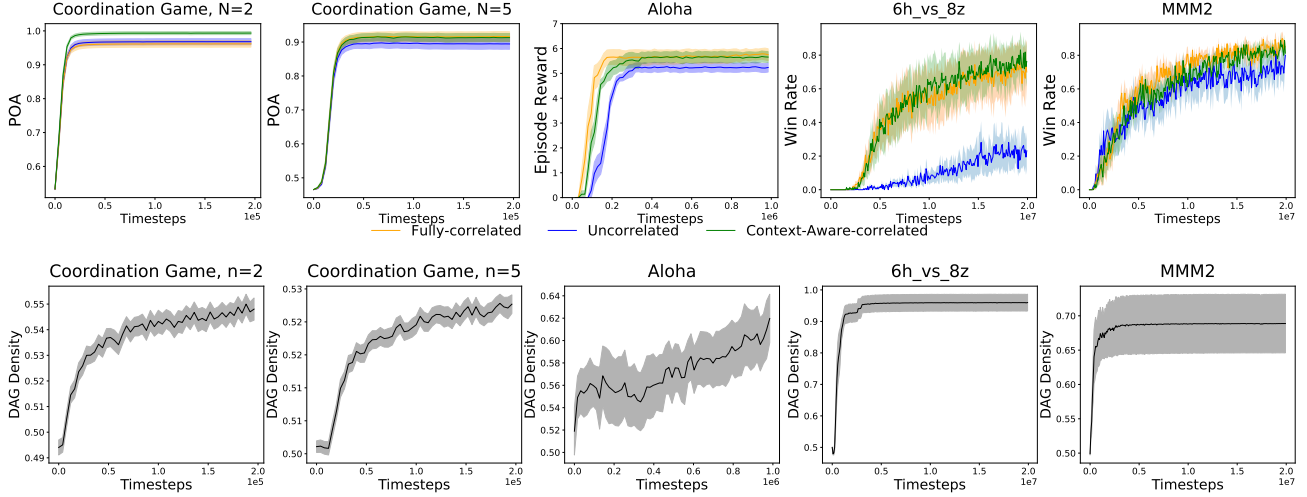
Figure 4. Top: Performance of the learned context-aware-correlated against the Fully-correlated and Uncorrelated baselines. (means and standard errors over 50 random seeds for Coordination Game, 10 random seeds for Aloha, and 5 random seeds for 6h_vs_8z and MMM2.) *Bottom:* Changes in DAG Density of the learned context-aware BN policy during training with no regularization.

the whole trajectory, in both actor and critic. To have the uniform dimensionality required by the MLP-based actor, we handle the actions of the agents not selected by the DAG as the parents by padding dummy vectors of zeros.

7.4.1. COORDINATION GAME

We run the experiments in the Coordination Game with $N = 2, 5$ under full observability, and no regularization (i.e., $\alpha = 0$), plotted in Figure 4(top). Remarkably, the result in $N = 2$ shows that our context-aware DAG learning outperforms the Fully-correlated. One possible explanation is that the dynamic graph leads to sufficient exploration of the augmented state defined in 5.6, and thus results in better performance. The context-aware DAG topology performs similarly to Fully-correlated in $N = 5$, and both outperform Uncorrelated. As shown in Figure 4(bottom), the density of the unregularized learned context-aware DAG is increasing in both $N = 2$ and $N = 5$ scenarios, from 50% to 55% and 50% to 53%, respectively.

7.4.2. COORDINATION GAME

We run the experiments in the Coordination Game with $N = 2, 5$ under full observability, and no regularization (i.e., $\alpha = 0$), plotted in Figure 4(top). Remarkably, the result in $N = 2$ shows that our context-aware DAG learning outperforms the Fully-correlated. One possible explanation is that the dynamic graph leads to sufficient exploration of the augmented state defined in 5.6, and thus results in better performance. The context-aware DAG topology performs similarly to Fully-correlated in $N = 5$, and both outperform Uncorrelated. As shown in Figure 4(bottom), the density of the unregularized learned context-aware DAG is increasing

in both $N = 2$ and $N = 5$ scenarios, from 50% to 55% and 50% to 53%, respectively.

7.4.3. ALOHA

We run the experiments in the Aloha with $N = 10$ under partial observability (each agent observes the backlog of its own messages), and no sparsity regularization (i.e., $\alpha = 0$). The results in Figure 4(top) show that our context-aware DAG learning performs comparably to Fully-correlated, and both outperform Uncorrelated. Note that the initial policy at timestep 0 with a random initialization will generate collisions resulting in large negative rewards. The policy will soon learn to avoid collisions, and we only show the performance when the policy can generate positive rewards. As shown in Figure 4(bottom), the density of the unregularized learned context-aware DAG is also increasing from 50% to 62% which is larger than the ones learned in the Coordination Game. This suggests that the action dependencies in Aloha may be more important than the ones in the Coordination Game.

Analysis: Learned DAG topologies. In the timestep shown in Figure 5, each agent has a backlog of one message to send. For Context-Aware-Correlated, guided by the learned
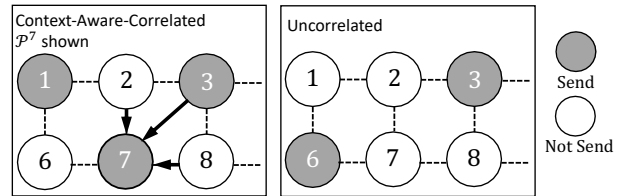

Figure 5. Learned DAG topology in Aloha. Only $\mathcal{P}^7$ is shown.

topology, agent 7 obtains the extra parent action (and observation) dependencies from nearby agents 2 and agent 8 , which do not send, and far-away agent 3 which sends but causes no collision. Thus, agent 7 is therefore more confident to send its message. For Uncorrelated, agents need to be more careful to avoid collisions. Both agents 2 and 7 choose not to send in this case to make it safer for agent 3 and 6 to send. This results in one less message sent for the shown agents.

7.4.4. SMAC

We run the experiments in the *Super Hard* SMAC scenario 6h_vs_8z and MMM2 under partial observability with no last actions stored, plotted in Figure 4(top). 6h_vs_8z and MMM2 are noisier than the Coordination Game and Aloha, and we find no benefit of using permutation matrix Π to change the topological ordering. Therefore, we use a fixed topological ordering where Π is the identity matrix. The action dependencies in both scenarios are crucial, as we can see in Figure 4(bottom) that the unregularized context-aware graph degenerates to an almost Full-Dependency graph in 6h_vs_8z and densely correlated graph with around 70% DAG density in MMM2, respectively. Therefore, we regularize it to control the DAG density with an annealing strategy, which gradually decreases threshold η and increases regularization weight α. Specifically, in the first $a\%$ training steps, the sparsity threshold η is set to 1, which encourages the agents to learn that the action dependencies are useful. Then, from $a\%$ total training steps to $b\%$ total training steps, we decrease sparsity threshold η from 1 to 0 uniformly in l_η times. From $b\%$ total training steps to $c\%$ total training steps, we uniformly increase in l_α times the regularization weight α from 0.1 to 1 in 6h_vs_8z and 0.05 to 0.5 in MMM2.

As shown in Figure 6, from 0% to a% total training steps, the performance is similar to the Fully-correlated baseline in both scenarios, with DAG density quickly becoming close to 1. From a% to b%, as we decrease sparsity threshold η, the performance fluctuates but still be much better than Uncorrelated. From b% to c%, we increase the regularization weight α. For 6h_vs_8z, the performance decreases quickly close to Uncorrelated, but then recovers quickly to be better than Uncorrelated. For MMM2, the transition is more smooth and the performance consistently beat the Uncorrelated baseline. This multi-phase regularization strategy results in purely uncorrelated policies as shown in Figure 4(bottom), but achieves better performance than Uncorrelated, which is trained with purely uncorrelated policies during the whole training phase as shown in Figure 6.

Analysis: Visibility. The dependency on the allies which are not visible is meaningless. Since action dependencies in both 6h_vs_8z and MMM2 are important, one simple strategy that an agent can learn to maintain a good perfor-
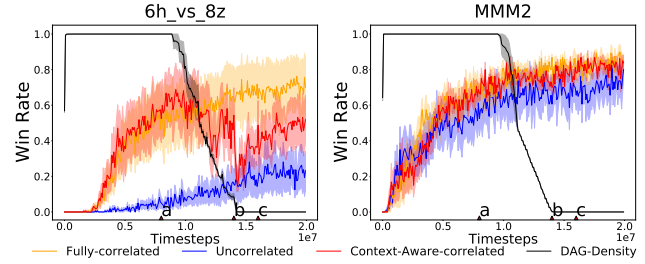

Figure 6. The performance of the sparsity regularized context-aware-correlated (with density annealing) in 6h_vs_8z and MMM2, against the Fully-correlated and Uncorrelated baselines. The black lines show the Changes in DAG Density of sparsity regularized context-aware-correlated (with density annealing) during training.

mance while decreasing the DAG density is to only output dependency on the visible agents. This is indeed the case, with the percentage of the visible agents that an agent wants to depend on almost consistently increasing in Figure 7.
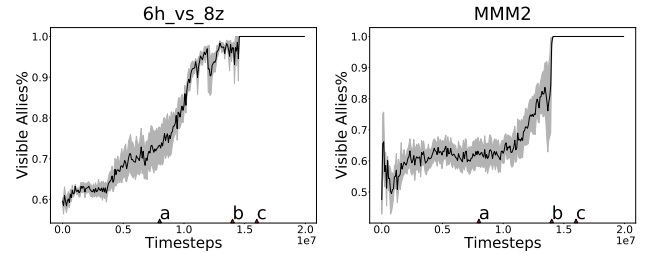

Figure 7. Visibility of allies during training (context-aware-correlated with density annealing) in 6h_vs_8z and MMM2.

Analysis: Average health. As shown in Figure 8, for 6h_vs_8z, agents 5 and 6 tend to depend on the actions of agents with a relatively low health bar, while agent 2 tends to depend on the actions of agents with relatively high health. This may be due to that we fixed the topological ordering of $(1, \cdots, 6)$, so agents 5 and 6 can potentially have more dependencies, and it learns to depend on the actions of agents with low health. On the other hand, agent 2 can only depend on the action of agent 1 which may not always have low health. For MMM2, health does not differentiate agents' selections of parent actions until in the middle of a% to b%, where the increase of the regularization causes all agents to depend on actions of agents with relatively high health, with agent 7 to the extreme.

Analysis: Average distance. As shown in Figure 9, in both scenarios, agent 6 tends to depend on the actions of agents in relatively long distances. In 6h_vs_8z, agent 2 tends to depend on the actions of agents with relatively short distances, while distance is relatively irrelevant for agents except agent 6 for the selection of parent actions. This also may be due to that we fixed the topological ordering of agent $(1, \cdots, N)$. Agent 6 can potentially have more
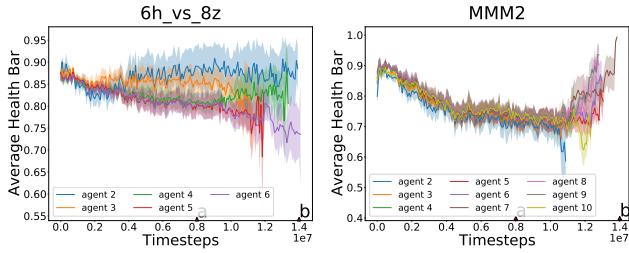
Figure 8. Average health during training (context-aware-correlated with sparsity annealing) in 6h_vs_8z and MMM2.

Figure 10. Emergence of multi-modality for BN policy (annealing) in SMAC.

dependencies, so it learns to depend on the actions of agents that are far away. On the other hand, agent 2 can only depend on the action of agent 1 which may be nearby sometimes. For MMM2, agent 8 consistently depends on the actions of agents from relatively longer distances, whereas agent 5 behaves the opposite.
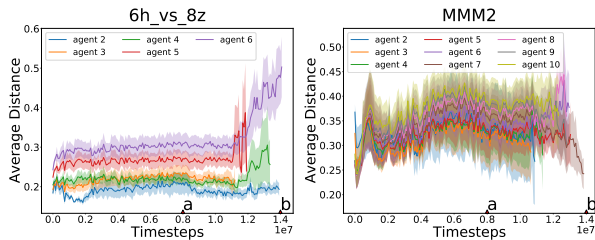

Figure 9. Average parent distance during training (context-aware-correlated with sparsity annealing) in 6h_vs_8z and MMM2.

Analysis: Emergence of multi-modality for BN policy
Previous works (Baker et al., 2019; Lowe et al., 2019; Tang et al., 2021) show that the emergence of diverse behaviors is prevalent in many multi-agent reinforcement learning problems. A recent work (Fu et al., 2022) shows the benefit of learning a multi-modal policy. Here we analyze the emergence of multi-modality for BN policy learned with the annealing strategy. To quantify multi-modality, we measure the KL Divergence between the distribution of the BN policy and the distribution of the same BN policy with empty DAG. The result in Figure 10 shows that in 6h_vs_8z, agent 6 with most possible parent actions has the largest multi-modality, whereas agent 2 with least possible parent actions except agent 1 has the smallest multi-modality. In MMM2, agent 8 emerges with the largest multi-modality, whereas agent 2 with least possible parent actions except agent 1 has the smallest multi-modality. It is also shown in both scenarios that increasing DAG density regularization also decreases multi-modality, where the purely decentralized one has zero multi-modality.
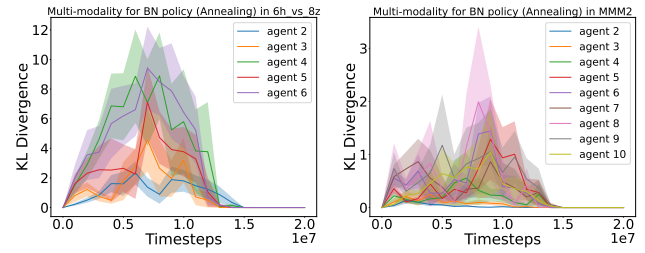
8. Conclusion

In this paper, we have motivated action correlations for co-operative MARL and proposed the notion of BN joint policy to introduce correlations. We have then derived the BN policy gradient formula and proved the convergence to Nash policy asymptotically under the tabular softmax BN policy parameterization. Further, we have proposed a practical algorithm to adapt any multi-agent actor-critic method to realize the BN joint policy and empirically demonstrated the benefits of the proposed method.

Acknowledgments

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

Böhmer, W., Kurin, V., and Whiteson, S. Deep coordination graphs. In *International Conference on Machine Learning*, pp. 980–991. PMLR, 2020.

Boutilier, C. Sequential optimality and coordination in multiagent systems. In *International Joint Conference on Artificial Intelligence*, 1999.

Callaway, D. S. and Hiskens, I. A. Achieving controllability of electric loads. *Proceedings of the IEEE*, 99(1):184–199, 2010.

Charpentier, B., Kibler, S., and Günnemann, S. Differentiable dag sampling. *arXiv preprint arXiv:2203.08509*, 2022.

Chen, D., Zhang, Q., and Doan, T. T. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games. *arXiv preprint arXiv:2206.07642*, 2022.

Chu, T., Wang, J., Codecà, L., and Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.

Corke, P., Peterson, R., and Rus, D. Networked robots: Flying robot navigation using a sensor net. In *Robotics research. The eleventh international symposium*, pp. 234–243. Springer, 2005.

Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4414–4425. PMLR, 2022.

Fu, W., Yu, C., Xu, Z., Yang, J., and Wu, Y. Revisiting some common practices in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2206.07505*, 2022.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. On the pitfalls of measuring emergent communication. *arXiv preprint arXiv:1903.05168*, 2019.

Mena, G., Belanger, D., Linderman, S., and Snoek, J. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

Peshkin, L., Kim, K.-E., Meuleau, N., and Kaelbling, L. P. Learning to cooperate via policy search. *arXiv preprint cs/0105032*, 2001.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International conference on machine learning*, pp. 4295–4304. PMLR, 2018.

Rogers, A., Farinelli, A., Stranders, R., and Jennings, N. R. Bounded approximate decentralised coordination via the max-sum algorithm. *Artif. Intell.*, 175:730–759, 2009.

Roughgarden, T. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015.

Ruan, J., Du, Y., Xiong, X., Xing, D., Li, X., Meng, L., Zhang, H., Wang, J., and Xu, B. Gcs: Graph-based coordination strategy for multi-agent reinforcement learning. *arXiv preprint arXiv:2201.06257*, 2022.

Samvelyan, M., Rashid, T., De Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J., and Whiteson, S. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

Tang, Z., Yu, C., Chen, B., Xu, H., Wang, X., Fang, F., Du, S., Wang, Y., and Wu, Y. Discovering diverse multi-agent strategic behavior via reward randomization. *arXiv preprint arXiv:2103.04564*, 2021.

Wang, T., Zeng, L., Dong, W., Yang, Q., Yu, Y., and Zhang, C. Context-aware sparse deep coordination graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=wQfgfb8VKTn.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games, 2021.

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. On the effect of log-barrier regularization in decentralized softmax gradient play in multiagent systems. *arXiv preprint arXiv:2202.00872*, 2022.

A. Proof of Theorem 5.5 and Corollary 5.8

Lemma A.1.

$$|Q^{\pi_{\tilde\theta}}(s,a) - Q^{\pi_\theta}(s,a)| \le \frac{r_{\max} - r_{\min}}{(1-\gamma)^2} \max_s \|\pi_{\tilde\theta_s} - \pi_{\theta_s}\|_1 \tag{4}$$

$$|\overline{A}^{\pi_{\tilde\theta},i}(s,a) - \overline{A}^{\pi_\theta,i}(s,a)| \le \frac{2(r_{\max} - r_{\min})}{(1-\gamma)^2} \max_s \|\pi_{\tilde\theta_s} - \pi_{\theta_s}\|_1 \tag{5}$$

where $\overline{A}^{\pi_\theta,i}(s,a) = Q^{\pi_\theta}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,\bar a^i, a^{-i})$.

Proof. Equation (4) is proved in lemma 32 in (Zhang et al., 2022).
For Equation (5),

$$|A^{\pi_{\tilde\theta},i}(s,a) - A^{\pi_\theta,i}(s,a)| = |Q^{\pi_{\tilde\theta}}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\tilde\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_{\tilde\theta}}(s,\bar a^i, a^{-i}) - (Q^{\pi_\theta}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,\bar a^i, a^{-i}))|$$

$$\le |Q^{\pi_{\tilde\theta}}(s,a) - Q^{\pi_\theta}(s,a)| + |\mathbb{E}_{\bar a^i \sim \pi^i_{\tilde\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_{\tilde\theta}}(s,\bar a^i, a^{-i}) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,\bar a^i, a^{-i})|$$

$$\le |Q^{\pi_{\tilde\theta}}(s,a) - Q^{\pi_\theta}(s,a)| + \max_a |Q^{\pi_{\tilde\theta}}(s,a) - Q^{\pi_\theta}(s,a)|$$

By Equation (4),

$$\le \frac{r_{\max} - r_{\min}}{(1-\gamma)^2} \max_s \|\pi_{\tilde\theta_s} - \pi_{\theta_s}\|_1 + \frac{r_{\max} - r_{\min}}{(1-\gamma)^2} \max_s \|\pi_{\tilde\theta_s} - \pi_{\theta_s}\|_1 = \frac{2(r_{\max} - r_{\min})}{(1-\gamma)^2} \max_s \|\pi_{\tilde\theta_s} - \pi_{\theta_s}\|_1$$

$\square$

Lemma A.2. $\|\nabla_{\theta^i} V(\tilde\theta) - \nabla_{\theta^i} V(\theta)\|_1 \le \frac{8(r_{\max} - r_{\min})}{(1-\gamma)^3} \sum_{i=1}^N \|\tilde\theta^i - \theta^i\|_2$

Proof.

$$\|\nabla_{\theta^i} V(\tilde\theta) - \nabla_{\theta^i} V(\theta)\|_1$$

$$= \frac{1}{1-\gamma} \sum_{s,a^{\mathcal{P}^i},a^i} |d^{\pi_{\tilde\theta}}_\mu(s,a^{\mathcal{P}^i})\pi^i_{\tilde\theta^i}(a^i|s,a^{\mathcal{P}^i})A^{\pi_{\tilde\theta},i}(s,a^{\mathcal{P}^i},a^i) - d^{\pi_\theta}_\mu(s,a^{\mathcal{P}^i})\pi^i_{\theta^i}(a^i|s,a^{\mathcal{P}^i})A^{\pi_\theta,i}(s,a^{\mathcal{P}^i},a^i)|$$

$$= \frac{1}{1-\gamma} \sum_{s,a^{\mathcal{P}^i},a^i} |d^{\pi_{\tilde\theta}}_\mu(s,a^{\mathcal{P}^i})\pi^i_{\tilde\theta^i}(a^i|s,a^{\mathcal{P}^i})*$$

$$\sum_{a^{-\mathcal{P}^i_+}} \pi_{\tilde\theta}(a^{-\mathcal{P}^i_+}|s,a^{\mathcal{P}^i},a^i)\big(Q^{\pi_{\tilde\theta}}(s,a^{\mathcal{P}^i},a^i,a^{-\mathcal{P}^i_+}) - \mathbb{E}_{\bar a^i \sim \pi^i_{\tilde\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_{\tilde\theta}}(s,a^{\mathcal{P}^i},\bar a^i, a^{-\mathcal{P}^i_+})\big) -$$

$$d^{\pi_\theta}_\mu(s,a^{\mathcal{P}^i})\pi^i_{\theta^i}(a^i|s,a^{\mathcal{P}^i}) \sum_{a^{-\mathcal{P}^i_+}} \pi_\theta(a^{-\mathcal{P}^i_+}|s,a^{\mathcal{P}^i},a^i)\big(Q^{\pi_\theta}(s,a^{\mathcal{P}^i},a^i,a^{-\mathcal{P}^i_+}) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,a^{\mathcal{P}^i},\bar a^i, a^{-\mathcal{P}^i_+})\big)|$$

$$\le \frac{1}{1-\gamma} \sum_{s,a} |d^{\pi_{\tilde\theta}}_\mu(s)\pi_{\tilde\theta}(a|s)\big(Q^{\pi_{\tilde\theta}}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\tilde\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_{\tilde\theta}}(s,\bar a^i, a^{-i})\big) -$$

$$d^{\pi_\theta}_\mu(s)\pi_\theta(a|s)\big(Q^{\pi_\theta}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,\bar a^i, a^{-i})\big)|$$

Denote $\overline{A}^{\pi_\theta,i}(s,a) = Q^{\pi_\theta}(s,a) - \mathbb{E}_{\bar a^i \sim \pi^i_{\theta^i}(\cdot|s,a^{\mathcal{P}^i})} Q^{\pi_\theta}(s,\bar a^i, a^{-i})$,

$$= \frac{1}{1-\gamma} \sum_{s,a} |d^{\pi_{\tilde\theta}}_\mu(s)\pi_{\tilde\theta}(a|s)\overline{A}^{\pi_{\tilde\theta},i}(s,a) - d^{\pi_\theta}_\mu(s)\pi_\theta(a|s)\overline{A}^{\pi_\theta,i}(s,a)|$$

$$\le \frac{1}{1-\gamma}\left(\sum_{s,a} |d^{\pi_{\tilde\theta}}_\mu(s)\pi_{\tilde\theta}(a|s) - d^{\pi_{\tilde\theta}}_\mu(s)\pi_{\tilde\theta}(a|s)||\overline{A}^{\pi_{\tilde\theta},i}(s,a)| + \sum_{s,a} d^{\pi_\theta}_\mu(s)\pi_\theta(a|s)|\overline{A}^{\pi_{\tilde\theta},i}(s,a) - \overline{A}^{\pi_\theta,i}(s,a)|\right)$$

Since $\overline{A}^{\pi_{\theta},i}(s,a) \leq \frac{2(r_{\max}-r_{\min})}{1-\gamma}$,

$$\leq \frac{1}{1-\gamma}\left(\sum_{s,a}\frac{2(r_{\max}-r_{\min})}{1-\gamma}|d_{\mu}^{\pi_{\tilde{\theta}}}(s)\pi_{\tilde{\theta}}(a|s) - d_{\mu}^{\pi_{\tilde{\theta}}}(s)\pi_{\tilde{\theta}}(a|s) + \max_{s,a}\left|\overline{A}^{\pi_{\tilde{\theta}},i}(s,a) - \overline{A}^{\pi_{\theta},i}(s,a)\right|\right)$$

By Equation (5),

$$\leq \frac{1}{1-\gamma}\left(\sum_{s,a}\frac{2(r_{\max}-r_{\min})}{1-\gamma}|d_{\mu}^{\pi_{\tilde{\theta}}}(s)\pi_{\tilde{\theta}}(a|s) - d_{\mu}^{\pi_{\tilde{\theta}}}(s)\pi_{\tilde{\theta}}(a|s) + \frac{2(r_{\max}-r_{\min})}{(1-\gamma)^2}\max_{s}\|\pi_{\tilde{\theta}_s} - \pi_{\theta_s}\|_1\right)$$

By corollary (35) in (Zhang et al., 2022),

$$\leq \frac{1}{1-\gamma}\left(\frac{2(r_{\max}-r_{\min})}{(1-\gamma)^2}\max_{s}\|\pi_{\tilde{\theta}_s} - \pi_{\theta_s}\|_1 + \frac{2(r_{\max}-r_{\min})}{(1-\gamma)^2}\max_{s}\|\pi_{\tilde{\theta}_s} - \pi_{\theta_s}\|_1\right)$$

$$= \frac{4(r_{\max}-r_{\min})}{(1-\gamma)^3}\max_{s}\|\pi_{\tilde{\theta}_s} - \pi_{\theta_s}\|_1$$

$$\leq \frac{4(r_{\max}-r_{\min})}{(1-\gamma)^3}\max_{s}\sum_{i,a^{\mathcal{P}^i}}\|\pi_{\tilde{\theta}^i_{s,a^{\mathcal{P}^i}}} - \pi_{\theta^i_{s,a^{\mathcal{P}^i}}}\|_1$$

By corollary (37) in (Zhang et al., 2022),

$$\leq \frac{8(r_{\max}-r_{\min})}{(1-\gamma)^3}\max_{s}\sum_{i,a^{\mathcal{P}^i}}\|\tilde{\theta}^i_{s,a^{\mathcal{P}^i}} - \theta^i_{s,a^{\mathcal{P}^i}}\|_2$$

$$\leq \frac{8(r_{\max}-r_{\min})}{(1-\gamma)^3}\max_{s}\sum_{i=1}^{N}\|\tilde{\theta}^i_{s} - \theta^i_{s}\|_2$$

$$\leq \frac{8(r_{\max}-r_{\min})}{(1-\gamma)^3}\sum_{i=1}^{N}\|\tilde{\theta}^i - \theta^i\|_2$$

$$\square$$

Lemma A.3 (Smoothness of V under tabular Baysian softmax).

$$\|\nabla_{\theta}V(\tilde{\theta}) - \nabla_{\theta}V(\theta)\|_2 \leq \frac{8N(r_{\max}-r_{\min})}{(1-\gamma)^3}\|\tilde{\theta} - \theta\|_2$$

Proof.

$$\|\nabla_{\theta}V(\tilde{\theta}) - \nabla_{\theta}V(\theta)\|_2^2 = \sum_{i=1}^{N}\|\nabla_{\theta^i}V(\tilde{\theta}) - \nabla_{\theta^i}V(\theta)\|_2^2$$

$$\leq \sum_{i=1}^{N}\|\nabla_{\theta^i}V(\tilde{\theta}) - \nabla_{\theta^i}V(\theta)\|_1^2$$

By lemma (A.2),

$$\leq \sum_{i=1}^{N}\left(\frac{8(r_{\max}-r_{\min})}{(1-\gamma)^3}\sum_{j=1}^{N}\|\tilde{\theta}^j - \theta^j\|_2\right)^2$$

$$= \frac{64N(r_{\max}-r_{\min})^2}{(1-\gamma)^6}\left(\sum_{i=1}^{N}\|\tilde{\theta}^i - \theta^i\|_2\right)^2$$

$$\leq \frac{64N^2(r_{\max} - r_{\min})^2}{(1-\gamma)^6} \sum_{i=1}^{N} \|\tilde{\theta}^i - \theta^i\|_2^2$$

$$= \frac{64N^2(r_{\max} - r_{\min})^2}{(1-\gamma)^6} \|\tilde{\theta} - \theta\|_2^2$$

Therefore,

$$\|\nabla_\theta V(\tilde{\theta}) - \nabla_\theta V(\theta)\|_2 \leq \frac{8N(r_{\max} - r_{\min})}{(1-\gamma)^3} \|\tilde{\theta} - \theta\|_2$$

$\square$

Lemma A.4. *For a Baysian policy defined by $\mathcal{G}$, $\forall s, a^{\mathcal{P}^i}, a^i$,*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta^i}\bigg|_{s,a^{\mathcal{P}^i},a^i} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i}) A^{\pi_\theta,i}(s, a^{\mathcal{P}^i}, a^i)$$

,where $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i})$ $=$ $d_\mu^{\pi_\theta}(s) \sum_{a^{-\mathcal{P}^i}} \pi_\theta(a^{-\mathcal{P}^i}, a^{\mathcal{P}^i}|s)$, $A^{\pi_\theta,i}(s, a^{\mathcal{P}^i}, a^i)$ $=$ $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i}, a^i)$ $-$ $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i})$, $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i}, a^i)$ $=$ $\mathbb{E}_{\bar{a}^{-\mathcal{P}_+^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)} \left[Q^{\pi_\theta}(s, a^{\mathcal{P}^i}, a^i, \bar{a}^{-\mathcal{P}_+^i}) \right]$, $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i})$ $=$ $\mathbb{E}_{\bar{a}^{-\mathcal{P}^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i})} \left[Q^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^{-\mathcal{P}^i}) \right]$.

Note that the policy gradient formula in Lemma A.4 is the same as the formula in Lemma 5.1, but with different notations. Here we uses $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i}, a^i)$ instead of $Q^{\pi_\theta}(s, a^{\mathcal{P}_+^i})$ to highlight the local action a^i, and it is only used in the proof. They define the same quantity. We also uses $Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i})$ instead of $Q^{\pi_\theta}(s, a^{\mathcal{P}^i})$ for the proof.

Proof. For agent i,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta^i}\bigg|_{s,a^{\mathcal{P}^i},a^i} = \frac{1}{1-\gamma} \mathbb{E}_{\bar{s} \sim d_\mu^{\pi_\theta}} \mathbb{E}_{\bar{a} \sim \pi_\theta(\cdot|\bar{s})} \left[A^{\pi_\theta}(\bar{s}, \bar{a}) \frac{\partial \log \pi_{\theta^i}^i(\bar{a}^i|s, a^{\mathcal{P}^i})}{\partial \theta^i}\bigg|_{s,a^{\mathcal{P}^i},a^i} \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{\bar{s} \sim d_\mu^{\pi_\theta}} \mathbb{E}_{\bar{a} \sim \pi_\theta(\cdot|\bar{s})} \left[A^{\pi_\theta}(\bar{s}, \bar{a}) \mathbb{1}[\bar{s} = s] \mathbb{1}[\bar{a}^{\mathcal{P}^i} = a^{\mathcal{P}^i}] (\mathbb{1}[\bar{a}^i = a^i] - \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i})) \right]$$

$$= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \pi_\theta^{\mathcal{P}^i}(a^{\mathcal{P}^i}|s) \mathbb{E}_{\bar{a}^i \sim \pi_{\theta^i}^i(\cdot|s,a^{\mathcal{P}^i}), \bar{a}^{-\mathcal{P}_+^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)}$$

$$\left[A^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^i, \bar{a}^{-\mathcal{P}_+^i})(\mathbb{1}[\bar{a}^i = a^i] - \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i})) \right]$$

$$= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) \mathbb{E}_{\bar{a}^i \sim \pi_{\theta^i}^i(\cdot|s,a^{\mathcal{P}^i}), \bar{a}^{-\mathcal{P}_+^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)}$$

$$\left[A^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^i, \bar{a}^{-\mathcal{P}_+^i})(\mathbb{1}[\bar{a}^i = a^i] - \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i})) \right]$$

$$= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i})$$

$$\left(\mathbb{E}_{\bar{a}^i \sim \pi_{\theta^i}^i(\cdot|s,a^{\mathcal{P}^i}), \bar{a}^{-\mathcal{P}_+^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)} \left[A^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^i, \bar{a}^{-\mathcal{P}_+^i}) \mathbb{1}[\bar{a}^i = a^i] \right] \right.$$

$$\left. - \mathbb{E}_{\bar{a}^{-\mathcal{P}^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i})} \left[A^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^{-\mathcal{P}^i}) \pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i}) \right] \right)$$

$$= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i})$$

$$\left(\pi_{\theta^i}^i(a^i|s, a^{\mathcal{P}^i}) \mathbb{E}_{\bar{a}^{-\mathcal{P}_+^i} \sim \pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)} \left[A^{\pi_\theta}(s, a^{\mathcal{P}^i}, a^i, \bar{a}^{-\mathcal{P}_+^i}) \right] \right.$$

13

$$-\pi_{\theta^i}^i(a^i|s,a^{\mathcal{P}^i})\,\mathbb{E}_{\bar{a}^{-\mathcal{P}^i}\sim\pi_\theta(\cdot|s,a^{\mathcal{P}^i})}\left[A^{\pi_\theta}(s,a^{\mathcal{P}^i},\bar{a}^{-\mathcal{P}^i})\right]\Bigg)$$

$$=\frac{1}{1-\gamma}d_\mu^{\pi_\theta}(s,a^{\mathcal{P}^i})\pi_{\theta^i}^i(a^i|s,a^{\mathcal{P}^i})$$

$$\left(\mathbb{E}_{\bar{a}^{-\mathcal{P}_+^i}\sim\pi_\theta(\cdot|s,a^{\mathcal{P}^i},a^i)}\left[Q^{\pi_\theta}(s,a^{\mathcal{P}^i},a^i,\bar{a}^{-\mathcal{P}_+^i})\right]-\mathbb{E}_{\bar{a}^{-\mathcal{P}^i}\sim\pi_\theta(\cdot|s,a^{\mathcal{P}^i})}\left[Q^{\pi_\theta}(s,a^{\mathcal{P}^i},\bar{a}^{-\mathcal{P}^i})\right]\right)$$

$$=\frac{1}{1-\gamma}d_\mu^{\pi_\theta}(s,a^{\mathcal{P}^i})\pi_{\theta^i}^i(a^i|s,a^{\mathcal{P}^i})\left[Q^{\pi_\theta,i}(s,a^{\mathcal{P}^i},a^i)-Q^{\pi_\theta,i}(s,a^{\mathcal{P}^i})\right]$$

$$=\frac{1}{1-\gamma}d_\mu^{\pi_\theta}(s,a^{\mathcal{P}^i})\pi_{\theta^i}^i(a^i|s,a^{\mathcal{P}^i})A^{\pi_\theta,i}(s,a^{\mathcal{P}^i},a^i)$$

$\square$

Lemma A.5. *For all agents i with a round of update*

$$\theta^{t+1,i}=\theta^{t,i}+\eta\nabla V_{\theta^{t,i}}^i(\mu)$$

with learning rates $\eta\leq\frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, we have

$$V^{t+1}(s)\geq V^t(s);Q^{t+1}(s,a)\geq Q^t(s,a).$$

Proof. Since $V(\theta)$ is $\frac{8N(r_{\max}-r_{\min})}{(1-\gamma)^3}$-smooth, we know that with learning rate $\eta\leq\frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, V is monotonic increasing and therefore Q is also monotonic increasing. $\square$

Lemma A.6. *For all states s and actions a, there exists values $V^\infty(s)$ and $Q^\infty(s,a)$ such that as $t\to\infty,V^t(s)\to V^\infty(s),Q^t(s,a)\to Q^\infty(s,a)$. For all agents i, states s, parent actions $a^{\mathcal{P}^i}$, local action a^i, there exists values $Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)$ and $Q^{\infty,i}(s,a^{\mathcal{P}^i})$ such that as $t\to\infty,Q^{t,i}(s,a^{\mathcal{P}^i},a^i)\to Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i),Q^{t,i}(s,a^{\mathcal{P}^i})\to Q^{\infty,i}(s,a^{\mathcal{P}^i})$. Define*

$$\Delta^i=\min_{\{s,a^{\mathcal{P}^i},a^i|A^{\infty,i}(s,a^{\mathcal{P}^i},a^i)\neq0\}}|A^{\infty,i}(s,a^{\mathcal{P}^i},a^i)|.$$

$$\Delta=\min_i\Delta^i.$$

Further, there exists a T_0 such that for all $t>T_0$, agents i, states s, parent actions $a^{\mathcal{P}^i}$, local action a^i,

$$Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)-\frac{\Delta}{4}\leq Q^{t,i}(s,a^{\mathcal{P}^i},a^i)\leq Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)+\frac{\Delta}{4}$$

Proof. $\{V^t(s)\}$ is bounded and monotonically increasing, therefore $V^t(s)\to V^\infty(s)$. Similarly, we know $Q^t(s,a)\to Q^\infty(s,a)$. Since the Bayesian policy is assumed to converge, we have that both $\{Q^{t,i}(s,a^{\mathcal{P}^i},a^i)\}$ and $\{Q^{t,i}(s,a^{\mathcal{P}^i})\}$ are convergent. For all agents i, states s, parent actions $a^{\mathcal{P}^i}$, categorize the local action a^i into three groups:

$$I_0^{s,a^{\mathcal{P}^i},i}=\left\{a^i|Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)=Q^{\infty,i}(s,a^{\mathcal{P}^i})\right\}$$

$$I_+^{s,a^{\mathcal{P}^i},i}=\left\{a^i|Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)>Q^{\infty,i}(s,a^{\mathcal{P}^i})\right\}$$

$$I_-^{s,a^{\mathcal{P}^i},i}=\left\{a^i|Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)<Q^{\infty,i}(s,a^{\mathcal{P}^i})\right\}$$

Since $Q^{t,i}(s,a^{\mathcal{P}^i},a^i)\to Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)$ as $t\to\infty$, there exists a T_0 such that for all $t>T_0$, agents i, states s, parent actions $a^{\mathcal{P}^i}$,

$$Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)-\frac{\Delta}{4}\leq Q^{t,i}(s,a^{\mathcal{P}^i},a^i)\leq Q^{\infty,i}(s,a^{\mathcal{P}^i},a^i)+\frac{\Delta}{4}$$

$\square$

Lemma A.7. $\exists T_1$ *such that* $\forall t > T_1, i, s, a^{\mathcal{P}^i}, a^i$, *we have*

$$A^{t,i}(s, a^{\mathcal{P}^i}, a^i) < -\frac{\Delta}{4} \text{ for } a^i \in I_-^{s,a^{\mathcal{P}^i},i}; A^{t,i}(s, a^{\mathcal{P}^i}, a^i) > \frac{\Delta}{4} \text{ for } a^i \in I_+^{s,a^{\mathcal{P}^i},i}$$

Proof. Since $\forall s, i, a^{\mathcal{P}^i}, Q^{t,i}(s, a^{\mathcal{P}^i}) \to Q^{\infty,i}(s, a^{\mathcal{P}^i})$, we have that there exists $T_1 > T_0$ such that for all $t > T_1$,

$$Q^{\infty,i}(s, a^{\mathcal{P}^i}) - \frac{\Delta}{4} \le Q^{t,i}(s, a^{\mathcal{P}^i}) \le Q^{\infty,i}(s, a^{\mathcal{P}^i}) + \frac{\Delta}{4}$$

For $a^i \in I_-^{s,a^{\mathcal{P}^i},i}$,

$$\begin{aligned}
A^{t,i}(s, a^{\mathcal{P}^i}, a^i) &= Q^{t,i}(s, a^{\mathcal{P}^i}, a^i) - Q^{t,i}(s, a^{\mathcal{P}^i}) \\
&\le Q^{\infty,i}(s, a^{\mathcal{P}^i}, a^i) + \frac{\Delta}{4} - Q^{t,i}(s, a^{\mathcal{P}^i}) \\
&\le Q^{\infty,i}(s, a^{\mathcal{P}^i}, a^i) + \frac{\Delta}{4} - Q^{\infty,i}(s, a^{\mathcal{P}^i}) + \frac{\Delta}{4} \\
&\le -\Delta + \frac{\Delta}{4} + \frac{\Delta}{4} \\
&< -\frac{\Delta}{4}
\end{aligned} \tag{6}$$

For $a^i \in I_+^{s,a^{\mathcal{P}^i},i}$,

$$\begin{aligned}
A^{t,i}(s, a^{\mathcal{P}^i}, a^i) &= Q^{t,i}(s, a^{\mathcal{P}^i}, a^i) - Q^{t,i}(s, a^{\mathcal{P}^i}) \\
&\ge Q^{\infty,i}(s, a^{\mathcal{P}^i}, a^i) - \frac{\Delta}{4} - Q^{t,i}(s, a^{\mathcal{P}^i}) \\
&\ge Q^{\infty,i}(s, a^{\mathcal{P}^i}, a^i) - \frac{\Delta}{4} - Q^{\infty,i}(s, a^{\mathcal{P}^i}) - \frac{\Delta}{4} \\
&\ge \Delta - \frac{\Delta}{4} - \frac{\Delta}{4} \\
&> \frac{\Delta}{4}
\end{aligned} \tag{7}$$

$\square$

Lemma A.8. $\frac{\partial V^t(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i},a^i}} \to 0$ *as* $t \to \infty$ *for all agents* i, *states* s, *parent actions* $a^{\mathcal{P}^i}$, *local action* a^i. *This implies that* $\forall i, \forall a^{\mathcal{P}^i}$, *if* $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, *then* $\forall a^i \in I_-^{s,a^{\mathcal{P}^i},i} \cup I_+^{s,a^{\mathcal{P}^i},i}, \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 0$ *and that* $\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 1$.

Proof. Since $V^{\pi_\theta}(\mu)$ is $\frac{8N(r_{\max}-r_{\min})}{(1-\gamma)^3}$-smooth, we know that with learning rate $\eta < \frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}, \frac{\partial V^t(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i},a^i}} \to 0$ for all $i, s, a^{\mathcal{P}^i}, a^i$. From lemma A.4 we have

$$\frac{\partial V^t(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i},a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) A^{t,i}(s, a^{\mathcal{P}^i}, a^i)$$

Since from lemma A.7, we know that $|A^{\pi_\theta}(s, a^{\mathcal{P}^i}, a^i)| > \frac{\Delta}{4}$ for all $t > T_1$, for all $a^i \in I_-^{s,a^{\mathcal{P}^i},i} \cup I_+^{s,a^{\mathcal{P}^i},i}$, which together with the assumption that $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$ proves $\pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 0$. Then we also know for all $\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 1$. $\square$

From Lemma A.9 to Lemma A.15, we prove the properties under the condition that $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, so that with Lemma A.8, we know that $\pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 0$.

Lemma A.9. *For $t \geq T_1$, $\forall i, \forall a^{\mathcal{P}^i}$, if $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, then $\theta_{s,a^{\mathcal{P}^i},a^i}^i$ is strictly decreasing $\forall a^i \in I_{-}^{s,a^{\mathcal{P}^i},i}$ and $\theta_{s,a^{\mathcal{P}^i},a^i}^i$ is strictly increasing $\forall a^i \in I_{+}^{s,a^{\mathcal{P}^i},i}$.*

Proof. From A.4 we have

$$\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i})\pi^{t,i}(a^i|s, a^{\mathcal{P}^i})A^{t,i}(s, a^{\mathcal{P}^i}, a^i)$$

From lemma A.7, we know for all $t > T_1$, $a^i \in I_{-}^{s,a^{\mathcal{P}^i},i}$, $A^{t,i}(s, a^{\mathcal{P}^i}, a^i) < -\frac{\Delta}{4}$; For all $a^i \in I_{+}^{s,a^{\mathcal{P}^i},i}$, $A^{t,i}(s, a^{\mathcal{P}^i}, a^i) > \frac{\Delta}{4}$. This implies that after iteration T_1, $\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} < 0 \forall a^i \in I_{-}^{s,a^{\mathcal{P}^i},i}$; $\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} > 0 \forall a^i \in I_{+}^{s,a^{\mathcal{P}^i},i}$. $\longrightarrow$ After iteration T_1, $\theta_{s,a^{\mathcal{P}^i},a^i}^i$ is strictly decreasing $\forall a^i \in I_{-}^{s,a^{\mathcal{P}^i},i}$ and $\theta_{s,a^{\mathcal{P}^i},a^i}^i$ is strictly increasing $\forall a^i \in I_{+}^{s,a^{\mathcal{P}^i},i}$. $\qquad\square$

Lemma A.10. *For all $i, s, a^{\mathcal{P}^i}, a^i$, if $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$ and $I_{+}^{s,a^{\mathcal{P}^i},i} \neq \emptyset$, then we have:*

$$\max_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty, \quad min_{a^i \in \mathbb{A}^i}\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to -\infty$$

Proof. Since $I_{+}^{s,a^{\mathcal{P}^i},i} \neq \emptyset$, we have some action $a_+^i \in I_{+}^{s,a^{\mathcal{P}^i},i}$. From lemma 12, we know

$$\pi^{t,i}(a_+^i|s, a^{\mathcal{P}^i}) \to 0 \text{ as } t \to \infty$$

$$\longrightarrow \frac{\exp(\theta_{s,a^{\mathcal{P}^i},a_+^i}^{t,i})}{\sum_{a^i \in \mathbb{A}^i} \exp(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i})} \to 0 \text{ as } t \to \infty$$

From lemma A.9 we know $\theta_{s,a^{\mathcal{P}^i},a_+^i}^{t,i}$ is monotonically increasing, which implies

$$\sum_{a^i \in \mathbb{A}^i} \exp(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i}) \to \infty \text{ as } t \to \infty$$

From lemma A.8, we also know

$$\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}) \to 1$$

$$\longrightarrow \frac{\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \exp(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i})}{\sum_{a^i \in \mathbb{A}^i} \exp(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i})} \to 1$$

Since denominator does to ∞, we know

$$\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \exp(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i}) \to \infty$$

which implies

$$\max_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty$$

Note this also implies $\max_{a^i \in \mathbb{A}^i} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty$. The sum of the gradient is always zero: $\sum_{a^i \in \mathbb{A}^i} \frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s)\pi^{t,\mathcal{P}^i}(a^{\mathcal{P}^i}|s) \sum_{a^i \in \mathbb{A}^i} \pi^{t,i}(a^i|s, a^{\mathcal{P}^i})A^{t,i}(s, a^{\mathcal{P}^i}, a^i) = 0$. Thus, $\sum_{a^i \in \mathbb{A}^i} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} = \sum_{a^i \in \mathbb{A}^i} \theta_{s,a^{\mathcal{P}^i},a^i}^{0,i}$ which is a constant. Since $\max_{a^i \in \mathbb{A}^i} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty$, we know

$$\min_{a^i \in \mathbb{A}^i} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to -\infty$$

$\square$

Lemma A.11. *For some $s, i, a^{\mathcal{P}^i}$, suppose $a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}$. $\forall a \in I_0^{s,a^{\mathcal{P}^i},i}$, if $\exists t \geq T_1$ such that $\pi^{t,i}(a|s,a^{\mathcal{P}^i}) \leq \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})$, then $\forall \tau \geq t, \pi^{\tau,i}(a|s,a^{\mathcal{P}^i}) \leq \pi^{\tau,i}(a_+^i|s,a^{\mathcal{P}^i})$.*

Proof. Suppose $a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}, a \in I_0^{s,a^{\mathcal{P}^i},i}$, if $\pi^{t,i}(a|s,a^{\mathcal{P}^i}) \leq \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})$, then

$$\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s,a^{\mathcal{P}^i})\pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a}^i} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s,a^{\mathcal{P}^i})\pi^{t,i}(a^i|s,a^{\mathcal{P}^i})\Big[Q^{t,i}(s,a^{\mathcal{P}^i},a^i) - Q^{t,i}(s,a^{\mathcal{P}^i})\Big]$$

$$\leq \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s,a^{\mathcal{P}^i})\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\Big[Q^{t,i}(s,a^{\mathcal{P}^i},a_+^i) - Q^{t,i}(s,a^{\mathcal{P}^i})\Big] = \frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a_+^i}^i}$$

where the last step holds because $Q^{t,i}(s,a^{\mathcal{P}^i},a_+^i) \geq Q^{\infty,i}(s,a^{\mathcal{P}^i},a_+^i) - \frac{\Delta}{4} \geq Q^{\infty,i}(s,a^{\mathcal{P}^i},a) + \Delta - \frac{\Delta}{4} \geq Q^{t,i}(s,a^{\mathcal{P}^i},a) - \frac{\Delta}{4} + \Delta - \frac{\Delta}{4} > Q^{t,i}(s,a^{\mathcal{P}^i},a)$ for $t > T_0$.

We can then partition $I_0^{s,a^{\mathcal{P}^i},i}$ into $B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$ and $\bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$ as follows:

$$B_0^{s,a^{\mathcal{P}^i},i}(a_+^i) : \{a|a \in I_0^{s,a^{\mathcal{P}^i},i} \text{ and } \forall t \geq T_0, \pi^{t,i}(a_+^i|s,\mathcal{P}(i)) < \pi^{t,i}(a|s,\mathcal{P}(i))\}$$

$$\bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i) : I_0^{s,a^{\mathcal{P}^i},i} \setminus B_0^{s,\mathcal{P}(i),i}(a_+^i).$$

$\square$

Lemma A.12. *For some $s, i, a^{\mathcal{P}^i}$, if $\lim_{t\to\infty} d_\mu^{\pi^t}(s,a^{\mathcal{P}^i}) > 0$, then suppose $I_+^{s,a^{\mathcal{P}^i},i} \neq \emptyset$. $\forall a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}$, we have that $B_0^{s,a^{\mathcal{P}^i},i}(a_+^i) \neq \emptyset$ and that*

$$\sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i}) \to 1, \text{ as } t \to \infty.$$

This implies that:

$$\max_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty.$$

Proof. Let $a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}$. Consider any $\bar{a}^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$. Then by definition of $\bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$, there exists $t' > T_0$ such that $\pi^{t',i}(a_+^i|s,a^{\mathcal{P}^i}) \geq \pi^{t',i}(\bar{a}^i|s,a^{\mathcal{P}^i})$. From lemma A.11, we know $\forall \tau > t', \pi^{\tau,i}(a_+^i|s,a^{\mathcal{P}^i}) \geq \pi^{\tau,i}(\bar{a}^i|s,a^{\mathcal{P}^i})$. From lemma A.8, we know $\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i}) \to 0$ as $t \to \infty$, which implies

$$\pi^{t,i}(\bar{a}^i|s,a^{\mathcal{P}^i}) \to 0 \text{ as } t \to \infty.$$

Since $B_0^{s,a^{\mathcal{P}^i},i}(a_+^i) \cup \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i) = I_0^{s,i}$ and $\sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i}) \to 1$, we know

$$\sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i}) \to 1$$

$$B_0^{s,a^{\mathcal{P}^i},i}(a_+^i) \neq \emptyset$$

Using the same techniques in A.10, we know

$$\max_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} \to \infty$$

$\square$

17

Lemma A.13. *Consider any* $s, a^{\mathcal{P}^i}$, *where* $I_+^{s,a^{\mathcal{P}^i},i} \neq \emptyset$. *Then,* $\forall a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}, \exists T_{a^{\mathcal{P}^i},a_+^i}$ *such that* $\forall t > T_{a^{\mathcal{P}^i},a_+^i}, \forall a^i \in \bar{B}_0^{s,i}(a_+^i)$,

$$\pi^{t,i}(a_+^i|s, a^{\mathcal{P}^i})) > \pi^{t,i}(a^i|s, a^{\mathcal{P}^i}))$$

Proof. By the definition of $\bar{B}_0^{s,i}(a_+^i)$ and lemma A.11, $\forall a^i \in \bar{B}_0^{s,i}(a_+^i)$, there exists $t_{a^{\mathcal{P}^i},a^i} > T_0$ such that $\forall \tau > t_{a^{\mathcal{P}^i},a^i}$, $\pi^{\tau,i}(a_+^i|s, a^{\mathcal{P}^i}) > \pi^{\tau,i}(a^i|s, a^{\mathcal{P}^i})$. We can choose $T_{a^{\mathcal{P}^i},a_+^i} = \max_{a^{\mathcal{P}^i},a^i \in B_0^{s,i}(a_+^i)} t_{a^i}$. $\qquad\square$

Lemma A.14. $\forall i, a^{\mathcal{P}^i}, a^i$, *if* $\lim_{t\to\infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, *then we have* $\forall a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}$, $\theta_{s,a^{\mathcal{P}^i},a_+^i}^i$ *is lower bounded as* $t \to \infty$ *and* $\forall a_-^i \in I_-^{s,a^{\mathcal{P}^i},i}$, $\theta_{s,a^{\mathcal{P}^i},a_-^i}^i \to -\infty$ *as* $t \to \infty$.

Proof. From lemma A.9, we know that $\forall a_+^i \in I_+^{s,a^{\mathcal{P}^i},i}$, after T_1, $\theta_{s,a^{\mathcal{P}^i},a_+^i}^i$ is strictly increasing, and is therefore bounded from below.

For the second claim, we know from lemma A.9 that $\forall a_-^i \in I_-^{s,a^{\mathcal{P}^i},i}$, after T_1, $\theta_{s,a^{\mathcal{P}^i},a_-^i}^i$ is strictly decreasing. Then, by monotone convergence theorem, we know $\lim_{t\to\infty} \theta_{s,a^{\mathcal{P}^i},a_-^i}^i$ exists and is either $-\infty$ or some constant θ_0^i. We now prove by contraction that $\lim_{t\to\infty} \theta_{s,a^{\mathcal{P}^i},a_-^i}^i$ cannot be some constant θ_0^i. Suppose $\lim_{t\to\infty} \theta_{s,a^{\mathcal{P}^i},a_-^i}^i = \theta_0^i$. We immediately know that $\forall t \geq T_1, \theta_{s,a^{\mathcal{P}^i},a_-^i}^i > \theta_0^i$. By lemma A.10, we know $\exists a^i \in \mathbb{A}^i$ such that

$$\lim_{t\to\infty} \inf \theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} = -\infty \tag{8}$$

Let us consider some $\delta^i > 0$ such that $\theta_{s,a^{\mathcal{P}^i},a^i}^{T_1,i} \geq \theta_0^i - \delta^i$. Now for $t \geq T_1$, define $\tau^i(t)$ to be the largest iteration in $[T_1, t]$ such that $\theta_{s,a^{\mathcal{P}^i},a^i}^{\tau^i(t),i} \geq \theta_0^i - \delta^i$. Define $\mathcal{T}^{t,i}$ to be subsequence $\{t'\}$ of the interval $(\tau^i(t), t)$ such that $\theta_{s,a^{\mathcal{P}^i},a^i}^{t',i}$ decreases. Define

$$Z^{t,a^{\mathcal{P}^i},i} = \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial V^{t'}(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i}$$

where $Z^{t,a^{\mathcal{P}^i},i} = 0$ if $\mathcal{T}^{t,i} = \emptyset$.
For non-empty $\mathcal{T}^{t,i}$, we have:

$$Z^{t,a^{\mathcal{P}^i},i} = \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial V^{t'}(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} \leq \sum_{t'=\tau^i(t)+1}^{t-1} \frac{\partial V^{t'}(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} \leq \sum_{t'=\tau^i(t)}^{t-1} \frac{\partial V^{t'}(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} + \frac{1}{(1-\gamma)}(V_{\max} - V_{\min})$$

$$= \frac{1}{\eta}(\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} - \theta_{s,a^{\mathcal{P}^i},a^i}^{\tau^i(t),i}) + \frac{1}{(1-\gamma)}(V_{\max} - V_{\min})$$

where we have used that $|\frac{\partial V^{t'}(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i}| \leq \frac{1}{(1-\gamma)}(V_{\max} - V_{\min})$.
By equation (8), we know

$$\lim_{t\to\infty} \inf Z^{t,a^{\mathcal{P}^i},i} = -\infty \tag{9}$$

$$\frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi^{t,i}(a^i|s, a^{\mathcal{P}^i})\pi^{t,\mathcal{P}^i}(a^{\mathcal{P}^i}|s)A^{t,i}(s, a^{\mathcal{P}^i}, a^i)$$

For any $\mathcal{T}^{t,i} \neq \emptyset, \forall t' \in \mathcal{T}^{t,i}$, from lemma A.4, we know:

$$\left| \frac{\partial V^{t'}(\mu)/\partial \theta_{s,a_-^i}^i}{\partial V^{t'}(\mu)/\partial \theta_{s,a^i}^i} \right| = \left| \frac{\pi^{t',i}(a_-^i|s, a^{\mathcal{P}^i})A^{t',i}(s, a^{\mathcal{P}^i}, a_-^i)}{\pi^{t',i}(a^i|s, a^{\mathcal{P}^i})A^{t',i}(s, a^{\mathcal{P}^i}, a^i)} \right| \geq \exp(\theta_0^i - \theta_{s,a^{\mathcal{P}^i},a^i}^{t',i}) \frac{\Delta}{4(V_{\max} - V_{\min})}$$

$$\geq \exp(\delta^i) \frac{\Delta}{4(V_{\max} - V_{\min})}$$

where we have used that $|A^{t',i}(s, a^{\mathcal{P}^i}, a^i)| \leq V_{\max} - V_{\min}$ and $\forall t' > T_1, |A^{t',i}(s, a^{\mathcal{P}^i}, a^i_-)| \geq \frac{\Delta}{4}$.

Since both $\frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i_-}}$ and $\frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i}}$ are negative, we can get:

$$\frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i_-}} \leq \exp(\delta^i) \frac{\Delta}{4(V_{\max} - V_{\min})} \frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i}} \tag{10}$$

For non-empty $\mathcal{T}^{t,i}$,

$$\frac{1}{\eta}(\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_-} - \theta^{T_1,i}_{s,a^{\mathcal{P}^i}, a^i_-}) = \sum_{t'=T_1}^{t-1} \frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i_-}} \leq \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i_-}}$$

By Equation (10)

$$\leq \exp(\delta^i) \frac{\Delta}{4(V_{\max} - V_{\min})} \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial V^{t'}(\mu)}{\partial \theta^i_{s,a^{\mathcal{P}^i}, a^i}}$$

$$= \exp(\delta^i) \frac{\Delta}{4(V_{\max} - V_{\min})} Z^{t, a^{\mathcal{P}^i}, i}$$

which together with the fact that $\theta^{T_1,i}_{s,a^{\mathcal{P}^i}, a^i_-}$ is some finite constant and equation (9) lead to

$$\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_-} \to -\infty \text{ as } t \to \infty$$

this contradicts the assumption that $\{\theta^i_{s,a^{\mathcal{P}^i}, a^{t,i}_-}\}_{t \geq T_1}$ is lower bounded by θ^i_0 and complete the proof. □

Lemma A.15. *Consider any* $s, a^{\mathcal{P}^i}$ *where* $I_+^{s,a^{\mathcal{P}^i}, i} \neq \emptyset$. *Then, if* $\lim_{t \to \infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, *we have* $\forall a^i_+ \in I_+^{s,a^{\mathcal{P}^i}, i}$,

$$\sum_{a^i \in B_0^{s,a^{\mathcal{P}^i}, i}(a^i_+)} \theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i} \to \infty$$

Proof. For any $a^i \in B_0^{s,a^{\mathcal{P}^i}, i}(a^i_+)$. By definition, we know that $\forall t > T_0, \pi^{t,i}(a^i_+|s, a^{\mathcal{P}^i}) < \pi^{t,i}(a^i|s, a^{\mathcal{P}^i})$, which implies that $\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_+} < \theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i}$. Since in lemma A.14, $\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_+}$ is lower bounded as $t \to \infty$, we know that $\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i}$ is lower bounded as $t \to \infty$. This together with lemma A.12 proves that

$$\sum_{a^i \in B_0^{s,a^{\mathcal{P}^i}, i}(a^i_+)} \theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i} \to \infty$$

□

Lemma A.16. $\forall i, a^{\mathcal{P}^i}, a^i$, *if* $\lim_{t \to \infty} d_\mu^{\pi^t}(s, a^{\mathcal{P}^i}) > 0$, *then* $I_+^{s,a^{\mathcal{P}^i}, i} = \emptyset$.

Proof. Suppose $I_+^{s,a^{\mathcal{P}^i}, i}$ is non-empty for some $s, i, a^{\mathcal{P}^i}$, else the proof is complete. Let $a^i_+ \in I_+^{s,a^{\mathcal{P}^i}, i}$. Then, by lemma A.15, we know

$$\sum_{a^i \in B_0^{s,a^{\mathcal{P}^i}, i}(a^i_+)} \theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i} \to \infty \tag{11}$$

For $a^i \in I_-^{s,a^{\mathcal{P}^i}, i}$, since $\frac{\pi^{t,i}(a^i|s,a^{\mathcal{P}^i})}{\pi^{t,i}(a^i_+|s,a^{\mathcal{P}^i})} = \exp(\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i} - \theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_+}) \to 0$ (as $\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i_+}$ is lower bounded and $\theta^{t,i}_{s,a^{\mathcal{P}^i}, a^i} \to -\infty$ by lemma A.14), there exists $T_2 > T_0$ such that

$$\frac{\pi^{t,i}(a^i|s, a^{\mathcal{P}^i})}{\pi^{t,i}(a^i_+|s, a^{\mathcal{P}^i})} < \frac{\Delta}{8|\mathcal{A}^i|(V_{\max} - V_{\min})}$$

$$\longrightarrow -\sum_{a^i \in I_-^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})(V_{\max} - V_{\min}) > -\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{8} \tag{12}$$

For $a^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$, by definition of $\bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)$, we have $A^{t,i}(s,a^{\mathcal{P}^i},a^i) \to 0$ and by lemma A.13, $\forall t > T_{a^{\mathcal{P}^i},a_+^i}\, 1 < \frac{\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})}{\pi^{t,i}(a^i|s,a^{\mathcal{P}^i})}$. Then, $\exists T_3 > T_2, T_{a^{\mathcal{P}^i},a_+^i}$ such that

$$|A^{t,i}(s,a^{\mathcal{P}^i},a^i)| < \frac{\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})}{\pi^{t,i}(a^i|s,a^{\mathcal{P}^i})}\frac{\Delta}{16|\mathcal{A}^i|}$$

$$\longrightarrow \sum_{a^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})|A^{t,i}(s,a^{\mathcal{P}^i},a^i)| < \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{16}$$

$$\longrightarrow -\pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{16} < \sum_{a^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) < \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{16} \tag{13}$$

For $t > T_3$,

$$0 = \sum_{a^i \in \mathcal{A}^i} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$= \sum_{a^i \in I_0^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) + \sum_{a^i \in I_+^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$+ \sum_{a^i \in I_-^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$\overset{(a)}{\geq} \sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) + \sum_{a^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$+ \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a_+^i) + \sum_{a^i \in I_-^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

$$\overset{(b)}{\geq} \sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) + \sum_{a^i \in \bar{B}_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) + \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{4}$$

$$- \sum_{a^i \in I_-^{s,a^{\mathcal{P}^i},i}} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})(V_{\max} - V_{\min})$$

$$\overset{(c)}{\geq} \sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i) - \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{16} + \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{4} - \pi^{t,i}(a_+^i|s,a^{\mathcal{P}^i})\frac{\Delta}{8}$$

$$> \sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \pi^{t,i}(a^i|s,a^{\mathcal{P}^i})A^{t,i}(s,a^{\mathcal{P}^i},a^i)$$

where (a) uses $\forall a^i \in I_+^{s,a^{\mathcal{P}^i},i}$ and $t > T_3 > T_1, A^{t,i}(s,a^{\mathcal{P}^i},a^i) > 0$ from lemma A.7, (b) uses $\forall t > T_3 > T_1, A^{t,i}(s,a^{\mathcal{P}^i},a_+^i) > \frac{\Delta}{4}$ from lemma A.7 and $A^{t,i}(s,a^{\mathcal{P}^i},a^i) \geq -(V_{\max} - V_{\min})$, (c) uses equation (12) and equation (13). This implies that

$$\forall t > T_3, \sum_{a^i \in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} < 0$$

which contradicts with equation (11) which leads to

$$\lim_{t\to\infty} \sum_{a^i\in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} (\theta_{s,a^{\mathcal{P}^i},a^i}^{t,i} - \theta_{s,a^{\mathcal{P}^i},a^i}^{T_3,i}) = \eta \sum_{t=T_3}^{\infty} \sum_{a^i\in B_0^{s,a^{\mathcal{P}^i},i}(a_+^i)} \frac{\partial V^t(\mu)}{\partial \theta_{s,a^{\mathcal{P}^i},a^i}^i} \to \infty$$

Therefore, the set $I_+^{s,a^{\mathcal{P}^i},i} = \emptyset$. □

Theorem A.17. *Under Assumptions 5.2 - 5.4, suppose every agent i follows the policy gradient dynamics (2), which results in the update dynamics (3) for each each agent i, parent actions $a^{\mathcal{P}^i}$, and local action a^i, with $\eta \le \frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, then the converged BN policy $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is a Nash policy.*

Proof. For convenience, denote $\sum_{a^{-\mathcal{P}^i}} \pi_\theta(a^{-\mathcal{P}^i}, a^{\mathcal{P}^i}|s)$ as $\overline{\pi}_\theta^{\mathcal{P}^i}(\cdot|s)$ so that $d_\mu^{\pi_\theta}(s, a^{\mathcal{P}^i}) = d_\mu^{\pi_\theta}(s)\overline{\pi}_\theta^{\mathcal{P}^i}(\cdot|s)$.
$\forall i \in \mathcal{N}$, let $\theta'_* = [\theta_*^{-i}, \tilde{\theta}_*^i]$ be the parameters of any joint policy where only agent i's parameters are changed.
By performance difference lemma,

$$V^{\pi_{\theta'_*}} - V^{\pi_{\theta_*}} = \frac{1}{1-\gamma} \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}} \mathbb{E}_{\bar{a}\sim\pi_{\theta'_*}} \left[A^{\pi_{\theta_*}}(\bar{s}, \bar{a}) \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}(\cdot)} \mathbb{E}_{\bar{a}^{\mathcal{P}^i}\sim\overline{\pi}_{\theta'_*}^{\mathcal{P}^i}(\cdot|\bar{s})} \mathbb{E}_{\bar{a}^i\sim\pi_{\tilde{\theta}_*^i}^i(\cdot|\bar{s},\bar{a}^{\mathcal{P}^i})} \mathbb{E}_{\bar{a}^{-\mathcal{P}_+^i}\sim\pi_{\theta'_*}^{-\mathcal{P}_+^i}(\cdot|\bar{s},a_+^{\mathcal{P}^i})} \left[Q^{\pi_{\theta_*}}(\bar{s}, \bar{a}^{\mathcal{P}^i}, \bar{a}^i, \bar{a}^{-\mathcal{P}_+^i}) - V^{\pi_{\theta_*}}(\bar{s}) \right]$$

Since $(\theta')^{-i} = \theta_*^{-i}$ which means $\overline{\pi}_{\theta'_*}^{\mathcal{P}^i}(\cdot|\bar{s}) = \overline{\pi}_{\theta_*}^{\mathcal{P}^i}(\cdot|\bar{s}), \pi_{\theta'_*}^{-\mathcal{P}_+^i}(\cdot|\bar{s}, a_+^{\mathcal{P}^i}) = \pi_{\theta_*}^{-\mathcal{P}_+^i}(\cdot|\bar{s}, a_+^{\mathcal{P}^i})$,

$$= \frac{1}{1-\gamma} \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}(\cdot)} \mathbb{E}_{\bar{a}^{\mathcal{P}^i}\sim\overline{\pi}_{\theta_*}^{\mathcal{P}^i}(\cdot|\bar{s})} \mathbb{E}_{\bar{a}^i\sim\pi_{\tilde{\theta}_*^i}^i(\cdot|\bar{s},\bar{a}^{\mathcal{P}^i})} \mathbb{E}_{\bar{a}^{-\mathcal{P}_+^i}\sim\pi_{\theta_*}^{-\mathcal{P}_+^i}(\cdot|\bar{s},a_+^{\mathcal{P}^i})} \left[Q^{\pi_{\theta_*}}(\bar{s}, \bar{a}^{\mathcal{P}^i}, \bar{a}^i, \bar{a}^{-\mathcal{P}_+^i}) - V^{\pi_{\theta_*}}(\bar{s}) \right]$$

By lemma A.16 which proves either $\overline{\pi}_{\theta_*}^{\mathcal{P}^i}(\cdot|\bar{s}) = 0$ or $I_+^{s,a^{\mathcal{P}^i},i} = \emptyset$,

$$\le \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}(\cdot)} \mathbb{E}_{\bar{a}^{\mathcal{P}^i}\sim\overline{\pi}_{\theta_*}^{\mathcal{P}^i}(\cdot|\bar{s})} \mathbb{E}_{\bar{a}^i\sim\pi_{\tilde{\theta}_*^i}^i(\cdot|\bar{s},\bar{a}^{\mathcal{P}^i})} \mathbb{E}_{\bar{a}^{-\mathcal{P}^i}\sim\pi_{\theta_*}^{-\mathcal{P}^i}(\cdot|\bar{s},a^{\mathcal{P}^i})} \left[Q^{\pi_{\theta_*}}(\bar{s}, \bar{a}^{\mathcal{P}^i}, \bar{a}^{-\mathcal{P}^i}) - V^{\pi_{\theta_*}}(\bar{s}) \right]$$

$$= \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}(\cdot)} \mathbb{E}_{\bar{a}^{\mathcal{P}^i}\sim\overline{\pi}_{\theta_*}^{\mathcal{P}^i}(\cdot|\bar{s})} \mathbb{E}_{\bar{a}^{-\mathcal{P}^i}\sim\pi_{\theta_*}^{-\mathcal{P}^i}(\cdot|\bar{s},a^{\mathcal{P}^i})} \left[Q^{\pi_{\theta_*}}(\bar{s}, \bar{a}^{\mathcal{P}^i}, \bar{a}^{-\mathcal{P}^i}) - V^{\pi_{\theta_*}}(\bar{s}) \right]$$

$$= \mathbb{E}_{\bar{s}\sim d_\mu^{\pi_{\theta'_*}}(\cdot)} \left[V^{\pi_{\theta_*}}(\bar{s}) - V^{\pi_{\theta_*}}(\bar{s}) \right] = 0$$

.

$$\longrightarrow V^{\pi_{\theta'_*}} \le V^{\pi_{\theta_*}}$$

Therefore, $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is a Nash policy. □

Corollary A.18 (Asymptotic convergence of BN policy gradient to optimal fully-correlated BN joint policy). *Under Assumptions 5.2 - 5.4 and additional Assumption 5.6 that assumes positive visitation measure for any augmented state, suppose every agent $i \in \mathcal{N}$ follows the policy gradient dynamics (2), which results in the update dynamics (3) for each each agent i, parent actions $a^{\mathcal{P}^i}$, and local action a^i, with $\eta \le \frac{(1-\gamma)^3}{8N(r_{\max}-r_{\min})}$, then the converged fully-correlated BN policy $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is an optimal policy.*

Proof. We can assume without loss of generality that agents $1 \cdots N$ in $\mathcal{G}$ has a topological ordering of $1 \cdots N$ (This means that agent $i \in \mathcal{N}$ is the source of $N - i$ edges and target of $i - 1$ edges).
Note that in this case, $\forall i, a^{\mathcal{P}^i} = [a_+^{\mathcal{P}^{i-1}}]$,

$$Q^{\pi_\theta,i}(s, a^{\mathcal{P}^i}) = \mathbb{E}_{\bar{a}^{-\mathcal{P}^i}\sim\pi_\theta(\cdot|s,a^{\mathcal{P}^i})} \left[Q^{\pi_\theta}(s, a^{\mathcal{P}^i}, \bar{a}^{-\mathcal{P}^i}) \right]$$

$$= \mathbb{E}_{\bar{a}^{-\mathcal{P}_+^{i-1}}\sim\pi_\theta(\cdot|s,a_+^{\mathcal{P}^{i-1}})} \left[Q^{\pi_\theta}(s, a^{\mathcal{P}_+^{i-1}}, \bar{a}^{-\mathcal{P}_+^{i-1}}) \right] = Q^{\pi_\theta,i-1}(s, a_+^{\mathcal{P}^{i-1}}) \quad (14)$$

With assumption 5.6, we know that $\forall i, a^{\mathcal{P}^i}, I_+^{s,a^{\mathcal{P}^i},i} = \emptyset$.

$\forall a = [a^{\mathcal{P}^N}, a^N]$,

$$Q^{\pi_{\theta*}}(s,a) = Q^{\pi_{\theta*}}(s, a^{\mathcal{P}^N}, a^N) = Q^{\pi_{\theta*},N}(s, a^{\mathcal{P}^N}, a^N)$$

By $I_+^{s,a^{\mathcal{P}^N},N} = \emptyset$,

$$\leq Q^{\pi_{\theta*},N}(s, a^{\mathcal{P}^N})$$

By Equation (14),

$$= Q^{\pi_{\theta*},N-1}(s, a^{\mathcal{P}^{N-1}}, a^{N-1})$$

By $I_+^{s,a^{\mathcal{P}^{N-1}},N-1} = \emptyset$,

$$\leq Q^{\pi_{\theta*},N-1}(s, a^{\mathcal{P}^{N-1}})$$

By Equation (14),

$$= Q^{\pi_{\theta*},N-2}(s, a^{\mathcal{P}^{N-2}}), a^{N-2})$$

By keep doing the same procedure above,

$$\leq Q^{\pi_{\theta*},1}(s, a^{\mathcal{P}^1})$$

Since $a^{\mathcal{P}^1} = \emptyset$,

$$= V^{\pi_{\theta*}}(s)$$

Then, since $\forall s, a, Q^{\pi_{\theta*}}(s,a) \leq V^{\pi_{\theta*}}(s)$, we know that $(\pi_{\theta_*^1}^1, \cdots, \pi_{\theta_*^N}^N, \mathcal{G})$ is an optimal policy. □

B. Experiment details

B.1. Tabular Coordination Game

B.1.1. PSEUDOCODE FOR THE REWARD FUNCTION IN COORDINATION GAME

Algorithm 1 Calculate the team reward for N agents in state s

---

  if $(N = 2)$ or $(N = 3)$ **then**
    difference_bound=1
  else
    difference_bound=2
  end if
  if $\text{abs}(s.\text{count}(0) - s.\text{count}(1)) \leq$ difference_bound **then**
    if $s.\text{count}(0) < s.\text{count}(1)$ **then**
      reward$= 1$
    else
      reward$= 0$
    end if
  else if $s.\text{count}(0) > s.\text{count}(1)$ **then**
    reward$= 3$
  else
    reward$= 2$
  end if

B.1.2. COORDINATION GAME ENVIRONMENT HYPERPARAMETERS

Table 1. CG Env Hyperparameters

Hyperparameter	Value
γ (discount factor)	0.95
μ (initial state distribution)	Uniform
ϵ	0.1

B.1.3. HYPERPARAMETERS FOR COORDINATION GAME (TABULAR)

Table 2. Hyperparameters

Hyperparameter	Value
Environment steps	2e5, 1e6, 2e7 for CG,Aloha, and SMAC, respectively.
Episode length	20, 25, 400 for CG,Aloha, and SMAC, respectively.
PPO epoch	5 for all environments.
Critic Learning rate	7e-4 for CG and aloha, and 5e-4 for SMAC.
Actor Learning rate	7e-4 for CG and aloha, and 5e-4 for SMAC.
Optimizer	Adam.
#Episodes for evaluation	100 for CG and aloha, and 32 for SMAC.
#Rollout threads	32 for CG and aloha, 8 for SMAC.
#Training threads	32 for CG and aloha, 1 for SMAC.
Hidden size	64 for all environments.
Actor architecture for CG	Concat(Base(s), $a^{\mathcal{P}^i}$))-FC(action dim)-softmax
Actor architecture for Aloha	Concat(Base(o^i), Base(Concat($o^{\mathcal{P}^i}$, $a^{\mathcal{P}^i}$))-FC(action dim)-softmax
Actor architecture for SMAC	Concat(o^i, $a^{\mathcal{P}^i}$)-Base(hidden)-softmax
Edge Net architecture for CG	Concat($\{o^i\}_{i=1}^N$)-DeepSet
Edge Net architecture for Aloha and SMAC	Concat($\{o^i\}_{i=1}^N$)-FC(hidden)-Relu-FC($2N^2$)
Permutation Net architecture for CG and Aloha	Concat($\{o^i\}_{i=1}^N$)-FC(hidden)-Relu-FC($2N^2$)
Permutation Net architecture for SMAC	Always output identity matrix
Critic architecture for all environments	joint observation or state-Base(hidden)-FC(1)

Coordination Game is abbreviated as CG.
Base(hidden): FC(hidden)-Relu-FC(hidden)-Relu
DeepSetEncoder: FC(hidden)-Relu-FC(hidden)-Relu-FC(hidden)
DeepSetDecoder: FC(hidden)-Relu-FC(hidden)-Relu-FC($2N^2$)
DeepSet: input-DeepSetEncoder-mean(dim for agents)-DeepSetEncoder