Natural Policy Gradient for Bayesian Network Policies in Markov Potential Games

Dingyang Chen, Zhenyu Zhang, Yuan Ling, Qi Zhang

Keywords: natural policy gradient, Bayesian network, Markov potential games

Summary

We study natural policy gradient (NPG) with Bayesian network (BN) policies in Markov potential games (MPGs) under softmax parameterization. BN policies represent correlated joint policies via a directed acyclic graph, generalizing the independent (product) policies used in prior work. We prove that BN-NPG converges to near-Nash policies at an $O(1/K)$ rate averaged over K iterations for any fixed DAG, and that for fully correlated BN policies, every near-Nash policy is also near-optimal, recovering the global optimality guarantee of single-agent NPG. Unlike product policies, BN policies condition each agent's action on parent actions, requiring sufficient exploration over all parent-action profiles. We impose this as an assumption in the unregularized setting, and then remove it by designing a log-barrier regularizer, weighted by the frequency of encountering each parent-action profile under the current policy, that provably maintains exploration throughout training. We further extend these results to partially observable settings with factored state spaces, where each agent observes only its local state.

Contribution(s)

1. $O(1/K)$ convergence of natural policy gradient under softmax parameterization with Bayesian network (BN) policies to near-Nash policies in Markov potential games, for any fixed DAG structure, averaged over K iterations.
   Context: Prior work established $O(1/K)$ convergence for NPG with product policies (Sun et al., 2023). BN policy gradient (without natural gradient) was studied in Chen & Zhang (2023); Chen et al. (2026).
2. For fully correlated BN policies, every near-Nash policy is also near-optimal with controlled ratio, recovering the global optimality guarantee of single-agent NPG.
   Context: Product-policy NPG only guarantees convergence to Nash, which can be arbitrarily suboptimal (Zhang et al., 2022; Sun et al., 2023).
3. A log-barrier regularizer, weighted by the frequency of encountering each parent-action profile under the current policy, that provably maintains exploration throughout training, removing the need for a priori policy exploration assumptions.
   Context: The unweighted log-barrier in Zhang et al. (2022) suffices for product policies but introduces uncontrollable terms for BN policies due to the parent action marginal probability.
4. Extension to partially observable settings with factored state spaces, where each agent observes only its local state. Convergence bounds incur an additive term proportional to ϵ_{po}, the maximum discrepancy between full-state and local-state advantages.
   Context: All prior NPG convergence results for MPGs assume full state observation (Fox et al., 2022; Zhang et al., 2022; Sun et al., 2023).

Natural Policy Gradient for Bayesian Network Policies in Markov Potential Games

Dingyang Chen[1], **Zhenyu Zhang**[1], **Yuan Ling**[1], **Qi Zhang**[2]
{dingerwa, zhenyuzh, yualing}@amazon.com, qzhang9@wpi.edu

[1]**Amazon**
[2]**Worcester Polytechnic Institute**

Abstract

We study natural policy gradient (NPG) with Bayesian network (BN) policies in Markov potential games (MPGs) under softmax parameterization. BN policies represent correlated joint policies via a directed acyclic graph, generalizing the independent (product) policies used in prior work. We prove that BN-NPG converges to near-Nash policies at an $O(1/K)$ rate averaged over K iterations for any fixed DAG, and that for fully correlated BN policies, every near-Nash policy is also near-optimal, recovering the global optimality guarantee of single-agent NPG. Unlike product policies, BN policies condition each agent's action on parent actions, requiring sufficient exploration over all parent-action profiles. We impose this as an assumption in the unregularized setting, and then remove it by designing a log-barrier regularizer, weighted by the frequency of encountering each parent-action profile under the current policy, that provably maintains exploration throughout training. We further extend these results to partially observable settings with factored state spaces, where each agent observes only its local state.

1 Introduction

Multi-agent reinforcement learning has achieved remarkable empirical success across domains such as robotics (Corke et al., 2005), traffic control (Chu et al., 2019), and power systems (Callaway & Hiskens, 2010). A central theoretical challenge is finding Nash policies in Markov games (MGs) Shapley (1953), where each agent plays a best response to all others. In the single-agent case, Nash policies coincide with optimal policies, and natural policy gradient (NPG) under softmax parameterization is known to converge to the global optimum at an $O(1/K)$ rate after K iterations Agarwal et al. (2021).

For multi-agent settings, Markov potential games (MPGs) (Macua et al., 2018; Leonardos et al., 2021; Zhang et al., 2021) provide a tractable subclass that includes fully cooperative games as a special case. Recent work has established that independent NPG with product policies converges to Nash policies in MPGs, with rates improving from $O(1/\sqrt{K})$ (Zhang et al., 2022) to $O(1/K)$ (Sun et al., 2023). However, product-policy Nash equilibria can be arbitrarily suboptimal: the potential function, which captures social welfare in cooperative settings, may be far below its maximum at a Nash policy. This gap is inherent to the product-policy class, since product policies cannot represent correlations between agents' actions, and which Nash equilibrium the algorithm converges to is highly sensitive to initialization.

In this work, we close this gap by combining NPG with Bayesian network (BN) policies, which represent correlated joint policies via a directed acyclic graph (DAG) and interpolate between product policies and fully expressive joint policies. We establish that BN-NPG converges to near-Nash policies at an $O(1/K)$ rate averaged over K iterations for any fixed DAG, and prove that for fully correlated BN policies, every near-Nash policy is also near-optimal, recovering the single-agent

optimality guarantee. A key technical contribution is a log-barrier regularizer, weighted by the frequency of encountering each parent-action profile under the current policy, that provably maintains exploration throughout training rather than assuming it holds a priori. All prior NPG convergence results for MPGs assume full state observation (Fox et al., 2022; Zhang et al., 2022; Sun et al., 2023); we further extend our results to partially observable settings with factored state spaces, where each agent observes only its local state. The convergence bounds incur an additive term proportional to the maximum discrepancy between the full-state and local-state advantages, recovering the full-observation guarantees when this discrepancy vanishes.

2 Related work

Natural policy gradient under softmax parameterization. In the single-agent setting, Agarwal et al. (2021) established that unregularized natural policy gradient converges to the global optimum at an $O(1/K)$ rate after K iterations. Extending this to MPGs with independent (product) policies, Fox et al. (2022) proved asymptotic convergence of independent NPG to Nash policies. Zhang et al. (2022) provided the first finite-time convergence rates, albeit at a slower $O(1/\sqrt{K})$ rate, and introduced log-barrier regularization to remove dependence on trajectory-dependent constants. Sun et al. (2023) further improved the rate to $O(1/K)$, matching the single-agent result. However, all of these multi-agent results only guarantee convergence to a Nash policy, which can be arbitrarily suboptimal in terms of the potential function, whereas the single-agent counterpart achieves global optimality.

Multi-agent correlated policies in multi-agent RL. The suboptimality of product policies has motivated a line of work on correlated joint policies, which offer greater expressiveness and thus stronger optimality guarantees. Ye et al. (2023) proved the suboptimality of product-policy gradient methods and proposed a sequential (autoregressive) policy transformation to recover global optimality. Christianos et al. (2023) addressed equilibrium selection by designing an actor-critic algorithm that steers convergence toward Pareto-optimal Nash equilibria, though still within the product-policy class.

A natural and structured way to represent correlated joint policies is through Bayesian networks (BNs) (Heckerman, 2020), which interpolate between product policies (empty DAG) and fully correlated policies (fully connected DAG) via the choice of dependency graph. Ruan et al. (2022) empirically demonstrated the benefits of BN policies for coordination. Chen & Zhang (2023) showed that policy gradient with fully correlated BN policies asymptotically converges to global optima in cooperative Markov games, a special case of MPGs. Chen et al. (2026) extended this to finite-time convergence rates via log-barrier regularization. However, all existing BN-policy results are based on (vanilla) policy gradient; no convergence results exist for natural policy gradient with BN policies, which is the gap our work addresses. Additionally, all of the above works assume full state observation; we further extend our results to partially observable settings with factored state spaces.

3 Preliminaries

We consider a stochastic game $\mathcal{M} = (n, \mathcal{S}, \mathcal{A}, P, \{r_i\}_{i\in[n]}, \gamma, \rho)$ consisting of n agents denoted by a set $[n] = \{1, \dots, n\}$. The global state space is $\mathcal{S}$, and the global action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is the product of individual action spaces, with the global action denoted as $a := (a_1, \dots, a_n)$. The transition model is $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\cdot)$ denotes the probability simplex. Each agent i has a reward function $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We use $\gamma \in (0, 1)$ to denote the discount factor and $\rho \in \Delta(\mathcal{S})$ to denote the initial state distribution. With full observability, meaning each agent observes the global state $s \in \mathcal{S}$, a general joint policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to distributions over the joint action space. The state value function for agent i is defined as: $V_i^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^\infty \gamma^t r_i(s_t, a_t) \mid s_0 = s\right]$, and its expectation over the initial distribution is $V_i^\pi(\rho) := \mathbb{E}_{s\sim\rho}[V_i^\pi(s)]$. The state-action value function is: $Q_i^\pi(s, a) := \mathbb{E}^\pi \left[\sum_{t=0}^\infty \gamma^t r_i(s_t, a_t) \mid s_0 = s, a_0 = a\right]$. The discounted state visitation distribution is

$d_\rho^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid s_0 \sim \rho, \pi)$. The commonly used joint policy subclass is the *product policy*, $\pi = (\pi_1, \ldots, \pi_n) : \mathcal{S} \to \times_{i\in[n]}\Delta(\mathcal{A}^i)$, where joint policy π is factored as a product of local policies $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, such that $\pi(a|s) = \prod_{i\in[n]}\pi_i(a_i|s)$. For a product policy $\pi = (\pi_1, \ldots, \pi_n)$ and an agent i, we write $\pi_{-i} := (\pi_j)_{j\neq i}$ for the local policies of all agents other than i, so that the joint policy factors as $\pi = (\pi_i, \pi_{-i})$. We use (π_i', π_{-i}) to denote the joint policy obtained by replacing agent i's local policy π_i with π_i' while keeping all other agents' policies π_{-i} fixed. We focus on the subclass of Markov potential games (Zhang et al., 2022) defined as:

Definition 1 (Markov potential game)**.** The stochastic game $\mathcal{M}$ is a Markov potential game (MPG) if there exists a bounded potential function $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ such that for any agent i, state s, and policies π_i, π_i', π_{-i}:

$$V_i^{\pi_i',\pi_{-i}}(s) - V_i^{\pi_i,\pi_{-i}}(s) = \Phi^{\pi_i',\pi_{-i}}(s) - \Phi^{\pi_i,\pi_{-i}}(s),$$

where $\Phi^\pi(s) := \mathbb{E}^\pi[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t) \mid s_0 = s]$ is the potential value function.

We will use shorthand notation $\Phi^\pi(\rho) := \mathbb{E}_{s\sim\rho}[\Phi^\pi(s)]$. We assume $\phi(s,a) \in [0, \phi_{\max}]$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, and consequently $\Phi^\pi(\rho) \in [0, \frac{\phi_{\max}}{1-\gamma}]$. The state-action value function and advantage function for the potential are: $Q_\phi^\pi(s,a) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t) \mid s_0 = s, a_0 = a\right]$ and $A_\phi^\pi(s,a) := Q_\phi^\pi(s,a) - \Phi^\pi(s)$, respectively.

We now formalize the two solution concepts used throughout the paper. The first is the Nash equilibrium, which requires that no agent can unilaterally improve its value:

Definition 2 (ϵ-Nash policy)**.** The *NE-gap* of a policy π is:

$$\text{NE-gap}(\pi) := \max_{i\in[n],\pi_i'}\left[V_i^{\pi_i',\pi_{-i}}(\rho) - V_i^\pi(\rho)\right].$$

A policy π is an *ϵ-Nash policy* if NE-gap$(\pi) \leq \epsilon$.

The second measures how far a policy is from maximizing the total potential function:

Definition 3 (ϵ-optimal policy)**.** The *optimality-gap* of a policy π is:

$$\text{optimality-gap}(\pi) := \max_{\pi^*}\Phi^{\pi^*}(\rho) - \Phi^\pi(\rho).$$

A policy π is *ϵ-optimal* if optimality-gap$(\pi) \leq \epsilon$.

Potential maximization is often desirable in its own right; for instance, in fully cooperative games the potential equals the shared reward, so a higher potential directly improves every agent's payoff. In general, a Nash policy need not be optimal: the NE-gap can be zero while the optimality-gap remains large. In such settings, convergence to a near-optimal policy is the relevant goal, and convergence to a near-Nash policy alone is insufficient.

Bayesian network policy. Product policies assume agents act independently, which limits expressiveness. To enable controlled correlation between agents, we follow prior work (Chen & Zhang, 2023) to consider a Bayesian Network (BN) structure represented by a directed acyclic graph (DAG) $G = ([n], \mathcal{E})$ connecting the agents as vertices. For each agent i, let $\mathcal{P}^i := \{j : (j,i) \in \mathcal{E}\}$ denote the set of parent agents. The BN policy π_G is defined as:

$$\pi_G(a|s) = \prod_{i=1}^{n}\pi_i(a_i|s, a_{\mathcal{P}^i}), \tag{1}$$

where $a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i} := \times_{j\in\mathcal{P}^i}\mathcal{A}_j$ denotes the joint action of parent agents. Therefore, a BN joint policy is also factorized into local policies $(\pi_1, \ldots, \pi_n)$, thus generalizing product policies and allows Definitions 1, 2, and 3 to extend naturally to BN policies, which is the setting we consider in this work.

Notations. For a subset $\mathcal{M} \subseteq [n]$ of agents and its complement $-\mathcal{M}$, a joint action is decomposed as $a = (a_{\mathcal{M}}, a_{-\mathcal{M}})$. The conditional distribution over $a_{\mathcal{M}}$ given $a_{-\mathcal{M}}$ is: $\pi_{\mathcal{M}}(a_{\mathcal{M}}|s, a_{-\mathcal{M}}) \propto \pi(a_{\mathcal{M}}, a_{-\mathcal{M}}|s)$. The marginal distribution over $a_{\mathcal{M}}$ is $\pi_{\mathcal{M}}(a_{\mathcal{M}}|s) := \sum_{a_{-\mathcal{M}}} \pi(a_{-\mathcal{M}}, a_{\mathcal{M}}|s)$. The augmented state visitation distribution is: $d^{\pi}_{\rho}(s, a_{\mathcal{M}}) := d^{\pi}_{\rho}(s)\pi_{\mathcal{M}}(a_{\mathcal{M}}|s)$. The marginalized Q-function for agent i is $Q^{\pi}_i(s, a_{\mathcal{M}}) := \sum_{a_{-\mathcal{M}}} \pi(a_{-\mathcal{M}}|s, a_{\mathcal{M}})Q^{\pi}_i(s, a_{\mathcal{M}}, a_{-\mathcal{M}})$. The marginalized advantage function for agent i is $A^{\pi}_i(s, a_{\mathcal{P}^i}, a_i) := Q^{\pi}_i(s, a_{\mathcal{P}^i} \cup a_i) - Q^{\pi}_i(s, a_{\mathcal{P}^i})$.

We write $\kappa := \max_{i\in[n]} |\mathcal{P}^i|$ for the maximum in-degree of G, which controls the DAG sparsity: $\kappa = 0$ for product policies (empty DAG) and $\kappa = n - 1$ for the fully connected DAG. Without loss of generality, we assume agents are indexed according to a topological ordering of G, i.e., if $j \in \mathcal{P}^i$ then $j < i$.

4 Convergence of tabular BN unregularized natural policy gradient ascent

Fixing DAG G, we consider parameterizing local policies of a BN policy in the tabular softmax manner from the global state and parent actions as in Chen & Zhang (2023). For each local policy π_i is parameterized using a tabular softmax function with parameters: $\theta_i = \{\theta_i(s, a_{\mathcal{P}^i}, a_i) \in \mathbb{R} : s \in \mathcal{S}, a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i}, a_i \in \mathcal{A}_i\}$, leading to the induced softmax local policy:

$$\pi_{\theta_i}(a_i|s, a_{\mathcal{P}^i}) := \frac{\exp\left(\theta_i(s, a_{\mathcal{P}^i}, a_i)\right)}{\sum_{a'_i} \exp\left(\theta_i(s, a_{\mathcal{P}^i}, a'_i)\right)}. \tag{2}$$

The BN policy is thus parameterized as $\pi_\theta = (\pi_{\theta_1}, \ldots, \pi_{\theta_n})$. Prior work (Sun et al., 2023) established the convergence rate results of tabular natural policy gradient ascent with product policy. In this section, we generalize the results to BN policy with any fixed DAG.

Formally, consider the maximization over the parameterized BN policy class of potential $L(\theta) := \Phi^{\pi_\theta}(\rho)$. The natural policy gradient (NPG) update for agent i is:

$$\theta^{k+1}_i := \theta^k_i + \eta F_i(\theta^k)^{\dagger} \nabla_{\theta_i} L(\theta^k), \tag{3}$$

where $F_i(\theta)$ is the Fisher information matrix:

$$F_i(\theta) := \mathbb{E}_{s\sim d^{\pi_\theta}_{\rho}, a_{\mathcal{P}^i}\sim\pi_{\theta,\mathcal{P}^i}(\cdot|s), a_i\sim\pi_{\theta_i}(\cdot|s,a_{\mathcal{P}^i})} \left[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s, a_{\mathcal{P}^i}) \nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s, a_{\mathcal{P}^i})^{\top}\right].$$

Lemma 1. The NPG update (3) is equivalent to:

$$\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) \propto \pi^k_i(a_i|s, a_{\mathcal{P}^i}) \exp\left(\frac{\eta A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma}\right). \tag{4}$$

The block-diagonal structure of $F_i(\theta)$ and the cancellation of the augmented visitation factor $d^{\pi^k}_{\rho}(s, a_{\mathcal{P}^i})$ under the pseudo-inverse, which underlie this equivalence, are illustrated on a concrete $n = 3$ chain in Appendix A. The gradient form in Lemma 1 enables us to extend the finite-time convergence guarantees with product policy (Sun et al., 2023) to BN joint policies under the same assumptions in addition to one assumption unique to the BN policy on the exploration of parent actions, which we state below.

Assumption 1 (State exploration)**.** The MPG satisfies $\inf_\pi \min_s d^{\pi}_{\rho}(s) > 0$.

Assumption 1 is standard in the analysis of natural policy gradient methods for MPGs (Zhang et al., 2022; Sun et al., 2023) and holds if the initial distribution ρ assigns positive mass to every state.

Assumption 2 (Advantage gap)**.** The advantage gap, as defined below, is positive:

$$\delta := \inf_{\pi} \min_{i\in[n], s\in\mathcal{S}, a_{\mathcal{P}^i}\in\mathcal{A}_{\mathcal{P}^i}} \left[A^{\pi}_i(s, a_{\mathcal{P}^i}, a_{i_p}) - A^{\pi}_i(s, a_{\mathcal{P}^i}, a_{i_q})\right] > 0,$$

where $a_{i_p} \in \operatorname{argmax}_{a_j\in\mathcal{A}_i} A^{\pi}_i(s, a_{\mathcal{P}^i}, a_j)$ and $a_{i_q} \in \operatorname{argmax}_{a_j\in\mathcal{A}_i\setminus\{a_{i_p}\}} A^{\pi}_i(s, a_{\mathcal{P}^i}, a_j)$ denote the best and second-best actions for agent i.

The advantage gap δ generalizes its product-policy counterpart in Sun et al. (2023), where the minimum is taken over agents i and states s only, to additionally include parent actions $a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i}$. It measures the minimum separation between the best and second-best actions in terms of the advantage, uniformly across all agents, states, parent action profiles, and policies.

Assumption 3 (Policy exploration)**.** For all iterations k, states s, agents i, parent actions $a_{\mathcal{P}^i}$, and actions a_i:

$$\inf_{k\geq 0} \min_{s,i,a_{\mathcal{P}^i},a_i} \pi_i^k(a_i|s,a_{\mathcal{P}^i}) =: \nu > 0.$$

For product policies ($\mathcal{P}^i = \emptyset$ for all i), the convergence analyses of Zhang et al. (2022); Sun et al. (2023) rely on the quantity $c := \inf_{i\in[n],s\in\mathcal{S},k\geq 0} \sum_{a_j\in\mathcal{A}_{i_p}^k} \pi_i^k(a_j|s) > 0$, where $\mathcal{A}_{i_p}^k$ denotes the set of actions maximizing the advantage $A_i^{\pi^k}(s,a_i)$, the product-policy counterpart of our $A_i^{\pi}(s,a_{\mathcal{P}^i},a_i)$, for agent i at iteration k; ensuring $c > 0$ typically requires an assumption such as Assumption 3. For BN policies, this condition is additionally necessary for our proof because it requires bounding the parent marginal ratio $\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s)/\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)$, which depends on the probability of all actions (including suboptimal ones) through the parent marginal $\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \geq \nu^{|\mathcal{P}^i|}$. This assumption is removed in Section 4.2 via the log-barrier regularizer.

The following convergence analysis generalizes the product-policy framework of Sun et al. (2023) to BN policies by treating $(s,a_{\mathcal{P}^i})$ as an augmented state. The proof relies on two key lemmas, stated and proved in Appendix B: Lemma 5 establishes monotonic ascent of the potential function Φ along the BN-NPG iterates, and Lemma 6 introduces an interpolating function $f^k(\alpha)$ that bridges the BN-NPG update ($\alpha = \eta$) with the greedy improvement ($\alpha \to \infty$), connecting the potential ascent to the NE-gap through the advantage gap δ. Combining these two lemmas yields the main convergence result.

Theorem 1 (Convergence of BN-NPG in MPGs)**.** *Consider a Markov potential game with BN policy structure satisfying Assumptions 1–3, with the policy update following BN-NPG (4). For any $K \geq 1$, choosing $\eta = \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$ where $A_{\max} := \max_i |\mathcal{A}_i|$, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \leq \frac{2M\phi_{\max}}{K\nu^{\kappa}(1-\gamma)}\left(1 + \frac{2\sqrt{n}(1+A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)}\right),$$

where $M := \sup_{\pi}\max_s \frac{1}{d_\rho^\pi(s)}$ and $\kappa = \max_i |\mathcal{P}^i|$ is the maximum in-degree of the DAG.

Remark 4.1 (Dependence on n and the DAG)**.** The factor $1/\nu^{\kappa}$ in Theorem 1 grows exponentially with the maximum in-degree $\kappa = \max_i |\mathcal{P}^i|$, and hence with n in the worst case $\kappa = n-1$. It originates solely from bounding the parent marginal ratio $\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s)/\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \leq 1/\nu^{|\mathcal{P}^i|} \leq 1/\nu^{\kappa}$ (see the step from (4) to the NE-gap bound in the proof of Theorem 1), a term that is absent in product-policy analyses (Sun et al., 2023) where $\mathcal{P}^i = \emptyset$ and thus $\kappa = 0$. Hence this part of the bound is controlled by the DAG sparsity: sparse DAGs with small κ avoid the exponential blowup, and the worst case $\kappa = n-1$ is attained only by the fully connected DAG. Other parts of the bound, such as the $\sqrt{n}$ factors, may still depend explicitly on n. In the Nash-to-optimality results for fully correlated BN policies (Theorem 2 and Corollary 1) to follow next, the DAG will be fully connected by definition, so $\kappa = n-1$ and one pays the full worst-case exponent together with an additional $1/\nu^n$ factor from the Nash-to-optimality transfer (Lemma 9).

4.1 From Nash to optimality for fully correlated BN policies in MPGs

Theorem 1 guarantees convergence to a near-Nash policy for BN-NPG with any fixed DAG, but in general, a Nash equilibrium in an MPG can be arbitrarily suboptimal in terms of the potential Φ. In the single-agent setting, tabular softmax NPG converges to the globally optimal policy (Agarwal et al., 2021). Intuitively, a fully correlated BN policy resembles the single-agent setting because $\pi(a|s) = \prod_{i=1}^n \pi_i(a_i|s,a_{\mathcal{P}^i})$ can represent any joint distribution over a given s, so there is no loss of expressiveness. We show here that this intuition is indeed correct. We first formalize the notion of a fully correlated BN policy.

Definition 4 (Maximally connected BN and fully-correlated BN policy)**.** A BN policy $(\pi, \mathcal{G})$ is *fully-correlated* if its dependency graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is a *maximally connected Bayesian network*, i.e., a DAG with the maximum number of edges while remaining acyclic: $|\mathcal{E}| = n(n-1)/2$.

For fully correlated BN policies, it turns out that any ϵ-Nash policy is also ϵ'-optimal with a controlled ratio ϵ'/ϵ. The key ingredients to this result are: (i) a uniform bound on the per-agent advantage for ϵ-Nash policies (Lemma 7), (ii) a telescoping recursion on the marginalized Q-function exploiting the nested parent structure $\mathcal{P}^i = \{1, \dots, i-1\}$ (Lemma 8), and (iii) their combination into a Nash-to-optimality transfer (Lemma 9). These lemmas are stated and proved in Appendix B.

Combining the NE-gap bound in Theorem 1 with the Nash-to-optimality transfer in Lemma 9 yields the following optimality-gap guarantee.

Theorem 2. *Consider a Markov potential game with fully correlated BN policy structure satisfying Assumptions 1–3, with the policy update following (4). For any $K \geq 1$, choosing $\eta = \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{optimality-gap}(\pi^k) \leq \frac{2M^2\phi_{\max}(1-\nu^n)}{\nu^{2n-1}(1-\nu)K(1-\gamma)}\left(1 + \frac{2\sqrt{n}(1+A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)}\right).$$

4.2 NPG with log-barrier regularizer to ensure action exploration

The convergence results in Theorems 1 and 2 require Assumption 3, which posits a uniform lower bound $\nu > 0$ on all action probabilities throughout training. This assumption serves two purposes: (i) it bounds the parent marginal ratio $\pi^k_{\mathcal{P}^i}/\pi^{k+1}_{\mathcal{P}^i}$ that arises from the BN structure when connecting the NE-gap to the potential ascent, and (ii) it enables the Nash-to-optimality transfer for fully correlated BN policies. To avoid making this assumption explicitly, we introduce a log-barrier regularizer that provably maintains action exploration.

Formally, for BN policy π_θ, consider the log-barrier regularized objective at iteration k:

$$L^k_\lambda(\theta) := \Phi^{\pi_\theta}(\rho) + \lambda \sum_{i=1}^n \sum_{s,a_{\mathcal{P}^i}} d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \sum_{a_i} \log \pi_{\theta_i}(a_i|s, a_{\mathcal{P}^i}),$$

where $\lambda > 0$ is the regularization parameter and augmented state visitation distribution of the current policy π^k, $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$, is treated as a fixed weight of the log-barrier term.

For product policies ($\mathcal{P}^i = \emptyset$), Zhang et al. (2022) use an unweighted barrier $\sum_{s,a_i} \log \pi_{\theta_i}(a_i|s)$, which after the Fisher pseudo-inverse produces a $1/d_\rho^{\pi^k}(s)$ factor in the update; this is acceptable since $1/d_\rho^{\pi^k}(s) \leq M$. For BN policies, an unweighted barrier would analogously produce $1/d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) = 1/(d_\rho^{\pi^k}(s)\,\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s))$ in the update, where the parent marginal $\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)$ can be arbitrarily small. Weighting by $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$ cancels this factor against the Fisher matrix (see the proof of Lemma 2), yielding an update that depends on λ/π_i^k alone and removing the need for Assumption 3. The following lemma confirms that this cancellation occurs, giving a closed-form update free of $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$. Appendix A provides a fully worked $n = 3$ chain example illustrating both the Fisher pseudo-inverse cancellation and why the $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$-weighting (rather than an unweighted barrier) is needed.

Lemma 2. The log-barrier regularized NPG update $\theta_i^{k+1} = \theta_i^k + \eta F_i(\theta^k)^\dagger \nabla_{\theta_i} L^k_\lambda(\theta^k)$ is equivalent to:

$$\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) \propto \pi_i^k(a_i|s, a_{\mathcal{P}^i}) \exp\left(\frac{\eta A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\eta\lambda}{\pi_i^k(a_i|s, a_{\mathcal{P}^i})} - \eta\lambda|\mathcal{A}_i|\right). \tag{5}$$

Lemma establishes that the log-barrier regularized BN-NPG update (5) guarantees a uniform lower bound on all action probabilities, removing the need for making Assumption 3 explicitly.

Lemma 3. For $\eta \leq \frac{1}{15\left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right)}$ and $\theta^0 = \mathbf{0}$, running the log-barrier regularized BN-NPG update (5) guarantees that $\pi_i^k(a_i|s, a_{\mathcal{P}^i}) \geq \nu_\lambda := \frac{\lambda}{4\left(\lambda A_{\max} + \frac{1}{(1-\gamma)^2}\right)}$ for all $k \geq 0$, $i \in [n]$, $s \in \mathcal{S}$, $a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i}$, and $a_i \in \mathcal{A}_i$.

Since the update (5) is free of $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$, the bounded-probability argument of Zhang et al. (2022) applies directly to each $(s, a_{\mathcal{P}^i})$ block, yielding a stronger bound that is independent of M. More specifically, with the probability lower bound ν_λ in hand, the potential ascent and NE-gap analysis follows the same structure as the unregularized case, with two modifications handled in the appendix: Lemma 10 extends the potential ascent (Lemma 5) to the regularized update, picking up an additional λ-dependent residual from the log-barrier perturbation; and Lemma 11 bounds the discrepancy between the regularized update π_i^{k+1} and the unregularized softmax interpolation $\pi_{i,\eta}$ used in the f^k-to-NE-gap connection (Lemma 6). The $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$-weighting in the regularizer is essential: an unweighted barrier would reintroduce the uncontrollable $1/\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)$ factor into the bridge bound. Combining Lemmas 10 and 11 with the f^k-to-NE-gap connection (Lemma 6), and substituting $\nu = \nu_\lambda$ for the parent marginal ratio bound, we obtain the main convergence result.

Theorem 3. *Consider a Markov potential game with BN policy structure satisfying Assumptions 1–2, with the policy update following the log-barrier regularized BN-NPG (5). For any $K \geq 1$, choosing $\eta \leq \min\{\frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}, \frac{1}{15\left(\frac{1}{(1-\gamma)^2}+\lambda A_{\max}\right)}\}$, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \leq \frac{2M\phi_{\max}}{K\nu_\lambda^\kappa(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\delta\eta}\right) + R_\lambda,$$

where $\kappa = \max_i |\mathcal{P}^i|$ is the maximum in-degree of the DAG and $R_\lambda := \frac{2Mn\eta\lambda(\sqrt{n}(1+A_{\max})\phi_{\max}+1)}{\nu_\lambda^{\kappa+1}(1-\gamma)^2}\left(1+\frac{1-\gamma}{\nu_\lambda\delta\eta}\right)$.

By Theorem 3, R_λ vanishes as $\lambda \to 0$, reflecting the tradeoff between exploration (larger ν_λ) and regularization cost.

Similar to the unregularized case (Theorem 2), we can apply the Nash-to-optimality transfer (Lemma 9) to convert the NE-gap bound into an optimality-gap guarantee. Since Lemma 3 guarantees $\pi_i^k(a_i|s, a_{\mathcal{P}^i}) \geq \nu_\lambda$ for all k, this replaces Assumption 3 with the explicit lower bound $\nu = \nu_\lambda$. Combining Theorem 3 with Lemma 9 yields the following, where the fully connected BN has maximum in-degree $\kappa = n - 1$.

Corollary 1. Consider a Markov potential game with fully correlated BN policy structure satisfying Assumptions 1–2, with the policy update following the log-barrier regularized BN-NPG (5). For any $K \geq 1$, with η chosen as in Theorem 3, we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{optimality-gap}(\pi^k) \leq \frac{M(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)}\left[\frac{2M\phi_{\max}}{K\nu_\lambda^{n-1}(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\delta\eta}\right) + R_\lambda\right].$$

5 Extension to partially observable local states

The preceding results assume each agent observes the global state s. All prior NPG convergence results for MPGs, whether with product policies (Fox et al., 2022; Zhang et al., 2022; Sun et al., 2023) or BN policies (Theorems 1–3 above), rely on this assumption, since the NPG update and its analysis require per-state advantage functions and Fisher information matrices indexed by the global state. In many multi-agent systems, however, each agent has access only to a local view of the environment, making full state observation impractical. We now extend the log-barrier regularized BN-NPG to such a partially observable setting. Specifically, we consider a factored state space $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$, where agent i observes only its local state $s_i \in \mathcal{S}_i$. The transition, rewards, and potential remain defined on the global state $s = (s_1, \ldots, s_n)$. Each agent's policy now conditions on $(s_i, a_{\mathcal{P}^i})$ rather than $(s, a_{\mathcal{P}^i})$, which requires replacing the full-state advantage with a marginalized version that aggregates over unobserved states.

Each agent's policy conditions on $(s_i, a_{\mathcal{P}^i})$ via tabular softmax, giving $\pi(a|s) = \prod_i \pi_i(a_i|s_i, a_{\mathcal{P}^i})$. Define the augmented local-state visitation $d_\rho^\pi(s_i, a_{\mathcal{P}^i}) := \sum_{s_{-i}} d_\rho^\pi(s_i, s_{-i})\, \pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s_i, s_{-i})$ and the marginalized advantage:

$$\bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i) := \frac{1}{d_\rho^\pi(s_i, a_{\mathcal{P}^i})}\sum_{s_{-i}} d_\rho^\pi(s_i, s_{-i})\, \pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s_i, s_{-i})\, A_i^\pi((s_i, s_{-i}), a_{\mathcal{P}^i}, a_i). \quad (6)$$

The marginalized advantage $\bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i)$ arises naturally from the policy gradient for local-state policies. Since π_i depends only on $(s_i, a_{\mathcal{P}^i})$, the policy gradient of $L(\theta)$ with respect to $\theta_i(s_i, a_{\mathcal{P}^i}, a_i)$ aggregates over unobserved states:

$$\frac{\partial L(\theta)}{\partial \theta_i(s_i, a_{\mathcal{P}^i}, a_i)} = \frac{1}{1-\gamma} d_\rho^\pi(s_i, a_{\mathcal{P}^i})\, \pi_i(a_i|s_i, a_{\mathcal{P}^i})\, \bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i).$$

This has the same form as the full-observation gradient (cf. the proof of Lemma 1), with $d_\rho^\pi(s, a_{\mathcal{P}^i})$ replaced by $d_\rho^\pi(s_i, a_{\mathcal{P}^i})$ and A_i^π replaced by $\bar{A}_i^\pi$. The Fisher information matrix is block-diagonal with blocks indexed by $(s_i, a_{\mathcal{P}^i})$, each of the form $d_\rho^\pi(s_i, a_{\mathcal{P}^i})\, F_{i,s_i,a_{\mathcal{P}^i}}(\theta_i)$, so the pseudo-inverse cancels $d_\rho^\pi(s_i, a_{\mathcal{P}^i})$ and yields $\bar{A}_i^\pi/(1-\gamma)$ as the natural gradient direction. When $s_i = s$ (full observation), $\bar{A}_i^\pi$ reduces to $A_i^\pi(s, a_{\mathcal{P}^i}, a_i)$. In general, $\bar{A}_i$ and A_i differ; we measure this discrepancy and define the appropriate solution concept below.

Definition 5 (Partial observability gap)**.** The *partial observability gap* measures the worst-case discrepancy between the full-state and marginalized advantages. It enters the convergence bounds as an additive bias that captures the cost of partial observability:

$$\epsilon_{\text{po}} := \sup_\pi \max_{i,s,a_{\mathcal{P}^i},a_i} |A_i^\pi(s, a_{\mathcal{P}^i}, a_i) - \bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i)|.$$

Since each agent's policy conditions only on $(s_i, a_{\mathcal{P}^i})$, the NE-gap (Definition 2) is evaluated with deviations restricted to the same local-state policy class. The partial observability gap ϵ_{po} is zero when advantages do not depend on s_{-i} (e.g., factored rewards and transitions), and small when unobserved states have limited influence on agent i's advantage. The update rule applies the same d-weighted log-barrier design to the local-state setting.

The log-barrier regularized objective in the local-state setting takes the same form as the full-observation case, with $(s, a_{\mathcal{P}^i})$ replaced by $(s_i, a_{\mathcal{P}^i})$ and the weight $d_\rho^{\pi^k}(s_i, a_{\mathcal{P}^i})$ in place of $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$. The derivation of the update follows Lemma 2: the gradient decomposes into blocks indexed by $(s_i, a_{\mathcal{P}^i})$, and the Fisher pseudo-inverse cancels $d_\rho^{\pi^k}(s_i, a_{\mathcal{P}^i})$ from both the potential gradient and the log-barrier gradient, exactly as in the full-observation case. The only difference is that the full-state advantage $A_i^{\pi^k}$ is replaced by the marginalized advantage $\bar{A}_i^{\pi^k}$, yielding:

$$\pi_i^{k+1}(a_i|s_i, a_{\mathcal{P}^i}) \propto \pi_i^k(a_i|s_i, a_{\mathcal{P}^i}) \exp\left(\frac{\eta \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\eta\lambda}{\pi_i^k(a_i|s_i, a_{\mathcal{P}^i})} - \eta\lambda|\mathcal{A}_i|\right). \tag{7}$$

Since $|\bar{A}_i| \le 1/(1-\gamma)$, the bounded probability guarantee (Lemma 3) holds verbatim with $(s, a_{\mathcal{P}^i})$ replaced by $(s_i, a_{\mathcal{P}^i})$: $\pi_i^k(a_i|s_i, a_{\mathcal{P}^i}) \ge \nu_\lambda$ for all k.

The convergence analysis requires the same advantage gap structure as the full-observation case (Assumption 2), now stated in terms of the marginalized advantage $\bar{A}_i$ as defined in Assumption 4 below. This is the natural analogue: since the local-state policy can only distinguish actions through $\bar{A}_i$, the relevant gap is between the best and second-best actions under this marginalized quantity.

Assumption 4 (Marginalized advantage gap)**.** The marginalized advantage gap, as defined below, is positive:

$$\bar{\delta} := \inf_\pi \min_{i,s_i,a_{\mathcal{P}^i}} \left[\bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, \bar{a}_{i_p}) - \bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, \bar{a}_{i_q})\right] > 0$$

where $\bar{a}_{i_p}$ and $\bar{a}_{i_q}$ denote the best and second-best actions in terms of the marginalized advantage $\bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, \cdot)$.

Applying the same potential ascent and bridge arguments as in Theorem 3, with $\bar{A}_i$ in place of A_i and $\bar{\delta}$ in place of δ, we obtain the following:

Theorem 4. *Under Assumptions 1 and 4, for η as in Theorem 3:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \le \underbrace{\frac{2M\phi_{\max}}{K\nu_\lambda^\kappa(1-\gamma)}\left(1 + \frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right) + R_\lambda}_{\text{same structure as Theorem 3}} + \underbrace{\frac{4Mn\epsilon_{\text{po}}}{\nu_\lambda^\kappa(1-\gamma)}\left(1 + \frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right)}_{\text{partial observability bias}}.$$

Theorem 4 preserves the $O(1/K)$ convergence structure of the full-observation result (Theorem 3), with an additive ϵ_{po}-dependent term that captures the cost of partial observability. When $\epsilon_{\text{po}} = 0$ (e.g., factored rewards and transitions), the bound reduces exactly to Theorem 3. For fully correlated BN policies, the Nash-to-optimality transfer (Lemma 9) applies similarly, with an additional irreducible $\frac{n\epsilon_{\text{po}}}{1-\gamma}$ floor from the mismatch between local and global advantages. Straightforwardly, Corollary 2 in Appendix D.7 extends Corollary 1 to this partial observability setting.

6 Conclusion

We established $O(1/K)$ convergence of natural policy gradient under softmax parameterization with Bayesian network policies in Markov potential games, matching the best known rate for product policies while enabling richer correlation structures. For fully correlated BN policies, we proved that every near-Nash policy is also near-optimal, recovering the global optimality guarantee of single-agent NPG. A key technical ingredient is a log-barrier regularizer, weighted by the frequency of encountering each parent-action profile under the current policy, which provably maintains exploration without a priori assumptions on the policy iterates. These results extend to partially observable settings with factored state spaces, where convergence bounds incur an additive term proportional to the maximum discrepancy between full-state and local-state advantages.

Acknowledgments

Qi Zhang acknowledges funding support from National Science Foundation (NSF) award 2544947 and NSF CAREER award 2544948.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

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

Dingyang Chen and Qi Zhang. Context-aware bayesian network actor-critic methods for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 5327–5350. PMLR, 2023.

Dingyang Chen, Jianing Ye, Zhenyu Zhang, Xiaolong Kuang, Xinyang Shen, Ozalp Ozer, Chongjie Zhang, and Qi Zhang. Correlated policy optimization in multi-agent subteams. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Tke3BVwUz6.

Filippos Christianos, Georgios Papoudakis, and Stefano V Albrecht. Pareto actor-critic for equilibrium selection in multi-agent reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=3AzqYa18ah.

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

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

Roy Fox, Stephen M. Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4414–4425. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/fox22a.html.

David Heckerman. A tutorial on learning with bayesian networks. *CoRR*, abs/2002.00269, 2020. URL https://arxiv.org/abs/2002.00269.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Jingqing Ruan, Yali Du, Xuantang Xiong, Dengpeng Xing, Xiyun Li, Linghui Meng, Haifeng Zhang, Jun Wang, and Bo Xu. Gcs: Graph-based coordination strategy for multi-agent reinforcement learning. *arXiv preprint arXiv:2201.06257*, 2022.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Youbang Sun, Tao Liu, Ruida Zhou, PR Kumar, and Shahin Shahrampour. Provably fast convergence of independent natural policy gradient for markov potential games. *Advances in neural information processing systems*, 36:43951–43971, 2023.

Jianing Ye, Chenghao Li, Jianhao Wang, and Chongjie Zhang. Towards global optimality in cooperative marl with the transformation and distillation framework, 2023. URL https://arxiv.org/abs/2207.11143.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1923–1935. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0cd4c8c7ba098b199242c6634f43f653-Paper-Conference.pdf.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Worked Example: Fisher Pseudo-Inverse Cancellation for $n = 3$

This appendix illustrates, on a concrete small example, the two facts that drive the closed-form updates in Lemmas 1 and 2: (i) the augmented visitation factor $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$ cancels between the gradient and the Fisher pseudo-inverse, and (ii) the $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$-weighting of the log-barrier is exactly what keeps this cancellation intact, whereas an unweighted barrier reintroduces the uncontrollable parent-marginal factor $1/\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)$.

Setup. Consider $n = 3$ agents on the chain DAG $1 \to 2 \to 3$ (Figure 1), so $\mathcal{P}^1 = \emptyset$, $\mathcal{P}^2 = \{1\}$, and $\mathcal{P}^3 = \{2\}$. Let each action space be binary, $\mathcal{A}_i = \{0, 1\}$, so $|\mathcal{A}_i| = 2$. We focus on agent 2 at a fixed (s, a_1); the same computation applies verbatim to any agent and any parent profile, and to product policies by taking $\mathcal{P}^i = \emptyset$. Abbreviate the local policy as the vector

$$\pi_2^k(\cdot|s, a_1) = (p,\ 1-p), \qquad p := \pi_2^k(a_2 = 0 \mid s, a_1) \in (0, 1),$$

and write $A_2(a_2) := A_2^{\pi^k}(s, a_1, a_2)$ and $d := d_\rho^{\pi^k}(s, a_1) = d_\rho^{\pi^k}(s)\,\pi_1^k(a_1|s)$ for the augmented visitation weight.

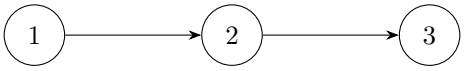

Figure 1: The $n = 3$ chain DAG G used in the worked example. Agent i conditions its action on the actions of its parents $\mathcal{P}^i$: $\mathcal{P}^1 = \emptyset$, $\mathcal{P}^2 = \{1\}$, $\mathcal{P}^3 = \{2\}$. The maximum in-degree is $\kappa = 1$.

Fisher block and its null space. For agent 2, the Fisher information matrix is block-diagonal across (s, a_1) pairs, with the block at (s, a_1) equal to $d\,F_{2,s,a_1}$, where

$$F_{2,s,a_1} = \mathrm{diag}(p, 1-p) - (p, 1-p)^\top (p, 1-p) = p(1-p)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Its null space is $\mathrm{span}\{\mathbf{1}\}$ with $\mathbf{1} = (1, 1)^\top$. The pseudo-inverse of the full block is $\frac{1}{d}F_{2,s,a_1}^\dagger$.

Gradients lie in the range of the Fisher block. By the gradient formula in the proof of Lemma 1, the potential and (weighted) log-barrier gradients with respect to $\theta_2(s, a_1, \cdot)$ are

$$g^\Phi = \frac{d}{1-\gamma}\big(p\,A_2(0),\ (1-p)\,A_2(1)\big), \qquad g^{\mathrm{lb}} = \lambda d\,\big(1-2p,\ 2p-1\big).$$

Both lie in the range of F_{2,s,a_1} (i.e., are orthogonal to $\mathbf{1}$): $\sum_{a_2} g^\Phi(a_2) = \frac{d}{1-\gamma}\big(pA_2(0) + (1-p)A_2(1)\big) = 0$ because the advantage is centered under π_2^k, and $\sum_{a_2} g^{\mathrm{lb}}(a_2) = \lambda d\big((1-2p) + (2p-1)\big) = 0$.

Applying the pseudo-inverse: the d factor cancels. For any g with $\sum_{a_2} g(a_2) = 0$, one verifies directly that $x(a_2) := g(a_2)/\pi_2^k(a_2) + c$ solves $F_{2,s,a_1}\,x = g$ for any constant c:

$$(F_{2,s,a_1}\,x)(a_2) = \pi_2^k(a_2)\,x(a_2) - \pi_2^k(a_2)\langle\pi_2^k, x\rangle = g(a_2) + \pi_2^k(a_2)c - \pi_2^k(a_2)\big(\underbrace{\textstyle\sum_{a_2'} g(a_2')}_{=0} + c\big) = g(a_2).$$

Hence $[F^{\dagger}_{2,s,a_1} g](a_2) = g(a_2)/\pi_2^k(a_2) + c'$ for a constant c' (fixed by orthogonality to $\mathbf{1}$). Applying $\frac{1}{d} F^{\dagger}_{2,s,a_1}$ to the total gradient $g = g^{\Phi} + g^{\mathrm{lb}}$ and using $g(a_2)/\pi_2^k(a_2) = d\big[\frac{A_2(a_2)}{1-\gamma} + \lambda(\frac{1}{\pi_2^k(a_2)} - 2)\big]$:

$$\left[\tfrac{1}{d} F^{\dagger}_{2,s,a_1}\, g\right](a_2) = \frac{A_2(a_2)}{1-\gamma} + \frac{\lambda}{\pi_2^k(a_2)} - \lambda\,|\mathcal{A}_2| + \frac{c'}{d}.$$

The augmented visitation weight $d = d_\rho^{\pi^k}(s, a_1)$ cancels completely (it multiplies the gradient and divides through the pseudo-inverse), the term $-\lambda|\mathcal{A}_2| = -2\lambda$ matches Lemma 2, and the constant c'/d is independent of a_2 and so cancels after the softmax normalization in $\theta_2^{k+1} = \theta_2^k + \eta[\cdots]$. This recovers exactly the regularized update (5); dropping the λ terms recovers the unregularized update (4).

Why the weighting matters. Suppose instead one uses the *unweighted* barrier $\lambda \sum_{s,a_2} \log \pi_{\theta_2}(a_2|s, a_1)$, whose gradient is $g^{\mathrm{lb,unw}} = \lambda(1-2p,\ 2p-1)$ *without* the factor d. The potential part is unchanged and its own d still cancels, but the barrier part now gives

$$\left[\tfrac{1}{d} F^{\dagger}_{2,s,a_1}\, g^{\mathrm{lb,unw}}\right](a_2) = \frac{1}{d}\left(\frac{\lambda}{\pi_2^k(a_2)} - 2\lambda\right) = \frac{1}{d_\rho^{\pi^k}(s)\,\pi_1^k(a_1|s)}\left(\frac{\lambda}{\pi_2^k(a_2)} - 2\lambda\right),$$

which carries the parent-marginal factor $1/\pi_1^k(a_1|s) = 1/\pi_{\mathcal{P}^2}(a_1|s)$ that can be arbitrarily small. This is precisely the uncontrollable term discussed in Section 4.2; weighting the barrier by $d_\rho^{\pi^k}(s, a_1)$ removes it.

B Proof of Unregularized NPG

B.1 Proof of Lemma 1

Proof. We first give explict forms of the gradient and Fisher information matrix. By Lemma 5.1 in Chen & Zhang (2023), the gradient of the objective function for BN policy is:

$$\frac{\partial L(\theta)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = \frac{1}{1-\gamma} d_\theta(s, a_{\mathcal{P}^i}) \pi_{\theta_i}(a_i|s, a_{\mathcal{P}^i}) A_i(s, a_{\mathcal{P}^i}, a_i).$$

The Fisher information matrix is block-diagonal, with blocks indexed by $(s, a_{\mathcal{P}^i})$:

$$F_i(\theta) = \mathrm{blkdiag}\left\{d_\theta(s, a_{\mathcal{P}^i}) F_{i,s,a_{\mathcal{P}^i}}(\theta_i)\right\}_{s,a_{\mathcal{P}^i}},$$

where each block is

$$F_{i,s,a_{\mathcal{P}^i}}(\theta_i) = \mathrm{diag}\{\pi_{\theta_i}(\cdot|s, a_{\mathcal{P}^i})\} - \pi_{\theta_i}(\cdot|s, a_{\mathcal{P}^i})\pi_{\theta_i}(\cdot|s, a_{\mathcal{P}^i})^\top.$$

For brevity, we will write $F_{i,s,a_{\mathcal{P}^i}}$ as shorthand for $F_{i,s,a_{\mathcal{P}^i}}(\theta_i)$.

It suffices to compute for each block $(s, a_{\mathcal{P}^i})$. Given block $(s, a_{\mathcal{P}^i})$, the update direction as given in (3) is therefore:

$$F^{\dagger}_{i,s,a_{\mathcal{P}^i}} \frac{\partial L(\theta)}{\partial \theta_i(s, a_{\mathcal{P}^i}, \cdot)} = F^{\dagger}_{i,s,a_{\mathcal{P}^i}} \left[\frac{1}{1-\gamma} d_\theta(s, a_{\mathcal{P}^i}) \pi_{\theta_i}(\cdot|s, a_{\mathcal{P}^i}) \odot A_i(s, a_{\mathcal{P}^i}, \cdot)\right]$$

where $\odot$ denotes element-wise multiplication. There are some facts about $F_{i,s,a_{\mathcal{P}^i}}$ and its pseudo-inverse $F^{\dagger}_{i,s,a_{\mathcal{P}^i}}$ that we will use:

- $F_{i,s,a_{\mathcal{P}^i}}$ is symmetric and positive semi-definite, with a one-dimensional null space spanned by the all-ones vector $\mathbf{1}$.

- For any symmetric matrix M, $M^\dagger M$ is the orthogonal projector onto the range of M. Since $F_{i,s,a_{\mathcal{P}^i}}$ is symmetric, $F^\dagger_{i,s,a_{\mathcal{P}^i}} F_{i,s,a_{\mathcal{P}^i}}$ is the orthogonal projector onto the range of $F_{i,s,a_{\mathcal{P}^i}}$, which is the orthogonal complement of the null space of $F_{i,s,a_{\mathcal{P}^i}}$ (i.e., the space orthogonal to $\mathbf{1}$). Therefore,

$$F^\dagger_{i,s,a_{\mathcal{P}^i}} F_{i,s,a_{\mathcal{P}^i}} = I - \mathrm{Proj}_{\mathrm{span}(\mathbf{1})} = I - \frac{1}{|\mathcal{A}_i|}\mathbf{1}\mathbf{1}^\top.$$

Now, let $f(a_i) := \frac{A_i(s,a_{\mathcal{P}^i},a_i)}{(1-\gamma)\pi_{\theta_i}(a_i|s,a_{\mathcal{P}^i})}$. Then:

$$\begin{aligned} F_{i,s,a_{\mathcal{P}^i}} f &= \mathrm{diag}\{\pi_{\theta_i}(\cdot|s,a_{\mathcal{P}^i})\} f - \pi_{\theta_i}(\cdot|s,a_{\mathcal{P}^i})\langle \pi_{\theta_i}(\cdot|s,a_{\mathcal{P}^i}), f\rangle \\ &= \frac{1}{1-\gamma} A_i(s,a_{\mathcal{P}^i},\cdot) - \pi_{\theta_i}(\cdot|s,a_{\mathcal{P}^i})\cdot 0 \\ &= \frac{1}{1-\gamma} A_i(s,a_{\mathcal{P}^i},\cdot), \end{aligned}$$

where we used $\sum_{a_i} \pi_{\theta_i}(a_i|s,a_{\mathcal{P}^i}) A_i(s,a_{\mathcal{P}^i},a_i) = 0$.

Therefore:

$$\begin{aligned} F^\dagger_{i,s,a_{\mathcal{P}^i}} \left[\frac{1}{1-\gamma} A_i(s,a_{\mathcal{P}^i},\cdot)\right] &= F^\dagger_{i,s,a_{\mathcal{P}^i}} F_{i,s,a_{\mathcal{P}^i}} f \\ &= \left(I - \frac{1}{|\mathcal{A}_i|}\mathbf{1}\mathbf{1}^\top\right) f \\ &= f - \bar{f}\mathbf{1}, \end{aligned}$$

where $\bar{f} = \frac{1}{|\mathcal{A}_i|}\sum_{a_i} f(a_i)$ is a constant depending only on $(s, a_{\mathcal{P}^i})$.

Multiplying back by $d_\theta(s, a_{\mathcal{P}^i})$ and converting back to the original coordinates:

$$\left[F_i(\theta)^\dagger \nabla_{\theta_i} L(\theta)\right]_{(s,a_{\mathcal{P}^i},a_i)} = \frac{A_i(s,a_{\mathcal{P}^i},a_i)}{1-\gamma} + c(s,a_{\mathcal{P}^i}),$$

where $c(s,a_{\mathcal{P}^i}) = -\bar{f}$ depends only on $(s,a_{\mathcal{P}^i})$ but not on a_i.

Therefore, the NPG update becomes:

$$\theta_i^{k+1}(s,a_{\mathcal{P}^i},a_i) = \theta_i^k(s,a_{\mathcal{P}^i},a_i) + \eta\left(\frac{A_i^k(s,a_{\mathcal{P}^i},a_i)}{1-\gamma} + c(s,a_{\mathcal{P}^i})\right).$$

Taking exponentials and normalizing:

$$\begin{aligned} \pi_i^{k+1}(a_i|s,a_{\mathcal{P}^i}) &= \frac{\exp(\theta_i^{k+1}(s,a_{\mathcal{P}^i},a_i))}{\sum_{a_i'}\exp(\theta_i^{k+1}(s,a_{\mathcal{P}^i},a_i'))} \\ &= \frac{\pi_i^k(a_i|s,a_{\mathcal{P}^i})\exp\left(\eta\frac{A_i^k(s,a_{\mathcal{P}^i},a_i)}{1-\gamma}\right)}{\sum_{a_i'}\pi_i^k(a_i'|s,a_{\mathcal{P}^i})\exp\left(\eta\frac{A_i^k(s,a_{\mathcal{P}^i},a_i')}{1-\gamma}\right)}, \end{aligned}$$

where the $\exp(\eta c(s,a_{\mathcal{P}^i}))$ terms cancel in the numerator and denominator. □

B.2 Lemma 4

Lemma 4 (BN Policy MPG Property)**.** Define $h_i(s,a) := r_i(s,a) - \phi(s,a)$, which implies $V_i^\pi(s) = \Phi^\pi(s) + V_{h_i}^\pi(s)$. Define $A_{h_i}^\pi(s,a_{\mathcal{P}^i},a_i)$ with respect to h_i analogously to the advantage function. We then have

$$\sum_{a_i}(\pi_i'(a_i|s,a_{\mathcal{P}^i}) - \pi_i(a_i|s,a_{\mathcal{P}^i}))A_{h_i}^\pi(s,a_{\mathcal{P}^i},a_i) = 0,$$

for all $s \in \mathcal{S}$, $a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i}$, $i \in [n]$, $\pi_i', \pi_i \in \Delta(\mathcal{A}_i)$.

Proof. By the definition of Markov potential game, for any μ, π_i', π_i, and π_{-i}:

$$V_i^{\pi_i',\pi_{-i}}(\mu) - V_i^{\pi_i,\pi_{-i}}(\mu) = \Phi^{\pi_i',\pi_{-i}}(\mu) - \Phi^{\pi_i,\pi_{-i}}(\mu).$$

It follows that

$$\begin{aligned} 0 &= V_{h_i}^{\pi_i',\pi_{-i}}(\mu) - V_{h_i}^{\pi_i,\pi_{-i}}(\mu) \\ &= \frac{1}{1-\gamma}\sum_s d_\mu^{\pi_i',\pi_{-i}}(s)\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)\sum_{a_i}(\pi_i'(a_i|s,a_{\mathcal{P}^i}) - \pi_i(a_i|s,a_{\mathcal{P}^i}))A_{h_i}^{\pi}(s,a_{\mathcal{P}^i},a_i). \end{aligned}$$

Since this holds for any initial distribution $\mu \in \Delta(\mathcal{S})$, including point masses on each state, we have for each s:

$$0 = \sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)\sum_{a_i}(\pi_i'(a_i|s,a_{\mathcal{P}^i}) - \pi_i(a_i|s,a_{\mathcal{P}^i}))A_{h_i}^{\pi}(s,a_{\mathcal{P}^i},a_i).$$

Since this holds for any choice of π_{-i} (and hence any distribution $\pi_{\mathcal{P}^i}(\cdot|s) \in \Delta(\mathcal{A}_{\mathcal{P}^i})$), we conclude:

$$0 = \sum_{a_i}(\pi_i'(a_i|s,a_{\mathcal{P}^i}) - \pi_i(a_i|s,a_{\mathcal{P}^i}))A_{h_i}^{\pi}(s,a_{\mathcal{P}^i},a_i), \quad \forall (s,a_{\mathcal{P}^i}).$$

□

B.3 Lemma 5

Lemma 5 (Potential Function Ascent for BN Policy)**.** Given policy π^k and advantage function $A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)$, for π^{k+1} generated using BN-NPG update (4), we have

$$\begin{aligned} &\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\ &\geq \left(\tfrac{1}{1-\gamma} - \tfrac{\sqrt{n}(1+A_{\max})\phi_{\max}\eta}{(1-\gamma)^3}\right)\textstyle\sum_s d_\rho^{\pi^{k+1}}(s)\sum_{i=1}^n\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\langle\pi_i^{k+1}(\cdot|s,a_{\mathcal{P}^i}), A_i^{\pi^k}(s,a_{\mathcal{P}^i},\cdot)\rangle. \end{aligned}$$

Proof. Without loss of generality, assume agents are indexed according to a topological ordering of the DAG G, i.e., if $j \in \mathcal{P}^i$, then $j < i$.

We introduce the intermediate policy:

$$\tilde{\pi}_{-i}^k(a_{-i}|s,a_i) := \prod_{j<i}\pi_j^{k+1}(a_j|s,a_{\mathcal{P}^j})\prod_{l>i}\pi_l^k(a_l|s,a_{\mathcal{P}^l}),$$

where all parent actions $a_{\mathcal{P}^j}$ and $a_{\mathcal{P}^l}$ are extracted from the joint action (a_i,a_{-i}). Under topological ordering:

- For $j < i$: $\mathcal{P}^j \subseteq \{1,\dots,j-1\} \subset \{1,\dots,i-1\}$, so $a_{\mathcal{P}^j}$ is determined by actions of agents in $\{1,\dots,j-1\}$, which are all contained in a_{-i} (since $j < i$).
- For $l > i$: $\mathcal{P}^l \subseteq \{1,\dots,l-1\}$, so $a_{\mathcal{P}^l}$ may include agent i's action (from a_i) and actions from agents in $\{1,\dots,i-1,i+1,\dots,l-1\}$ (contained in a_{-i}).

Thus $\tilde{\pi}_{-i}^k(\cdot|s,a_i)$ is well-defined as a distribution over a_{-i} conditioned on (s,a_i).

Let $\bar{A}_{i,\phi}^{\tilde{\pi}^k}(s,a_{\mathcal{P}^i},a_i)$ denote the marginalized advantage:

$$\begin{aligned} \bar{A}_{i,\phi}^{\tilde{\pi}^k}(s,a_{\mathcal{P}^i},a_i) &= \sum_{a_{-\mathcal{P}_+^i}}\tilde{\pi}_{-i}^k(a_{-\mathcal{P}_+^i}|s,a_{\mathcal{P}^i},a_i)A_\phi^{\pi^k}(s,a_{\mathcal{P}^i},a_i,a_{-\mathcal{P}_+^i}) \\ &= \frac{1}{\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)}\sum_{a_{-\mathcal{P}_+^i}}\tilde{\pi}_{-i}^k(a_{\mathcal{P}^i},a_{-\mathcal{P}_+^i}|s,a_i)A_\phi^{\pi^k}(s,a_{\mathcal{P}^i},a_i,a_{-\mathcal{P}_+^i}), \end{aligned}$$

where $a_{-\mathcal{P}^i_+}$ denotes the actions of agents not in $\mathcal{P}^i_+ := \mathcal{P}^i \cup \{i\}$, and

$$\tilde{\pi}^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i) = \frac{\tilde{\pi}^k_{-i}(a_{\mathcal{P}^i}, a_{-\mathcal{P}^i_+}|s, a_i)}{\sum_{a_{-\mathcal{P}^i_+}} \tilde{\pi}^k_{-i}(a_{\mathcal{P}^i}, a_{-\mathcal{P}^i_+}|s, a_i)} = \frac{\tilde{\pi}^k_{-i}(a_{\mathcal{P}^i}, a_{-\mathcal{P}^i_+}|s, a_i)}{\pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)}.$$

By the performance difference lemma:

$$\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) = \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_a (\pi^{k+1}(a|s) - \pi^k(a|s)) A^{\pi^k}_\phi(s, a).$$

Telescoping decomposition:

$$\begin{aligned}
&\pi^{k+1}(a|s) - \pi^k(a|s) \\
=&\sum_{i=1}^n \left[\prod_{j=1}^i \pi^{k+1}_j(a_j|s, a_{\mathcal{P}^j}) \prod_{l=i+1}^n \pi^k_l(a_l|s, a_{\mathcal{P}^l}) - \prod_{j=1}^{i-1} \pi^{k+1}_j(a_j|s, a_{\mathcal{P}^j}) \prod_{l=i}^n \pi^k_l(a_l|s, a_{\mathcal{P}^l}) \right] \\
=&\sum_{i=1}^n \left[\prod_{j<i} \pi^{k+1}_j(a_j|s, a_{\mathcal{P}^j}) \prod_{l>i} \pi^k_l(a_l|s, a_{\mathcal{P}^l}) \right] (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) \\
=&\sum_{i=1}^n \tilde{\pi}^k_{-i}(a_{-i}|s, a_i)(\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})).
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\
=& \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_a \sum_{i=1}^n \tilde{\pi}^k_{-i}(a_{-i}|s, a_i)(\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) A^{\pi^k}_\phi(s, a) \\
=& \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \sum_{a_i} \sum_{a_{-\mathcal{P}^i_+}} \tilde{\pi}^k_{-i}(a_{\mathcal{P}^i}, a_{-\mathcal{P}^i_+}|s, a_i) \\
&\quad \times (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) A^{\pi^k}_\phi(s, a_{\mathcal{P}^i}, a_i, a_{-\mathcal{P}^i_+}) \\
=& \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) \\
&\quad \times \underbrace{\sum_{a_{-\mathcal{P}^i_+}} \tilde{\pi}^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i) A^{\pi^k}_\phi(s, a_{\mathcal{P}^i}, a_i, a_{-\mathcal{P}^i_+})}_{= \bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i)} \\
=& \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) \bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) \\
=& \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) \\
&- \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) A^{\pi^k}_{h_i}(s, a_{\mathcal{P}^i}, a_i) \\
&+ \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} (\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})) \\
&\quad \times (\bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) - A^{\pi^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i)).
\end{aligned}$$

The decomposition $\bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) = A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) - A^{\pi^k}_{h_i}(s, a_{\mathcal{P}^i}, a_i) + (\bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) - A^{\pi^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i))$ follows from $A_i = A_{i,\phi} + A_{h_i}$ and $\bar{A}^{\pi^k}_{i,\phi} = A^{\pi^k}_{i,\phi}$.

By Lemma 4, the second term vanishes:

$$\sum_{a_i}(\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i}))A^{\pi^k}_{h_i}(s, a_{\mathcal{P}^i}, a_i) = 0.$$

For the third term, we bound the error. Since:

$$\begin{aligned}
&|\bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) - A^{\pi^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i)| \\
&= \left| \sum_{a_{-\mathcal{P}^i_+}} \tilde{\pi}^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i) A^{\pi^k}_{\phi}(s, a_{\mathcal{P}^i}, a_i, a_{-\mathcal{P}^i_+}) \right. \\
&\qquad \left. - \sum_{a_{-\mathcal{P}^i_+}} \pi^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i) A^{\pi^k}_{\phi}(s, a_{\mathcal{P}^i}, a_i, a_{-\mathcal{P}^i_+}) \right| \\
&= \left| \sum_{a_{-\mathcal{P}^i_+}} (\tilde{\pi}^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(a_{-\mathcal{P}^i_+}|s, a_{\mathcal{P}^i}, a_i)) A^{\pi^k}_{\phi}(s, a_{\mathcal{P}^i}, a_i, a_{-\mathcal{P}^i_+}) \right| \\
&\leq \frac{\phi_{\max}}{1-\gamma} \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1,
\end{aligned}$$

we have:

$$\begin{aligned}
&\left| \sum_{a_i} (\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i}))(\bar{A}^{\tilde{\pi}^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i) - A^{\pi^k}_{i,\phi}(s, a_{\mathcal{P}^i}, a_i)) \right| \\
&\leq \frac{\phi_{\max}}{1-\gamma} \sum_{a_i} |\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i})| \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\
&\geq \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) \\
&\quad - \frac{\phi_{\max}}{(1-\gamma)^2} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \\
&\qquad \times \sum_{a_i} |\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i})| \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1.
\end{aligned}$$

Since

$$\begin{aligned}
&\sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} |\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i})| \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1 \\
&\leq \frac{\sqrt{n}}{2} \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \|\pi_i^{k+1}(\cdot|s, a_{\mathcal{P}^i}) - \pi_i^k(\cdot|s, a_{\mathcal{P}^i})\|_1^2 \\
&\quad + \frac{1}{2\sqrt{n}} \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1^2,
\end{aligned}$$

where the inequality follows by first applying Cauchy–Schwarz on the sum over a_i:

$$\begin{aligned}&\sum_{a_i} |\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s, a_{\mathcal{P}^i})| \|\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1 \\ \leq &\|\pi_i^{k+1}(\cdot|s, a_{\mathcal{P}^i}) - \pi_i^k(\cdot|s, a_{\mathcal{P}^i})\|_1 \cdot b_{i,s,a_{\mathcal{P}^i}},\end{aligned}$$

with $b_{i,s,a_{\mathcal{P}^i}} := \sqrt{\sum_{a_i} \|\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1^2}$ (using $\|v\|_2 \leq \|v\|_1$), and then applying Young's inequality with rescaling $n^{1/4}$:

$$a \cdot b = (n^{1/4}a)(n^{-1/4}b) \leq \frac{\sqrt{n}}{2}a^2 + \frac{1}{2\sqrt{n}}b^2.$$

We now bound the two terms T_1 and T_2 separately. Define the shorthand

$$W_j(s) := \sum_{a_{\mathcal{P}^j}} \pi_{\mathcal{P}^j}^{k+1}(a_{\mathcal{P}^j}|s) \langle \pi_j^{k+1}(\cdot|s, a_{\mathcal{P}^j}),\, A_j^{\pi^k}(s, a_{\mathcal{P}^j}, \cdot)\rangle, \quad W(s) := \sum_{j=1}^n W_j(s).$$

Bounding T_1. By Pinsker's inequality $\|p - q\|_1^2 \leq 2\,\mathrm{KL}(p\|q)$:

$$\begin{aligned}T_1 &:= \frac{\sqrt{n}}{2} \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \|\pi_i^{k+1}(\cdot|s, a_{\mathcal{P}^i}) - \pi_i^k(\cdot|s, a_{\mathcal{P}^i})\|_1^2 \\ &\leq \sqrt{n} \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\, \mathrm{KL}(\pi_i^{k+1}(\cdot|s, a_{\mathcal{P}^i}) \| \pi_i^k(\cdot|s, a_{\mathcal{P}^i})).\end{aligned}$$

From the BN-NPG update (4), $\log \frac{\pi_i^{k+1}(a_i|s,a_{\mathcal{P}^i})}{\pi_i^k(a_i|s,a_{\mathcal{P}^i})} = \frac{\eta A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)}{1-\gamma} + c(s, a_{\mathcal{P}^i})$ where $c(s, a_{\mathcal{P}^i}) = -\log Z(s, a_{\mathcal{P}^i}) \leq 0$. Therefore:

$$\mathrm{KL}(\pi_i^{k+1}\|\pi_i^k) = \frac{\eta}{1-\gamma}\langle \pi_i^{k+1}, A_i^{\pi^k}\rangle + c(s, a_{\mathcal{P}^i}) \leq \frac{\eta}{1-\gamma}\langle \pi_i^{k+1}, A_i^{\pi^k}\rangle,$$

giving $T_1 \leq \frac{\sqrt{n}\,\eta}{1-\gamma} W(s)$.

Bounding T_2. By Pinsker's inequality, for each $(i, s, a_{\mathcal{P}^i}, a_i)$:

$$\|\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1^2 \leq 2\,\mathrm{KL}(\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) \| \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)).$$

By the chain rule of KL divergence for BN distributions, since agents $l > i$ use the same policy π_l^k in both $\tilde{\pi}_{-i}^k$ and π_{-i}^k (contributing zero KL):

$$\mathrm{KL}(\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) \| \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)) = \sum_{j<i} \sum_{a_{\mathcal{P}^j}} \tilde{\mu}_j(a_{\mathcal{P}^j}|s)\, \mathrm{KL}(\pi_j^{k+1}(\cdot|s, a_{\mathcal{P}^j}) \| \pi_j^k(\cdot|s, a_{\mathcal{P}^j})),$$

where $\tilde{\mu}_j(a_{\mathcal{P}^j}|s)$ is the marginal over $a_{\mathcal{P}^j}$ induced by agents $\{1, \ldots, j-1\}$ using π^{k+1}, conditioned on s. Note that under topological ordering, $\mathcal{P}^j \subseteq \{1, \ldots, j-1\}$ for $j < i$, so agent i's action does not appear in any parent set $\mathcal{P}^{j'}$ for $j' \leq j < i$; hence $\tilde{\mu}_j$ does not depend on a_i or $a_{\mathcal{P}^i}$.

Since the KL does not depend on a_i, summing over a_i gives a factor of $|\mathcal{A}_i|$:

$$\sum_{a_i} \|\tilde{\pi}_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi_{-i}^k(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1^2 \leq 2|\mathcal{A}_i| \sum_{j<i} \sum_{a_{\mathcal{P}^j}} \tilde{\mu}_j(a_{\mathcal{P}^j}|s)\, \mathrm{KL}(\pi_j^{k+1}(\cdot|s, a_{\mathcal{P}^j}) \| \pi_j^k(\cdot|s, a_{\mathcal{P}^j})).$$

Using $\mathrm{KL}(\pi_j^{k+1}\|\pi_j^k) \leq \frac{\eta}{1-\gamma}\langle \pi_j^{k+1}, A_j^{\pi^k}\rangle$, multiplying by $\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)$, and summing over $a_{\mathcal{P}^i}$, the marginal identity

$$\sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\, \tilde{\mu}_j(a_{\mathcal{P}^j}|s) = \pi_{\mathcal{P}^j}^{k+1}(a_{\mathcal{P}^j}|s)$$

holds because $\mathcal{P}^j \subseteq \{1,\dots,j-1\} \subset \{1,\dots,i-1\}$ and all agents $\{1,\dots,i-1\}$ use π^{k+1} in both $\tilde{\pi}^k_{-i}$ and π^{k+1}, so marginalizing the joint distribution of agents $\{1,\dots,i-1\}$ under π^{k+1} to $\mathcal{P}^j$ yields $\pi^{k+1}_{\mathcal{P}^j}$. Therefore:

$$
\begin{aligned}
T_2 &:= \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1^2 \\
&\le \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \frac{2|\mathcal{A}_i|\eta}{1-\gamma} \sum_{j<i} \sum_{a_{\mathcal{P}^j}} \pi^{k+1}_{\mathcal{P}^j}(a_{\mathcal{P}^j}|s) \langle \pi^{k+1}_j(\cdot|s, a_{\mathcal{P}^j}),\, A^{\pi^k}_j(s, a_{\mathcal{P}^j}, \cdot)\rangle \\
&= \frac{\eta}{\sqrt{n}(1-\gamma)} \sum_{i=1}^{n} |\mathcal{A}_i| \sum_{j<i} W_j(s) \le \frac{A_{\max}\eta}{\sqrt{n}(1-\gamma)} \sum_{j=1}^{n-1} (n-j) W_j(s) \le \frac{\sqrt{n}\, A_{\max}\, \eta}{1-\gamma} W(s),
\end{aligned}
$$

where we used $\sum_{i=1}^{n} \sum_{j<i} W_j = \sum_{j=1}^{n-1} (n-j) W_j \le n \sum_j W_j = n\, W(s)$.

Combining $T_1 + T_2$:

$$
\begin{aligned}
&\sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} |\pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) - \pi^k_i(a_i|s, a_{\mathcal{P}^i})| \|\tilde{\pi}^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i) - \pi^k_{-i}(\cdot|s, a_{\mathcal{P}^i}, a_i)\|_1 \\
&\le T_1 + T_2 \le \frac{\sqrt{n}(1 + A_{\max})\eta}{1-\gamma} \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \langle \pi^{k+1}_i(\cdot|s, a_{\mathcal{P}^i}),\, A^{\pi^k}_i(s, a_{\mathcal{P}^i}, \cdot)\rangle.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
&\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\
&\ge \frac{1}{1-\gamma} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) \\
&\quad - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) \\
&= \left(\frac{1}{1-\gamma} - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \right) \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi^{k+1}_i(a_i|s, a_{\mathcal{P}^i}) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i) \\
&= \left(\frac{1}{1-\gamma} - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \right) \sum_s d^{\pi^{k+1}}_\rho(s) \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \langle \pi^{k+1}_i(\cdot|s, a_{\mathcal{P}^i}), A^{\pi^k}_i(s, a_{\mathcal{P}^i}, \cdot)\rangle.
\end{aligned}
$$

□

B.4 Lemma 6

Lemma 6 (Function f^k for BN Policy)**.** At iteration k, define

$$
f^k(\alpha) = \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi_{i,\alpha}(a_i|s, a_{\mathcal{P}^i}) A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i),
$$

where

$$
\pi_{i,\alpha}(a_i|s, a_{\mathcal{P}^i}) \propto \pi^k_i(a_i|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A^{\pi^k}_i(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} \right).
$$

Note that $f^k(\alpha)$ depends on s; we suppress this dependence for notational brevity since the subsequent analysis sums over s weighted by $d^{\pi^{k+1}}_\rho(s)$ and the bound in this lemma holds uniformly for all $s \in \mathcal{S}$.

Then, for any $\alpha > 0$,

$$f^k(\alpha) \geq f^k(\infty)\left[1 - \frac{1}{\nu\left(\exp\left(\frac{\delta\alpha}{1-\gamma}\right) - 1\right) + 1}\right],$$

where $a^k_{i_p} \in \operatorname{argmax}_{a_j \in \mathcal{A}_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j) =: \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})$, $a^k_{i_q} \in \operatorname{argmax}_{a_j \in \mathcal{A}_i \setminus \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)$, and $\delta > 0$ is defined in Assumption 2.

Proof. For any $i \in [n]$ and any $(s, a_{\mathcal{P}^i})$ pair, we have

$$\begin{aligned}
&\sum_{a_i} \pi_{i,\alpha}(a_i|s, a_{\mathcal{P}^i}) A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i) \\
=&\frac{\sum_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)\pi_i^k(a_i|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma}\right)}{\sum_{a_i} \pi_i^k(a_i|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma}\right)} \\
\geq&\frac{\sum_{a_j \in \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)\pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)}{1-\gamma}\right)}{\sum_{a_j \in \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} \pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)}{1-\gamma}\right) + \sum_{a_j \in \mathcal{A}_i \setminus \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} \pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_q})}{1-\gamma}\right)} \\
&+ \frac{\sum_{a_j \in \mathcal{A}_i \setminus \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)\pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_q})}{1-\gamma}\right)}{\sum_{a_j \in \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} \pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_j)}{1-\gamma}\right) + \sum_{a_j \in \mathcal{A}_i \setminus \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} \pi_i^k(a_j|s, a_{\mathcal{P}^i}) \exp\left(\frac{\alpha A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_q})}{1-\gamma}\right)} \\
=&A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_p}) - \frac{\sum_{a_i}\left(A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_p}) - A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)\right)\pi_i^k(a_i|s, a_{\mathcal{P}^i})}{\left(\sum_{a_j \in \mathcal{A}^k_{i_p}(s, a_{\mathcal{P}^i})} \pi_i^k(a_j|s, a_{\mathcal{P}^i})\right)\left(\exp\left(\frac{\alpha\left(A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_p}) - A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a^k_{i_q})\right)}{1-\gamma}\right) - 1\right) + 1} \\
\geq&\max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)\left[1 - \frac{1}{\nu\left(\exp\left(\frac{\delta\alpha}{1-\gamma}\right) - 1\right) + 1}\right],
\end{aligned}$$

where the first inequality holds due to Lemma D.1 from the work Sun et al. (2023).

By taking sum over all agents $i \in [n]$ and all $(s, a_{\mathcal{P}^i})$ pairs weighted by $\pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s)$, we get

$$f^k(\alpha) \geq f^k(\infty)\left[1 - \frac{1}{\nu\left(\exp\left(\frac{\delta\alpha}{1-\gamma}\right) - 1\right) + 1}\right],$$

where $f^k(\infty) := \lim_{\alpha\to\infty} f^k(\alpha) = \sum_{i\in[n]}\sum_{a_{\mathcal{P}^i}} \pi^{k+1}_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)$. □

B.5 Proof of Theorem 1

Proof. Choose $\alpha = \eta = \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$ in Lemma 6. We have:

$$\begin{aligned}
\text{NE-gap}(\pi^k) &\le \frac{1}{1-\gamma}\sum_{i,s} d_\rho^{\pi_i^{*k},\pi_{-i}^k}(s)\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s)\max_{a_i} A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)\\
&\le \frac{M}{1-\gamma}\sum_{s} d_\rho^{\pi^{k+1}}(s)\sum_{i=1}^{n}\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s)\max_{a_i} A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)\\
&\le \frac{M}{\nu^\kappa(1-\gamma)}\sum_{s} d_\rho^{\pi^{k+1}}(s)\sum_{i=1}^{n}\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\max_{a_i} A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)\\
&\le \frac{M}{\nu^\kappa(1-\gamma)}\sum_{s} d_\rho^{\pi^{k+1}}(s)\lim_{\alpha\to\infty} f^k(\alpha)\\
&\le \frac{M}{\nu^\kappa(1-\gamma)}\sum_{s} d_\rho^{\pi^{k+1}}(s)\frac{f^k(\eta)}{1-\frac{1}{\nu(\exp(\frac{\delta\eta}{1-\gamma})-1)+1}}\\
&\le \frac{M}{\nu^\kappa(1-\gamma)}\sum_{s} d_\rho^{\pi^{k+1}}(s)\frac{f^k(\eta)}{1-\frac{1}{\nu\frac{\delta\eta}{1-\gamma}+1}}\\
&= \frac{M}{\nu^\kappa(1-\gamma)}\sum_{s} d_\rho^{\pi^{k+1}}(s) f^k(\eta)\left(1+\frac{1-\gamma}{\nu\delta\eta}\right).
\end{aligned}$$

In the third inequality we bound the parent marginal ratio termwise: for each agent i, $\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s) \le \nu^{-|\mathcal{P}^i|}\,\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \le \nu^{-\kappa}\,\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)$, using $\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \ge \nu^{|\mathcal{P}^i|}$ (each of the $|\mathcal{P}^i|$ parent conditionals is at least ν by Assumption 3), $|\mathcal{P}^i| \le \kappa$, and $\nu \le 1$. Since $\max_{a_i} A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i) \ge 0$ (advantages are centered under π^k, so the best action has nonnegative advantage), every summand is nonnegative and the common factor $\nu^{-\kappa}$ can be pulled out. By Lemma 5 with $\eta = \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$:

$$\begin{aligned}
\sum_s d_\rho^{\pi^{k+1}}(s) f^k(\eta) &= \sum_s d_\rho^{\pi^{k+1}}(s)\sum_{i=1}^{n}\sum_{a_{\mathcal{P}^i}}\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\langle \pi_i^{k+1}(\cdot|s,a_{\mathcal{P}^i}), A_i^{\pi^k}(s,a_{\mathcal{P}^i},\cdot)\rangle\\
&\le \frac{\Phi^{\pi^{k+1}}(\rho)-\Phi^{\pi^k}(\rho)}{\frac{1}{1-\gamma}-\frac{\sqrt{n}(1+A_{\max})\phi_{\max}\eta}{(1-\gamma)^3}} = 2(1-\gamma)(\Phi^{\pi^{k+1}}(\rho)-\Phi^{\pi^k}(\rho)).
\end{aligned}$$

Therefore:

$$\text{NE-gap}(\pi^k) \le 2M/\nu^\kappa\cdot(\Phi^{\pi^{k+1}}(\rho)-\Phi^{\pi^k}(\rho))\left(1+\frac{2\sqrt{n}(1+A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)}\right).$$

Summing over $k = 0,\dots,K-1$:

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\text{NE-gap}(\pi^k) &\le \frac{2M(\Phi^{\pi^K}(\rho)-\Phi^{\pi^0}(\rho))}{K\nu^\kappa}\left(1+\frac{2\sqrt{n}(1+A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)}\right)\\
&\le \frac{2M\phi_{\max}}{K\nu^\kappa(1-\gamma)}\left(1+\frac{2\sqrt{n}(1+A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)}\right).
\end{aligned}$$

□

B.6 Lemma 7

Lemma 7 (Advantage Bound for ϵ-Nash Policies in MPGs)**.** Under Assumptions 1 and 3, if π is an ϵ-Nash policy in a Markov potential game, then for all $i \in [n]$, $s \in \mathcal{S}$, $a_{\mathcal{P}^i} \in \mathcal{A}_{\mathcal{P}^i}$, and $a_i \in \mathcal{A}_i$:

$$A_i^\pi(s, a_{\mathcal{P}^i}, a_i) \leq \frac{M(1-\gamma)\epsilon}{\nu^{|\mathcal{P}^i|+1}}.$$

Proof. By the performance difference lemma for agent i's value function:

$$V_i^{\pi_i', \pi_{-i}}(\rho) - V_i^\pi(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_i', \pi_{-i}}(s) \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi_i'(a_i|s, a_{\mathcal{P}^i}) A_i^\pi(s, a_{\mathcal{P}^i}, a_i).$$

Since π is ϵ-Nash, by Definition 2, for all i and all π_i':

$$V_i^{\pi_i', \pi_{-i}}(\rho) - V_i^\pi(\rho) \leq \epsilon.$$

Fix any $(s, a_{\mathcal{P}^i}, a_i)$ and consider the deviation policy π_i' that places probability ν on action a_i at state-parent pair $(s, a_{\mathcal{P}^i})$ and distributes the remaining mass arbitrarily while maintaining $\pi_i'(\cdot|\tilde{s}, \tilde{a}_{\mathcal{P}^i}) \geq \nu$ for all $(\tilde{s}, \tilde{a}_{\mathcal{P}^i})$. By Assumption 1, $d_\rho^{\pi_i', \pi_{-i}}(s) \geq 1/M$. By Assumption 3, $\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \geq \nu^{|\mathcal{P}^i|}$. Therefore:

$$\epsilon \geq V_i^{\pi_i', \pi_{-i}}(\rho) - V_i^\pi(\rho) \geq \frac{\nu^{|\mathcal{P}^i|} \cdot \nu}{M(1-\gamma)} A_i^\pi(s, a_{\mathcal{P}^i}, a_i).$$

Rearranging gives $A_i^\pi(s, a_{\mathcal{P}^i}, a_i) \leq \frac{M(1-\gamma)\epsilon}{\nu^{|\mathcal{P}^i|+1}}$. □

B.7 Lemma 8

Lemma 8 (Q-function Recursion for Fully Connected BN)**.** For a fully connected BN with topological ordering $1, \ldots, n$, we have $\mathcal{P}_+^i = \mathcal{P}^{i+1}$ for all $i \in [n-1]$, and:

$$Q_{\phi,i}^\pi(s, a_{\mathcal{P}^i}) = Q_{\phi,i-1}^\pi(s, a_{\mathcal{P}_+^{i-1}}), \quad \forall i \geq 2,$$

where $Q_{\phi,i}^\pi(s, a_{\mathcal{P}^i}) := \sum_{a_{-\mathcal{P}^i}} \pi(a_{-\mathcal{P}^i}|s, a_{\mathcal{P}^i}) Q_\phi^\pi(s, a_{\mathcal{P}^i}, a_{-\mathcal{P}^i})$ is the marginalized Q-function for the potential.

Proof. Since $\mathcal{P}^i = \{1, \ldots, i-1\} = \mathcal{P}_+^{i-1}$:

$$\begin{aligned} Q_{\phi,i}^\pi(s, a_{\mathcal{P}^i}) &= \sum_{a_{-\mathcal{P}^i}} \pi(a_{-\mathcal{P}^i}|s, a_{\mathcal{P}^i}) Q_\phi^\pi(s, a_{\mathcal{P}^i}, a_{-\mathcal{P}^i}) \\ &= \sum_{a_{-\mathcal{P}_+^{i-1}}} \pi(a_{-\mathcal{P}_+^{i-1}}|s, a_{\mathcal{P}_+^{i-1}}) Q_\phi^\pi(s, a_{\mathcal{P}_+^{i-1}}, a_{-\mathcal{P}_+^{i-1}}) = Q_{\phi,i-1}^\pi(s, a_{\mathcal{P}_+^{i-1}}). \end{aligned}$$

□

B.8 Lemma 9

Lemma 9 (Nash-to-Optimality for Fully Connected BN in MPGs)**.** Under Assumptions 1 and 3, for a fully connected BN in a Markov potential game, if π is an ϵ-Nash policy, then π is an ϵ'-optimal policy with:

$$\epsilon' = \frac{M\epsilon(1-\nu^n)}{\nu^n(1-\nu)}.$$

Proof. By Lemma 7, for agent i with $|\mathcal{P}^i| = i - 1$ in a fully connected BN:

$$A_i^\pi(s, a_{\mathcal{P}^i}, a_i) \leq \frac{M(1-\gamma)\epsilon}{\nu^i}.$$

By the MPG definition, $V^i_{\pi'_i,\pi_{-i}}(\rho) - V^i_\pi(\rho) = \Phi_{\pi'_i,\pi_{-i}}(\rho) - \Phi_\pi(\rho)$ for any unilateral deviation. Define the marginalized potential advantage:

$$A^\pi_{\phi,i}(s, a_{\mathcal{P}^i}, a_i) := Q^\pi_{\phi,i}(s, a_{\mathcal{P}^i}, a_i) - Q^\pi_{\phi,i}(s, a_{\mathcal{P}^i}).$$

The MPG property implies $A^\pi_{\phi,i}(s, a_{\mathcal{P}^i}, a_i) = A^\pi_i(s, a_{\mathcal{P}^i}, a_i)$, so:

$$A^\pi_{\phi,i}(s, a_{\mathcal{P}^i}, a_i) \leq \frac{M(1-\gamma)\epsilon}{\nu^i}.$$

For any joint action $a = (a^1, \ldots, a^n)$, following reverse topological order and using Lemma 8:

$$\begin{aligned} Q^\pi_\phi(s,a) &= Q^\pi_{\phi,n}(s, a_{\mathcal{P}^n}, a_n) = Q^\pi_{\phi,n}(s, a_{\mathcal{P}^n}) + A^\pi_{\phi,n}(s, a_{\mathcal{P}^n}, a_n) \\ &\leq Q^\pi_{\phi,n-1}(s, a_{\mathcal{P}^{n-1}_+}) + \frac{M(1-\gamma)\epsilon}{\nu^n}. \end{aligned}$$

Continuing this recursion, at each step i we add $\frac{M(1-\gamma)\epsilon}{\nu^i}$. Noting $\mathcal{P}^1 = \emptyset$ so $Q^\pi_{\phi,1}(s, a_{\mathcal{P}^1}) = \Phi^\pi(s)$:

$$Q^\pi_\phi(s,a) \leq \Phi^\pi(s) + M(1-\gamma)\epsilon \sum_{i=1}^n \frac{1}{\nu^i} = \Phi^\pi(s) + M(1-\gamma)\epsilon \cdot \frac{1-\nu^n}{\nu^n(1-\nu)},$$

where we used the geometric series $\sum_{i=1}^n \nu^{-i} = \frac{1-\nu^n}{\nu^n(1-\nu)}$.

Let π^* be an optimal policy for the potential. By the performance difference lemma:

$$\begin{aligned} \Phi^{\pi^*}(\rho) - \Phi^\pi(\rho) &= \frac{1}{1-\gamma} \sum_s d^{\pi^*}_\rho(s) \sum_a \pi^*(a|s) \left(Q^\pi_\phi(s,a) - \Phi^\pi(s) \right) \\ &\leq \frac{1}{1-\gamma} \sum_s d^{\pi^*}_\rho(s) \cdot \frac{M(1-\gamma)\epsilon(1-\nu^n)}{\nu^n(1-\nu)} \\ &= \frac{M\epsilon(1-\nu^n)}{\nu^n(1-\nu)}, \end{aligned}$$

where we used $\sum_s d^{\pi^*}_\rho(s) = 1$. □

B.9 Proof of Theorem 2

Proof. By Theorem 1, since the fully connected BN has maximum in-degree $\kappa = n - 1$, we have

$$\frac{1}{K} \sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \leq \frac{2M\phi_{\max}}{K\nu^{n-1}(1-\gamma)} \left(1 + \frac{2\sqrt{n}(1 + A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)} \right) =: \bar{\epsilon}.$$

By Lemma 9, if π is an ϵ-Nash policy, then π is an $\frac{M(1-\nu^n)\epsilon}{\nu^n(1-\nu)}$-optimal policy. Therefore:

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \text{optimality-gap}(\pi^k) &\leq \frac{M(1-\nu^n)}{\nu^n(1-\nu)} \cdot \frac{1}{K} \sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \\ &\leq \frac{2M^2\phi_{\max}(1-\nu^n)}{\nu^{2n-1}(1-\nu)K(1-\gamma)} \left(1 + \frac{2\sqrt{n}(1 + A_{\max})\phi_{\max}}{\nu\delta(1-\gamma)} \right). \end{aligned}$$

□

C Proof of NPG with Log-Barrier Regularizer

C.1 Proof of Lemma 2

Proof. The gradient of the log-barrier regularized objective decomposes as:

$$\nabla_{\theta_i} L_\lambda^k(\theta) = \nabla_{\theta_i} \Phi^{\pi_\theta}(\rho) + \lambda \nabla_{\theta_i} \sum_{s, a_{\mathcal{P}^i}} d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \sum_{a_i} \log \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i}),$$

where $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$ is treated as a fixed weight (independent of θ).

From Lemma 1:

$$\frac{\partial \Phi^{\pi_\theta}(\rho)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s, a_{\mathcal{P}^i}) \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i}) A_i(s, a_{\mathcal{P}^i}, a_i).$$

For the log-barrier term, using the softmax parameterization (2):

$$\frac{\partial \log \pi_{\theta_i}(a_i' | s', a_{\mathcal{P}^i}')}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = \mathbb{1}\{(s', a_{\mathcal{P}^i}') = (s, a_{\mathcal{P}^i})\} \left(\mathbb{1}\{a_i' = a_i\} - \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i})\right).$$

Therefore:

$$\frac{\partial}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} \sum_{s', a_{\mathcal{P}^i}'} d_\rho^{\pi^k}(s', a_{\mathcal{P}^i}') \sum_{a_i'} \log \pi_{\theta_i}(a_i' | s', a_{\mathcal{P}^i}') = d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \left(1 - |\mathcal{A}_i| \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i})\right).$$

Combining both terms:

$$\frac{\partial L_\lambda^k(\theta)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = \frac{d_\rho^{\pi_\theta}(s, a_{\mathcal{P}^i}) \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i}) A_i(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \lambda d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \left(1 - |\mathcal{A}_i| \pi_{\theta_i}(a_i | s, a_{\mathcal{P}^i})\right).$$

Evaluating at $\theta = \theta^k$ (so that $d_\rho^{\pi_\theta}(s, a_{\mathcal{P}^i}) = d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$):

$$\frac{\partial L_\lambda^k(\theta^k)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \left[\frac{\pi_i^k(a_i | s, a_{\mathcal{P}^i}) A_i(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \lambda \left(1 - |\mathcal{A}_i| \pi_i^k(a_i | s, a_{\mathcal{P}^i})\right)\right].$$

One can verify that $\sum_{a_i} \frac{\partial L_\lambda^k(\theta^k)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)} = 0$ for each $(s, a_{\mathcal{P}^i})$, confirming the gradient lies in the range of the Fisher information matrix.

Applying the pseudo-inverse as in Lemma 1, for each block $(s, a_{\mathcal{P}^i})$. Let $g(s, a_{\mathcal{P}^i}, a_i) := \frac{\partial L_\lambda^k(\theta^k)}{\partial \theta_i(s, a_{\mathcal{P}^i}, a_i)}$. Since $\sum_{a_i} g(s, a_{\mathcal{P}^i}, a_i) = 0$, we have:

$$\begin{aligned}
\left[F_i(\theta^k)^\dagger \nabla_{\theta_i} L_\lambda^k(\theta^k)\right]_{(s, a_{\mathcal{P}^i}, a_i)} &= \frac{g(s, a_{\mathcal{P}^i}, a_i)}{d_\rho^{\pi^k}(s, a_{\mathcal{P}^i}) \pi_i^k(a_i | s, a_{\mathcal{P}^i})} + c(s, a_{\mathcal{P}^i}) \\
&= \frac{A_i(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\lambda}{\pi_i^k(a_i | s, a_{\mathcal{P}^i})} - \lambda |\mathcal{A}_i| + c(s, a_{\mathcal{P}^i}),
\end{aligned}$$

where $c(s, a_{\mathcal{P}^i})$ is a constant depending only on $(s, a_{\mathcal{P}^i})$. Note that $d_\rho^{\pi^k}(s, a_{\mathcal{P}^i})$ cancels between the gradient and the Fisher pseudo-inverse.

Therefore, the NPG update becomes:

$$\theta_i^{k+1}(s, a_{\mathcal{P}^i}, a_i) = \theta_i^k(s, a_{\mathcal{P}^i}, a_i) + \eta \left(\frac{A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\lambda}{\pi_i^k(a_i | s, a_{\mathcal{P}^i})} - \lambda |\mathcal{A}_i| + c(s, a_{\mathcal{P}^i})\right).$$

Taking exponentials and normalizing over a_i, the constant $c(s, a_{\mathcal{P}^i})$ cancels, yielding (5). □

C.2 Proof of Lemma 3

Proof. Define $f_i^k(s, a_{\mathcal{P}^i}, a_i) := \frac{A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\lambda}{\pi_i^k(a_i|s, a_{\mathcal{P}^i})} - \lambda|\mathcal{A}_i|$ so that the update (5) reads $\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) \propto \pi_i^k(a_i|s, a_{\mathcal{P}^i}) \exp(\eta f_i^k(s, a_{\mathcal{P}^i}, a_i))$.

We prove by induction. For $\theta^0 = \mathbf{0}$, we have $\pi_i^0(a_i|s, a_{\mathcal{P}^i}) = \frac{1}{|\mathcal{A}_i|}$ for all $(s, a_{\mathcal{P}^i}, a_i)$, which satisfies the lower bound.

Suppose that

$$\pi_i^k(a_i|s, a_{\mathcal{P}^i}) \geq \frac{\lambda}{4\left(\lambda A_{\max} + \frac{1}{(1-\gamma)^2}\right)},$$

then from the definition of $f_i^k(s, a_{\mathcal{P}^i}, a_i)$, using $|\mathcal{A}_i| \leq A_{\max}$, we have that

$$-\frac{1}{(1-\gamma)^2} - \lambda A_{\max} \leq f_i^k(s, a_{\mathcal{P}^i}, a_i) \leq 5\left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right).$$

Thus

$$-\frac{1}{15} \leq \eta f_i^k(s, a_{\mathcal{P}^i}, a_i) \leq \frac{1}{3},$$

which leads to the fact that

$$\begin{aligned} Z_k^{i,s,a_{\mathcal{P}^i}} &= \sum_{a_i'} \pi_i^k(a_i'|s, a_{\mathcal{P}^i}) \exp\left(\eta f_i^k(s, a_{\mathcal{P}^i}, a_i')\right) \\ &\leq \sum_{a_i'} \pi_i^k(a_i'|s, a_{\mathcal{P}^i}) \left(1 + \eta f_i^k(s, a_{\mathcal{P}^i}, a_i') + (\eta f_i^k(s, a_{\mathcal{P}^i}, a_i'))^2\right) \\ &= 1 + \sum_{a_i'} \pi_i^k(a_i'|s, a_{\mathcal{P}^i})(\eta f_i^k(s, a_{\mathcal{P}^i}, a_i'))^2 \qquad (8) \\ &\leq 1 + \frac{1}{9} = \frac{10}{9}. \end{aligned}$$

Thus we have that

$$\frac{\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i})}{\pi_i^k(a_i|s, a_{\mathcal{P}^i})} = \frac{\exp\left(\eta f_i^k(s, a_{\mathcal{P}^i}, a_i)\right)}{Z_k^{i,s,a_{\mathcal{P}^i}}} \geq \frac{1 + \eta f_i^k(s, a_{\mathcal{P}^i}, a_i)}{Z_k^{i,s,a_{\mathcal{P}^i}}} \geq \frac{1 - \frac{1}{15}}{\frac{10}{9}} = \frac{21}{25}.$$

Thus, for a_i such that $\pi_i^k(a_i|s, a_{\mathcal{P}^i}) \geq \frac{\lambda}{3(\lambda A_{\max} + \frac{1}{(1-\gamma)^2})}$, we have

$$\pi_i^{k+1}(a_i|s, a_{\mathcal{P}^i}) \geq \frac{21}{25} \cdot \frac{\lambda}{3(\lambda A_{\max} + \frac{1}{(1-\gamma)^2})} \geq \frac{\lambda}{4(\lambda A_{\max} + \frac{1}{(1-\gamma)^2})}.$$

On the other hand, for a_i such that $\pi_i^k(a_i|s, a_{\mathcal{P}^i}) < \frac{\lambda}{3(\lambda A_{\max} + \frac{1}{(1-\gamma)^2})}$, we have

$$f_i^k(s, a_{\mathcal{P}^i}, a_i) \geq -\frac{1}{(1-\gamma)^2} - \lambda A_{\max} + 3\left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right) = 2\left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right).$$

From inequality (8) we have that

$$\begin{aligned} Z_k^{i,s,a_{\mathcal{P}^i}} &\leq 1 + \eta^2 \sum_{a_i'} \pi_i^k(a_i'|s, a_{\mathcal{P}^i}) f_i^k(s, a_{\mathcal{P}^i}, a_i')^2 \\ &\leq 1 + 25\eta^2 \left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right)^2 \\ &\leq 1 + \frac{5}{3}\eta \left(\frac{1}{(1-\gamma)^2} + \lambda A_{\max}\right). \end{aligned}$$

Thus

$$\frac{\pi_i^{k+1}(a_i|s,a_{\mathcal{P}^i})}{\pi_i^k(a_i|s,a_{\mathcal{P}^i})} = \frac{\exp\left(\eta f_i^k(s,a_{\mathcal{P}^i},a_i)\right)}{Z_k^{i,s,a_{\mathcal{P}^i}}} \geq \frac{1+\eta f_i^k(s,a_{\mathcal{P}^i},a_i)}{Z_k^{i,s,a_{\mathcal{P}^i}}}$$
$$\geq \frac{1+2\eta\left(\frac{1}{(1-\gamma)^2}+\lambda A_{\max}\right)}{1+\frac{5}{3}\eta\left(\frac{1}{(1-\gamma)^2}+\lambda A_{\max}\right)} \geq 1,$$

then according to the induction assumption, we have

$$\pi_i^{k+1}(a_i|s,a_{\mathcal{P}^i}) \geq \frac{\lambda}{4\left(\lambda A_{\max}+\frac{1}{(1-\gamma)^2}\right)},$$

which completes the proof. □

C.3 Lemma 10

Lemma 10 (Potential Function Ascent for Log-Barrier Regularized BN-NPG)**.** For π^{k+1} generated using the log-barrier regularized BN-NPG update (5) with η satisfying the condition in Lemma 3, define

$$W(s) := \sum_{i=1}^{n}\sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\langle \pi_i^{k+1}(\cdot|s,a_{\mathcal{P}^i}), A_i^{\pi^k}(s,a_{\mathcal{P}^i},\cdot)\rangle.$$

Then:

$$\Phi^{\pi^{k+1}}(\rho)-\Phi^{\pi^k}(\rho) \geq \left(\frac{1}{1-\gamma}-\frac{\sqrt{n}(1+A_{\max})\phi_{\max}\eta}{(1-\gamma)^3}\right)\sum_s d_\rho^{\pi^{k+1}}(s)W(s) - \frac{\sqrt{n}(1+A_{\max})\phi_{\max}\eta\lambda n}{\nu_\lambda(1-\gamma)^2}.$$

Proof. The proof follows the structure of Lemma 5. The telescoping decomposition, the vanishing of the A_{h_i} term (Lemma 4), and the error bound via $\|\tilde{\pi}_{-i}^k - \pi_{-i}^k\|_1$ are identical, yielding:

$$\Phi^{\pi^{k+1}}(\rho)-\Phi^{\pi^k}(\rho)$$
$$\geq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^{k+1}}(s)W(s) - \frac{\phi_{\max}}{(1-\gamma)^2}\sum_s d_\rho^{\pi^{k+1}}(s)(T_1(s)+T_2(s)),$$

where $W(s) := \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)\langle \pi_i^{k+1}(\cdot|s,a_{\mathcal{P}^i}), A_i^{\pi^k}(s,a_{\mathcal{P}^i},\cdot)\rangle$ and T_1, T_2 are defined as in the proof of Lemma 5. The Cauchy–Schwarz and Young's inequality steps are unchanged.

The only difference is in bounding T_1 and T_2 via KL divergence. From the regularized update (5):

$$\log\frac{\pi_i^{k+1}(a_i|s,a_{\mathcal{P}^i})}{\pi_i^k(a_i|s,a_{\mathcal{P}^i})} = \frac{\eta A_i^{\pi^k}(s,a_{\mathcal{P}^i},a_i)}{1-\gamma} + \frac{\eta\lambda}{\pi_i^k(a_i|s,a_{\mathcal{P}^i})} - \eta\lambda|\mathcal{A}_i| + c(s,a_{\mathcal{P}^i}),$$

where $c(s,a_{\mathcal{P}^i}) = -\log Z_k^{i,s,a_{\mathcal{P}^i}} \leq 0$. Therefore:

$$\mathrm{KL}(\pi_i^{k+1}\|\pi_i^k) = \frac{\eta}{1-\gamma}\langle \pi_i^{k+1}, A_i^{\pi^k}\rangle + \eta\lambda\sum_{a_i}\frac{\pi_i^{k+1}(a_i)}{\pi_i^k(a_i)} - \eta\lambda|\mathcal{A}_i| + c(s,a_{\mathcal{P}^i})$$
$$\leq \frac{\eta}{1-\gamma}\langle \pi_i^{k+1}, A_i^{\pi^k}\rangle + \frac{\eta\lambda}{\nu_\lambda},$$

where we used $c \leq 0$, $-\eta\lambda|\mathcal{A}_i| \leq 0$, and $\sum_{a_i}\pi_i^{k+1}(a_i)/\pi_i^k(a_i) \leq 1/\nu_\lambda$ (since $\pi_i^k \geq \nu_\lambda$).

Bounding T_1. By Pinsker's inequality:

$$T_1 \leq \sqrt{n} \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \left(\frac{\eta}{1-\gamma} \langle \pi_i^{k+1}, A_i^{\pi^k} \rangle + \frac{\eta\lambda}{\nu_\lambda} \right) = \frac{\sqrt{n}\eta}{1-\gamma} W(s) + \frac{\sqrt{n}\eta\lambda n}{\nu_\lambda}.$$

Bounding T_2. The T_2 bound uses $\mathrm{KL}(\pi_j^{k+1} \| \pi_j^k)$ for agents $j < i$ in the same way as the unregularized proof (via KL chain rule and marginal identity). Substituting the regularized KL bound:

$$T_2 \leq \frac{\sqrt{n} A_{\max} \eta}{1-\gamma} W(s) + \frac{\sqrt{n} A_{\max} \eta\lambda n}{\nu_\lambda}.$$

Combining.

$$T_1 + T_2 \leq \frac{\sqrt{n}(1 + A_{\max})\eta}{1-\gamma} W(s) + \frac{\sqrt{n}(1 + A_{\max})\eta\lambda n}{\nu_\lambda}.$$

Substituting into the potential difference bound:

$$\begin{aligned} &\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\ &\geq \left(\frac{1}{1-\gamma} - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \right) \sum_s d_\rho^{\pi^{k+1}}(s) W(s) - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta\lambda n}{\nu_\lambda (1-\gamma)^2}. \end{aligned}$$

□

C.4 Lemma 11

Lemma 11 (Bridge Between Regularized Update and Greedy Improvement)**.** Let $W(s)$ be defined as in Lemma 10 and $f^k(\alpha)$ as in Lemma 6. Then:

$$f^k(\eta) \leq W(s) + \frac{2n\eta\lambda}{\nu_\lambda(1-\gamma)}.$$

Proof. Recall $f^k(\eta) = \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \langle \pi_{i,\eta}(\cdot|s, a_{\mathcal{P}^i}), A_i^{\pi^k}(s, a_{\mathcal{P}^i}, \cdot) \rangle$ where $\pi_{i,\eta}(a_i|s, a_{\mathcal{P}^i}) \propto \pi_i^k(a_i|s, a_{\mathcal{P}^i}) \exp(\eta A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)/(1-\gamma))$, and $W(s) = \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \langle \pi_i^{k+1}(\cdot|s, a_{\mathcal{P}^i}), A_i^{\pi^k}(s, a_{\mathcal{P}^i}, \cdot) \rangle$ where π_i^{k+1} is the regularized update.

It suffices to show that for each $(i, s, a_{\mathcal{P}^i})$:

$$\langle \pi_{i,\eta}, A_i^{\pi^k} \rangle \leq \langle \pi_i^{k+1}, A_i^{\pi^k} \rangle + \frac{\eta\lambda}{\nu_\lambda(1-\gamma)}.$$

Since $-\eta\lambda|\mathcal{A}_i|$ cancels in normalization, the regularized update satisfies $\pi_i^{k+1}(a_i) \propto \pi_{i,\eta}(a_i) \cdot Z_\eta \cdot \exp(\eta\lambda/\pi_i^k(a_i))$ where Z_η is the normalizer for $\pi_{i,\eta}$. After renormalization:

$$\pi_i^{k+1}(a_i) = \frac{\pi_{i,\eta}(a_i) \exp(\eta\lambda/\pi_i^k(a_i))}{\sum_{a_i'} \pi_{i,\eta}(a_i') \exp(\eta\lambda/\pi_i^k(a_i'))}.$$

Define $\beta(a_i) := \eta\lambda/\pi_i^k(a_i)$. Since $\pi_i^k \geq \nu_\lambda$, we have $0 \leq \beta(a_i) \leq \eta\lambda/\nu_\lambda$. Then:

$$\begin{aligned} \langle \pi_{i,\eta}, A_i \rangle - \langle \pi_i^{k+1}, A_i \rangle &= \sum_{a_i} A_i(a_i) \pi_{i,\eta}(a_i) \left(1 - \frac{e^{\beta(a_i)}}{\sum_{a_i'} \pi_{i,\eta}(a_i') e^{\beta(a_i')}} \right) \\ &= \langle \pi_{i,\eta}, A_i \rangle - \frac{\sum_{a_i} \pi_{i,\eta}(a_i) e^{\beta(a_i)} A_i(a_i)}{\sum_{a_i'} \pi_{i,\eta}(a_i') e^{\beta(a_i')}}. \end{aligned}$$

This equals $\langle A_i \rangle_{\pi_{i,\eta}} - \langle A_i \rangle_{\tilde{\pi}}$ where $\tilde{\pi}(a_i) \propto \pi_{i,\eta}(a_i) e^{\beta(a_i)}$. By the simulation lemma:

$$|\langle A_i \rangle_{\pi_{i,\eta}} - \langle A_i \rangle_{\tilde{\pi}}| \leq \frac{1}{1-\gamma} \|\pi_{i,\eta} - \tilde{\pi}\|_1.$$

To bound $\|\pi_{i,\eta} - \tilde{\pi}\|_1$, note that $\tilde{\pi}(a_i)/\pi_{i,\eta}(a_i) = e^{\beta(a_i)}/\bar{w}$ where $\bar{w} = \sum_{a_i'} \pi_{i,\eta}(a_i') e^{\beta(a_i')}$. Since $\beta \in [0, \eta\lambda/\nu_\lambda]$:

$$\|\pi_{i,\eta} - \tilde{\pi}\|_1 = \sum_{a_i} \pi_{i,\eta}(a_i) |1 - e^{\beta(a_i)}/\bar{w}| \leq \max_{a_i} |1 - e^{\beta(a_i)}/\bar{w}|.$$

Since $1 \leq e^{\beta(a_i)} \leq e^{\eta\lambda/\nu_\lambda}$ and $1 \leq \bar{w} \leq e^{\eta\lambda/\nu_\lambda}$, we have $e^{-\eta\lambda/\nu_\lambda} \leq e^{\beta(a_i)}/\bar{w} \leq e^{\eta\lambda/\nu_\lambda}$. Using $|1 - e^x| \leq |x| e^{|x|}$ and $\eta\lambda/\nu_\lambda \leq 4/15$ (from the step size condition in Lemma 3):

$$\|\pi_{i,\eta} - \tilde{\pi}\|_1 \leq \frac{\eta\lambda}{\nu_\lambda} e^{\eta\lambda/\nu_\lambda} \leq \frac{\eta\lambda}{\nu_\lambda} \cdot 2 = \frac{2\eta\lambda}{\nu_\lambda}.$$

Therefore $|\langle \pi_{i,\eta}, A_i \rangle - \langle \pi_i^{k+1}, A_i \rangle| \leq \frac{2\eta\lambda}{\nu_\lambda(1-\gamma)}$, and summing over n agents (with $\sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1} \leq 1$):

$$f^k(\eta) \leq W(s) + \frac{2n\eta\lambda}{\nu_\lambda(1-\gamma)}.$$

□

C.5 Proof of Theorem 3

Proof. Let η satisfy both conditions in the theorem statement. Since $\eta \leq \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$, the leading coefficient in Lemma 10 satisfies $\frac{1}{1-\gamma} - \frac{\sqrt{n}(1+A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \geq \frac{1}{2(1-\gamma)}$. Since η also satisfies the condition in Lemma 3, the probability lower bound $\pi_i^k \geq \nu_\lambda$ holds.

The NE-gap bound from Lemma 6 gives (using M for the state visitation ratio and the termwise parent marginal ratio bound $\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s)/\pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \leq \nu_\lambda^{-|\mathcal{P}^i|} \leq \nu_\lambda^{-\kappa}$ from the proof of Theorem 1, which is bounded since $\pi_i^k \geq \nu_\lambda$):

$$\text{NE-gap}(\pi^k) \leq \frac{M}{\nu_\lambda^\kappa(1-\gamma)} \sum_s d_\rho^{\pi^{k+1}}(s) f^k(\eta) \left(1 + \frac{1-\gamma}{\nu_\lambda \delta \eta}\right).$$

By Lemma 11:

$$\sum_s d_\rho^{\pi^{k+1}}(s) f^k(\eta) \leq \sum_s d_\rho^{\pi^{k+1}}(s) W(s) + \frac{2n\eta\lambda}{\nu_\lambda(1-\gamma)}.$$

By Lemma 10:

$$\sum_s d_\rho^{\pi^{k+1}}(s) W(s) \leq 2(1-\gamma)\left(\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho)\right) + \frac{2(1-\gamma)\sqrt{n}(1+A_{\max})\phi_{\max}\eta\lambda n}{\nu_\lambda(1-\gamma)^2}.$$

Combining:

$$\text{NE-gap}(\pi^k) \leq 2M/\nu_\lambda^\kappa \left(\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho)\right)\left(1 + \frac{1-\gamma}{\nu_\lambda \delta \eta}\right) + R_\lambda,$$

where R_λ collects the λ-dependent bias terms. Summing over $k = 0, \ldots, K-1$, dividing by K, and using $\Phi^{\pi^K}(\rho) - \Phi^{\pi^0}(\rho) \leq \phi_{\max}/(1-\gamma)$:

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \leq \frac{2M\phi_{\max}}{K\nu_\lambda^\kappa(1-\gamma)}\left(1 + \frac{1-\gamma}{\nu_\lambda\delta\eta}\right) + R_\lambda,$$

where $R_\lambda = \frac{2Mn\eta\lambda(\sqrt{n}(1+A_{\max})\phi_{\max}+1)}{\nu_\lambda^\kappa \nu_\lambda(1-\gamma)^2}\left(1 + \frac{1-\gamma}{\nu_\lambda\delta\eta}\right)$. The bounded probability guarantee from Lemma 3 provides $\pi_i^k \geq \nu_\lambda > 0$ for all k, replacing Assumption 3 with $\nu = \nu_\lambda$ in Lemma 9 and Theorem 2. □

D Proofs for Partially Observable Extension

The proofs for the local-state extension follow the same structure as the full-observation case (Sections B–C), with $(s, a_{\mathcal{P}^i})$ replaced by $(s_i, a_{\mathcal{P}^i})$ and A_i replaced by $\bar{A}_i$ in the NPG update. We state each result formally and highlight only the steps that differ from the full-observation proofs.

D.1 Update Rule and Bounded Probability

Lemma 12 (Exponential Form of Local-State Log-Barrier BN-NPG)**.** The $d_\rho^{\pi^k}(s_i, a_{\mathcal{P}^i})$-weighted log-barrier regularized NPG update is equivalent to (7).

Proof. The proof is identical to Lemma 2 with blocks indexed by $(s_i, a_{\mathcal{P}^i})$ instead of $(s, a_{\mathcal{P}^i})$. The policy gradient aggregates over s_{-i}:

$$\frac{\partial \Phi^{\pi_\theta}(\rho)}{\partial \theta_i(s_i, a_{\mathcal{P}^i}, a_i)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s_i, a_{\mathcal{P}^i}) \pi_{\theta_i}(a_i|s_i, a_{\mathcal{P}^i}) \bar{A}_i^{\pi_\theta}(s_i, a_{\mathcal{P}^i}, a_i),$$

and the log-barrier gradient contributes $\lambda d_\rho^{\pi^k}(s_i, a_{\mathcal{P}^i})(1 - |\mathcal{A}_i| \pi_{\theta_i}(a_i|s_i, a_{\mathcal{P}^i}))$. The Fisher information matrix has block-diagonal structure indexed by $(s_i, a_{\mathcal{P}^i})$ with blocks $d_\rho^{\pi_\theta}(s_i, a_{\mathcal{P}^i}) F_{i, s_i, a_{\mathcal{P}^i}}(\theta_i)$. At $\theta = \theta^k$, the pseudo-inverse cancels $d_\rho^{\pi^k}(s_i, a_{\mathcal{P}^i})$ from both terms, yielding (7). □

Lemma 13 (Bounded Probability for Local-State Log-Barrier BN-NPG)**.** Under the same step size condition as Lemma 3 and $\theta^0 = \mathbf{0}$, the local-state update (7) guarantees $\pi_i^k(a_i|s_i, a_{\mathcal{P}^i}) \geq \nu_\lambda$ for all $k, i, s_i, a_{\mathcal{P}^i}, a_i$.

Proof. Since $|\bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i)| \leq 1/(1-\gamma)$ (as $\bar{A}_i$ is a convex combination of A_i values bounded by $1/(1-\gamma)$), the induction proof of Lemma 3 applies verbatim with $(s, a_{\mathcal{P}^i})$ replaced by $(s_i, a_{\mathcal{P}^i})$ and A_i replaced by $\bar{A}_i$. □

D.2 Potential Function Ascent

Lemma 14 (Potential Function Ascent for Local-State Log-Barrier BN-NPG)**.** For π^{k+1} generated using the local-state log-barrier regularized BN-NPG update (7) with η satisfying the condition in Lemma 13, define

$$W(s) := \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \langle \pi_i^{k+1}(\cdot|s_i, a_{\mathcal{P}^i}), A_i^{\pi^k}(s, a_{\mathcal{P}^i}, \cdot) \rangle.$$

Then:

$$\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \geq \left(\frac{1}{1-\gamma} - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \right) \sum_s d_\rho^{\pi^{k+1}}(s) W(s) - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{(1-\gamma)^2}.$$

Proof. The performance difference lemma, telescoping decomposition, and MPG cancellation (Lemma 4) all hold at the full-state level identically to the proof of Lemma 10. Crucially, $\sum_{a_i} (\pi_i^{k+1}(a_i|s_i, a_{\mathcal{P}^i}) - \pi_i^k(a_i|s_i, a_{\mathcal{P}^i})) A_{h_i}^{\pi^k}(s, a_{\mathcal{P}^i}, a_i) = 0$ holds by Lemma 4 for *any* pair of distributions over a_i, regardless of whether they depend on s or s_i. The Cauchy–Schwarz and Young's inequality steps bounding $T_1 + T_2$ are also unchanged.

The only difference is in the KL bound. From the local-state update (7):

$$\log \frac{\pi_i^{k+1}(a_i|s_i, a_{\mathcal{P}^i})}{\pi_i^k(a_i|s_i, a_{\mathcal{P}^i})} = \frac{\eta \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i)}{1-\gamma} + \frac{\eta\lambda}{\pi_i^k(a_i|s_i, a_{\mathcal{P}^i})} - \eta\lambda|\mathcal{A}_i| + c(s_i, a_{\mathcal{P}^i}),$$

where $c(s_i, a_{\mathcal{P}^i}) \leq 0$. Therefore:

$$\mathrm{KL}(\pi_i^{k+1}(\cdot|s_i, a_{\mathcal{P}^i}) \| \pi_i^k(\cdot|s_i, a_{\mathcal{P}^i})) \leq \frac{\eta}{1-\gamma} \langle \pi_i^{k+1}(\cdot|s_i, a_{\mathcal{P}^i}), \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, \cdot) \rangle + \frac{\eta\lambda}{\nu_\lambda}.$$

Bounding T_1. By Pinsker's inequality and the KL bound above:

$$T_1 \leq \sqrt{n} \sum_{i=1}^{n} \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \left(\frac{\eta}{1-\gamma} \langle \pi_i^{k+1}, \bar{A}_i^{\pi^k}(s_i, \cdot) \rangle + \frac{\eta\lambda}{\nu_\lambda} \right).$$

We decompose $\langle \pi_i^{k+1}, \bar{A}_i^{\pi^k}(s_i, \cdot) \rangle = \langle \pi_i^{k+1}, A_i^{\pi^k}(s, \cdot) \rangle + \langle \pi_i^{k+1}, \bar{A}_i^{\pi^k}(s_i, \cdot) - A_i^{\pi^k}(s, \cdot) \rangle$, where the second term satisfies $|\langle \pi_i^{k+1}, \bar{A}_i^{\pi^k}(s_i, \cdot) - A_i^{\pi^k}(s, \cdot) \rangle| \leq \epsilon_{\text{po}}$ by the definition of ϵ_{po}. Therefore:

$$T_1 \leq \frac{\sqrt{n}\eta}{1-\gamma} W(s) + \frac{\sqrt{n}\eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{1-\gamma},$$

where we used $\sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) = 1$ and summed over n agents.

Bounding T_2. The T_2 bound uses $\mathrm{KL}(\pi_j^{k+1} \| \pi_j^k)$ for agents $j < i$ via the KL chain rule and marginal identity, exactly as in the proof of Lemma 10. Substituting the KL bound with the ϵ_{po} correction:

$$T_2 \leq \frac{\sqrt{n} A_{\max} \eta}{1-\gamma} W(s) + \frac{\sqrt{n} A_{\max} \eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{1-\gamma}.$$

Combining.

$$T_1 + T_2 \leq \frac{\sqrt{n}(1 + A_{\max})\eta}{1-\gamma} W(s) + \frac{\sqrt{n}(1 + A_{\max})\eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{1-\gamma}.$$

Substituting into the potential difference bound (identical to Lemma 10):

$$\begin{aligned} &\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho) \\ &\geq \left(\frac{1}{1-\gamma} - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta}{(1-\gamma)^3} \right) \sum_s d_\rho^{\pi^{k+1}}(s) W(s) - \frac{\sqrt{n}(1 + A_{\max})\phi_{\max}\eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{(1-\gamma)^2}. \end{aligned}$$

□

D.3 Bridge Lemma

Lemma 15 (Bridge Between Local-State Regularized Update and Greedy Improvement)**.** Define $\bar{f}^k(s, \eta) := \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \langle \bar{\pi}_{i,\eta}(\cdot|s_i, a_{\mathcal{P}^i}), A_i^{\pi^k}(s, a_{\mathcal{P}^i}, \cdot) \rangle$, where $\bar{\pi}_{i,\eta}(a_i|s_i, a_{\mathcal{P}^i}) \propto \pi_i^k(a_i|s_i, a_{\mathcal{P}^i}) \exp(\eta \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i)/(1-\gamma))$. Then:

$$\bar{f}^k(s, \eta) \leq W(s) + \frac{2n\eta\lambda}{\nu_\lambda(1-\gamma)}.$$

Proof. The proof is identical to Lemma 11. For each $(i, s_i, a_{\mathcal{P}^i})$, the regularized update π_i^{k+1} differs from $\bar{\pi}_{i,\eta}$ only by the log-barrier perturbation $\exp(\eta\lambda/\pi_i^k(a_i|s_i, a_{\mathcal{P}^i}))$. The same TV-distance bound gives $|\langle \bar{\pi}_{i,\eta}, A_i^{\pi^k}(s, \cdot) \rangle - \langle \pi_i^{k+1}, A_i^{\pi^k}(s, \cdot) \rangle| \leq \frac{2\eta\lambda}{\nu_\lambda(1-\gamma)}$, and summing over n agents yields the result. □

D.4 NE-gap Connection

Lemma 16 (NE-gap to Potential Ascent (Partial Observation))**.** Under Assumption 4 and $\pi_i^k \geq \nu_\lambda$, we have:

$$\textit{NE-gap}(\pi^k) \leq \frac{M}{\nu_\lambda^\kappa(1-\gamma)} \sum_s d_\rho^{\pi^{k+1}}(s) \left[(\bar{f}^k(s,\eta) + n\epsilon_{\text{po}}) \left(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta}\right) + n\epsilon_{\text{po}} \right].$$

Proof. By the performance difference lemma for agent i's value function under local-state deviation $\pi_i'(\cdot|s_i, a_{\mathcal{P}^i})$:

$$V_i^{\pi_i', \pi_{-i}}(\rho) - V_i^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_i', \pi_{-i}}(s) \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \sum_{a_i} \pi_i'(a_i|s_i, a_{\mathcal{P}^i}) A_i^{\pi}(s, a_{\mathcal{P}^i}, a_i).$$

Taking the maximum over π_i' and using $d_\rho^{\pi_i', \pi_{-i}}(s) \leq M \cdot d_\rho^{\pi^{k+1}}(s)$ and the termwise parent marginal ratio bound $\pi_{\mathcal{P}^i}^k(a_{\mathcal{P}^i}|s) \leq \nu_\lambda^{-|\mathcal{P}^i|} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \leq \nu_\lambda^{-\kappa} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s)$ (since $\pi_i^k \geq \nu_\lambda$ and $\max_{a_i} A_i^{\pi^k} \geq 0$, as in the proof of Theorem 1):

$$\text{NE-gap}(\pi^k) \leq \frac{M}{\nu_\lambda^\kappa(1-\gamma)} \sum_s d_\rho^{\pi^{k+1}}(s) \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i).$$

We now relate $\max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i)$ to $\bar{f}^k(s,\eta)$. By the definition of ϵ_{po}:

$$\max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i) \leq \max_{a_i} \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i) + \epsilon_{\text{po}}.$$

To bound $\max_{a_i} \bar{A}_i$, we apply Lemma 6 to the marginalized advantages $\bar{A}_i$ with gap $\bar{\delta}$ (Assumption 4). Define $\bar{f}_{\bar{A}}^k(s_i,\eta) := \sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \langle \bar{\pi}_{i,\eta}, \bar{A}_i^{\pi^k}(s_i, \cdot)\rangle$. By Lemma 6 and $\exp(x) - 1 \geq x$:

$$\sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \max_{a_i} \bar{A}_i^{\pi^k}(s_i, a_{\mathcal{P}^i}, a_i) \leq \bar{f}_{\bar{A}}^k(s_i,\eta) \left(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta}\right).$$

Since $|\langle \bar{\pi}_{i,\eta}, \bar{A}_i(s_i, \cdot) - A_i(s, \cdot)\rangle| \leq \epsilon_{\text{po}}$ for each $(i, a_{\mathcal{P}^i})$:

$$\bar{f}_{\bar{A}}^k(s_i,\eta) \leq \bar{f}^k(s,\eta) + n\epsilon_{\text{po}}.$$

Combining:

$$\begin{aligned} &\sum_{i=1}^n \sum_{a_{\mathcal{P}^i}} \pi_{\mathcal{P}^i}^{k+1}(a_{\mathcal{P}^i}|s) \max_{a_i} A_i^{\pi^k}(s, a_{\mathcal{P}^i}, a_i) \\ &\leq (\bar{f}^k(s,\eta) + n\epsilon_{\text{po}}) \left(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta}\right) + n\epsilon_{\text{po}}, \end{aligned}$$

where the first $n\epsilon_{\text{po}}$ comes from $\bar{f}_{\bar{A}}^k \leq \bar{f}^k + n\epsilon_{\text{po}}$ and the second from $\max_{a_i} A_i \leq \max_{a_i} \bar{A}_i + \epsilon_{\text{po}}$. □

D.5 Proof of Theorem 4

Proof. Let η satisfy both conditions in the theorem statement. By Lemma 13, $\pi_i^k \geq \nu_\lambda$ for all k.

By Lemma 16, using $(\bar{f}^k + n\epsilon_{\text{po}})(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta}) + n\epsilon_{\text{po}} \leq (\bar{f}^k + 2n\epsilon_{\text{po}})(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta})$, and then Lemma 15:

$$\text{NE-gap}(\pi^k) \leq \frac{M}{\nu_\lambda^\kappa (1-\gamma)} \left(1 + \frac{1-\gamma}{\nu_\lambda \bar{\delta}\eta}\right) \left[\sum_s d_\rho^{\pi^{k+1}}(s) \left(W(s) + \frac{2n\eta\lambda}{\nu_\lambda(1-\gamma)}\right) + 2n\epsilon_{\text{po}}\right].$$

By Lemma 14, since $\eta \leq \frac{(1-\gamma)^2}{2\sqrt{n}(1+A_{\max})\phi_{\max}}$ implies the leading coefficient is $\geq \frac{1}{2(1-\gamma)}$:

$$\sum_s d_\rho^{\pi^{k+1}}(s) W(s) \leq 2(1-\gamma)\left(\Phi^{\pi^{k+1}}(\rho) - \Phi^{\pi^k}(\rho)\right) + \frac{2(1-\gamma)\sqrt{n}(1+A_{\max})\phi_{\max}\eta n(\epsilon_{\text{po}} + \lambda/\nu_\lambda)}{(1-\gamma)^2}.$$

Substituting, summing over $k = 0, \ldots, K-1$, dividing by K, and using $\Phi^{\pi^K}(\rho) - \Phi^{\pi^0}(\rho) \leq \phi_{\max}/(1-\gamma)$:

$$\frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) \leq \frac{2M\phi_{\max}}{K\nu_\lambda^\kappa(1-\gamma)}\left(1 + \frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right) + R_\lambda + \frac{4Mn\epsilon_{\text{po}}}{\nu_\lambda^\kappa(1-\gamma)}\left(1 + \frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right),$$

where $R_\lambda = \frac{2Mn\eta\lambda(\sqrt{n}(1+A_{\max})\phi_{\max}+1)}{\nu_\lambda^\kappa \nu_\lambda (1-\gamma)^2}(1 + \frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta})$ collects the λ-dependent bias terms (including the ϵ_{po} contributions from the potential ascent that are absorbed into the R_λ term at the same order as the regularization bias). □

D.6 Nash-to-Optimality Transfer

Lemma 17 (Nash-to-Optimality for Fully Connected Local-State BN in MPGs)**.** Under Assumption 1 and $\pi_i^k \geq \nu_\lambda$, for a fully connected BN in a Markov potential game, if π is an ϵ-Nash policy (i.e., *NE-gap*$(\pi) \leq \epsilon$), then:

$$\textit{optimality-gap}(\pi) \leq \frac{M\epsilon(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)} + \frac{n\epsilon_{\text{po}}}{1-\gamma}.$$

Proof. By the performance difference lemma for agent i's value function under local-state deviation, analogous to Lemma 7: from NE-gap$(\pi) \leq \epsilon$, for any $(s_i, a_{\mathcal{P}^i}, a_i)$, choosing a deviation policy π_i' that concentrates on a_i at $(s_i, a_{\mathcal{P}^i})$ while maintaining $\pi_i' \geq \nu_\lambda$, and using $d_\rho^{\pi_i', \pi_{-i}}(s) \geq 1/M$ (Assumption 1) and $\pi_{\mathcal{P}^i}(a_{\mathcal{P}^i}|s) \geq \nu_\lambda^{|\mathcal{P}^i|}$:

$$\epsilon \geq \frac{\nu_\lambda^{|\mathcal{P}^i|} \cdot \nu_\lambda}{M(1-\gamma)} \bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i),$$

giving $\bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i) \leq M(1-\gamma)\epsilon/\nu_\lambda^{|\mathcal{P}^i|+1}$.

By the definition of ϵ_{po}, the full-state advantage satisfies:

$$A_i^\pi(s, a_{\mathcal{P}^i}, a_i) \leq \bar{A}_i^\pi(s_i, a_{\mathcal{P}^i}, a_i) + \epsilon_{\text{po}} \leq \frac{M(1-\gamma)\epsilon}{\nu_\lambda^{|\mathcal{P}^i|+1}} + \epsilon_{\text{po}}.$$

By the MPG property, $A_{\phi,i}^\pi(s, a_{\mathcal{P}^i}, a_i) = A_i^\pi(s, a_{\mathcal{P}^i}, a_i)$. For a fully connected BN with $|\mathcal{P}^i| = i-1$, the recursive Q-function decomposition (Lemma 8) gives:

$$\begin{aligned} Q_\phi^\pi(s,a) &= \Phi^\pi(s) + \sum_{i=1}^n A_{\phi,i}^\pi(s, a_{\mathcal{P}^i}, a_i) \\ &\leq \Phi^\pi(s) + \sum_{i=1}^n \left(\frac{M(1-\gamma)\epsilon}{\nu_\lambda^i} + \epsilon_{\text{po}}\right) \\ &= \Phi^\pi(s) + M(1-\gamma)\epsilon \cdot \frac{1-\nu_\lambda^n}{\nu_\lambda^n(1-\nu_\lambda)} + n\epsilon_{\text{po}}. \end{aligned}$$

By the performance difference lemma for the potential:

$$\begin{aligned}\Phi^{\pi^*}(\rho) - \Phi^{\pi}(\rho) &= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi^*}(s)\sum_a \pi^*(a|s)(Q_\phi^\pi(s,a) - \Phi^\pi(s)) \\ &\leq \frac{1}{1-\gamma}\left[M(1-\gamma)\epsilon \cdot \frac{1-\nu_\lambda^n}{\nu_\lambda^n(1-\nu_\lambda)} + n\epsilon_{\text{po}}\right] \\ &= \frac{M\epsilon(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)} + \frac{n\epsilon_{\text{po}}}{1-\gamma}.\end{aligned}$$

□

D.7 Corollary 2

Corollary 2 (Optimality for fully correlated local-state BN policies)**.** Under the conditions of Theorem 4 with fully connected BN:

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1} \text{optimality-gap}(\pi^k) \\ \leq &\frac{M(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)}\left[\frac{2M\phi_{\max}}{K\nu_\lambda^{n-1}(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right) + R_\lambda + \frac{4Mn\epsilon_{\text{po}}}{\nu_\lambda^{n-1}(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right)\right] + \frac{n\epsilon_{\text{po}}}{1-\gamma}.\end{aligned}$$

Proof. By Lemma 17, for each k:

$$\text{optimality-gap}(\pi^k) \leq \frac{M(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)} \cdot \text{NE-gap}(\pi^k) + \frac{n\epsilon_{\text{po}}}{1-\gamma}.$$

Averaging over $k = 0, \ldots, K-1$ and applying Theorem 4 (with maximum in-degree $\kappa = n-1$ for the fully connected BN):

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1} \text{optimality-gap}(\pi^k) \\ \leq &\frac{M(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)} \cdot \frac{1}{K}\sum_{k=0}^{K-1} \text{NE-gap}(\pi^k) + \frac{n\epsilon_{\text{po}}}{1-\gamma} \\ \leq &\frac{M(1-\nu_\lambda^n)}{\nu_\lambda^n(1-\nu_\lambda)}\left[\frac{2M\phi_{\max}}{K\nu_\lambda^{n-1}(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right) + R_\lambda + \frac{4Mn\epsilon_{\text{po}}}{\nu_\lambda^{n-1}(1-\gamma)}\left(1+\frac{1-\gamma}{\nu_\lambda\bar{\delta}\eta}\right)\right] + \frac{n\epsilon_{\text{po}}}{1-\gamma}.\end{aligned}$$

□