Independent Learning of Nash Equilibria in Partially Observable Markov Potential Games with Decoupled Dynamics

Philip Jordan
SYCAMORE, EPFL
philip.jordan@epfl.ch

Maryam Kamgarpour
SYCAMORE, EPFL
maryam.kamgarpour@epfl.ch

Abstract

We study Nash equilibrium learning in partially observable Markov games (POMGs), a multi-agent reinforcement learning framework in which agents cannot fully observe the underlying state. Prior work in this setting relies on centralization or information sharing, and suffers from sample and computational complexity that scales exponentially in the number of players. We focus on a subclass of POMGs with independent state transitions, where agents remain coupled through their rewards, and assume that the underlying fully observed Markov game is a Markov potential game. For this class, we present an independent learning algorithm in which players, observing only their own actions and observations and without communication, jointly converge to an approximate Nash equilibrium. Due to partial observability, optimal policies may in general depend on the full action-observation history. Under a filter stability assumption, we show that policies based on finite history windows provide sufficient approximation guarantees. This enables us to approximate the POMG by a surrogate Markov game that is near-potential, leading to quasi-polynomial sample and computational complexity for independent Nash equilibrium learning in the underlying POMG.

1 Introduction

Markov games have become the standard framework for modeling multi-agent reinforcement learning. In this framework, agents are typically assumed to have full access to the underlying state of the environment. However, in many practical settings, such as autonomous driving [1, 2], multi-robot control under sensor noise [1, 3], or strategic interactions in games of imperfect information [4], this assumption is violated due to noisy or partial state information. Partially observable Markov games (POMGs) extend the Markov game model by allowing each agent to receive only limited or noisy observations of the true state. Developing principled multi-agent learning methods for such partially observable settings remains a central challenge.

Despite the broad applicability of POMGs, our understanding of their tractability, as well as the design of efficient algorithms with provable guarantees, remains limited. In contrast, in the single-agent setting of partially observable Markov decision processes (POMDPs), it is known that optimal policies may in general depend on the entire action-observation history [5], leading to computational intractability of the planning problem even for a single agent [6], as well as exponential statistical complexity for learning near-optimal policies [7]. However, these hardness results are worst-case in nature, and recent advances paint a clearer picture of rich POMDP subclasses that admit (quasi-) efficient learning and planning [8, 9]. These developments motivate investigating whether similar progress can be achieved in the multi-agent setting.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Tractability under partial observability. Since the main challenge in POMDPs arises from the dependence of optimal policies on the full action-observation history, a large body of work studies structural conditions under which finite-window approximations become effective. One line of work achieves this by imposing richness conditions on the observations. In particular, the α-weakly revealing condition of [9] yields polynomial sample complexity, disregarding computational aspects. The related notion of γ-observability enables planning and learning algorithms with *quasi-polynomial* computational and sample complexity via intricate policy cover constructions [10, 8]. Furthermore, [10] shows that under this condition, and standard complexity-theoretic assumptions, polynomial-time algorithms are ruled out.

Another family of positive algorithmic results is based on a *filter stability* assumption, under which the posterior distribution over states given the action-observation history converges exponentially fast, regardless of the initial prior. This enables practical finite-window approximations and has led to provable learning algorithms closer to those used in practice, including Q-learning [11], gradient methods [12], and temporal difference learning [13], that achieve quasi-polynomial computational and sample complexity [12–14].

These positive results in the single-agent setting raise the question of whether similar assumptions can enable provable methods for multi-agent learning under partial observability. While a large body of practical work proposes algorithms for POMGs [15–18], only few provide theoretical guarantees. From a statistical perspective, [19] extends a sample-efficient POMDP approach [9] to POMGs under an α-weakly revealing condition, by replacing the policy optimization oracle used in [9] with an oracle for equilibrium computation. This leaves open the possibility of developing methods that are both statistically and computationally efficient.

Achieving computationally efficient guarantees in the multi-agent setting is challenging even under full observability. In particular, Nash equilibrium computation is intractable in general Markov games, and already PPAD-hard for normal-form games [20]. As a result, much of the literature focuses on structured subclasses, among which Markov potential games are particularly well studied [21, 22]. In this setting, several independent learning algorithms have been shown to converge provably to Nash equilibria without requiring centralized coordination and scaling to large numbers of players [23–25].

Partially observable games with a common reward structure, captured by decentralized POMDPs (Dec-POMDPs) [26], form an important subclass of partially observable Markov potential games. A common approach in this setting is to assume that agents share information, effectively reducing the problem to a centralized decision process [27–29]. More recently, [30] establishes quasi-polynomial-time Nash equilibrium learning under γ-observability in cooperative POMGs by leveraging information sharing. However, this comes at the cost of an exponential dependence on the number of players, the so-called *curse of multi-agency*. Moreover, their results suggest that observability assumptions sufficient in the single-agent setting do not directly extend to decentralized multi-agent settings, where agents may form inconsistent beliefs due to differing local observations.

Decoupled dynamics. Motivated by these challenges, we aim to design a communication-free independent learning algorithm that avoids the curse of multi-agency. To this end, we impose an additional structural assumption: while agents remain coupled through the reward function, their state dynamics are decoupled. Such models arise naturally in applications including wireless networks [31, 32], smart energy grids [33], autonomous driving, and multi-robot control [3]; see Appendix B for detailed illustrations. Decoupled dynamics have also been studied in the Markov game literature [34, 35] and their partially observable extension defines a rich subclass of POMGs. Two concrete instances of the above applications, a dynamic transmission game over a shared channel and a cooperative navigation task, serve as the environments of our numerical experiments in Appendix F.

Contributions. In this work, we study POMGs with decoupled dynamics and potential structure under a filter stability assumption. Our approach builds on recent advances in approximating POMDPs by finite-state MDPs based on finite observation-action windows [11, 10, 12–14]. Extending this idea to the multi-agent setting, we approximate a decoupled POMG $\mathcal{P}$ by a Markov game $\mathcal{G}^m$ defined over m-step history windows, termed superstate Markov game. We then establish the following:

1. We show that approximate Nash equilibria of the Markov game $\mathcal{G}^m$ induce approximate Nash equilibria of the original POMG with explicit error bounds (Proposition 1).

2. Based on decoupledness of transition dynamics, we establish that the superstate Markov game $\mathcal{G}^m$ admits a Markov near-potential structure [36, 37] (Proposition 2).
3. Building on recent independent learning methods for Markov potential games [34], we propose an independent learning algorithm for the superstate Markov game $\mathcal{G}^m$ based on finite-window model estimation and soft policy iteration (see Algorithm 1).
4. We prove that the proposed algorithm converges to an ϵ-approximate Nash equilibrium of the underlying POMG with quasi-polynomial computational and sample complexity, under a concentrability condition on the visitation of finite history windows. In particular, the complexity does not scale exponentially with the number of players (see Theorem 1).
5. In addition to our theoretical results, we provide simulations in two decoupled potential POMG environments, a dynamic transmission game and a cooperative gridworld, in which we observe convergence of the proposed algorithm and and increasingly accurate approximation of the equilibrium in the POMG as the window length m grows (see Appendix F).

The remainder of the paper is organized as follows. In Section 2, we introduce the formal problem setting. Section 3 develops the structural properties of the superstate Markov game. Building on these insights, Section 4 presents our independent learning algorithm, and Section 5 establishes its convergence guarantees. We conclude in Section 6.

2 Problem setting

Let $\mathbb{N} \coloneqq \{1, 2, \dots\}$ and $\mathbb{N}_0 \coloneqq \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$, let $[n] \coloneqq \{1, 2, \dots, n\}$ and $[n]_0 \coloneqq [n] \cup \{0\}$. For $a, b \in \mathbb{N}$ with $a \le b$, let $[a, b] \coloneqq \{a, a+1, \dots, b-1, b\}$.

POMG. We consider N-player tabular finite-horizon partially observable Markov games in which transition and observation kernels are independent across players, but rewards are coupled. Formally, a POMG within this class is defined as a tuple $\mathcal{P} = (\mathcal{N}, H, \mathcal{S}, \mathcal{A}, \mathcal{O}, r, \mathbb{P}, \mathbb{O}, \mu)$ with players $\mathcal{N} \coloneqq [N]$, horizon $H \in \mathbb{N}$, finite state space $\mathcal{S} \coloneqq \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$, finite action space $\mathcal{A} \coloneqq \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$, and finite observation space $\mathcal{O} \coloneqq \mathcal{O}_1 \times \cdots \times \mathcal{O}_N$. Player i's reward function at step $h \in [H]$ is defined as $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$. State transition probabilities are given by $\mathbb{P} = (\mathbb{P}_{i,h})_{i \in \mathcal{N}, h \in [H]}$ with each player's transition $\mathbb{P}_{i,h}(\cdot \mid s_{i,h}, a_{i,h}) \in \Delta(\mathcal{S}_i)$ for $s_{i,h} \in \mathcal{S}_i$, $a_i \in \mathcal{A}_i$, and observation probabilities are given by $\mathbb{O} = (\mathbb{O}_{i,h})_{i \in \mathcal{N}, h \in [H]}$ with $\mathbb{O}_{i,h}(\cdot \mid s_{i,h}) \in \Delta(\mathcal{O}_i)$ for each $i \in \mathcal{N}$. Observe that rewards are coupled through the dependence on joint states and actions, whereas state transitions and observations factor across players, as detailed below.

At step $h \in [H]$, when in state $s_h \in \mathcal{S}$, joint observation $o_h \in \mathcal{O}$ is emitted with probability $\mathbb{O}_h(o_h \mid s_h) = \prod_{i \in \mathcal{N}} \mathbb{O}_{i,h}(o_{i,h} \mid s_{i,h})$, reflecting the decoupled observation structure, and each player $i \in \mathcal{N}$ observes its component $o_{i,h}$. Then, player i chooses an action $a_{i,h} \in \mathcal{A}_i$ and receives reward $r_{i,h}(s_h, a_h)$. Next, the unobserved state transitions to $s_{h+1} \in \mathcal{S}$ with probability $\mathbb{P}_h(s_{h+1} \mid s_h, a_h) = \prod_{i \in \mathcal{N}} \mathbb{P}_{i,h}(s_{i,h+1} \mid s_{i,h}, a_{i,h})$, reflecting the decoupled transition structure. The initial state is drawn from the product distribution $\mu \in \Delta(\mathcal{S})$, i.e., $\mu(s_1) = \prod_{i \in \mathcal{N}} \mu_i(s_{1,i})$.

Histories and policies. For each player $i \in \mathcal{N}$ and $h \in \mathbb{N}$, let $\mathcal{H}_i^h \coloneqq (\mathcal{O}_i \times \mathcal{A}_i)^{h-1} \times \mathcal{O}_i$ be the set of h-step action-observation histories, and let $\mathcal{H}_i^0$ denote the empty history. Let $\mathcal{H}_i^{\le h} \coloneqq \bigcup_{h' \in [h]_0} \mathcal{H}_i^{h'}$. Following the standard POMDP convention (see, e.g., [10]), histories are observation-first and end with the current observation: let $\tau_i \in \mathcal{H}_i^h$ be a history written as $\tau_i = (o_{i,1}, a_{i,1}, \dots, a_{i,h-1}, o_{i,h})$. For $k, l \in \mathbb{N}$ with $k \le l$, we define the sub-history $\tau_{i,k:l} = (o_{i,k}, a_{i,k}, \dots, a_{i,l-1}, o_{i,l})$. In particular, $\tau_{i,1:h} = \tau_i$. We denote by $|\tau_i| = h$ the length of τ_i. For each player $i \in \mathcal{N}$, we define the class of history-dependent policies as

$$\Pi_i^H \coloneqq \left\{ \pi_i = (\pi_{i,h})_{h \in [H]} \mid \pi_{i,h} : \mathcal{H}_i^h \to \Delta(\mathcal{A}_i) \text{ for all } h \in [H] \right\}.$$

We consider localized product policies $\pi = (\pi_1, \dots, \pi_N) \in \Pi_1^H \times \cdots \times \Pi_N^H$, meaning that at step h, given local histories $\tau = (\tau_1, \dots, \tau_N) \in \mathcal{H}^h$, joint action $a = (a_1, \dots, a_N) \in \mathcal{A}$ is chosen by $\pi \in \Pi^H$ with probability $\pi_h(a \mid \tau) = \prod_{i \in \mathcal{N}} \pi_{i,h}(a_i \mid \tau_i)$.

Moreover, for any joint policy $\pi \in \Pi^H$, we define for each $i \in \mathcal{N}$ the value function

$$V_i(\pi) \coloneqq \mathbb{E}_{\pi,\, s_1 \sim \mu}\left[\sum_{h=1}^{H} r_{i,h}(s_h, a_h)\right], \tag{1}$$

where the expectation is taken over the distribution of trajectories induced by the POMG $\mathcal{P}$ under the joint policy π, including the randomness from state transitions and observation emissions.

Nash equilibria. Let $\epsilon > 0$. A joint policy $\pi \in \Pi^H$ is called an ϵ-approximate Nash equilibrium if for all players $i \in \mathcal{N}$ and all $\pi_i' \in \Pi_i^H$, it holds that $V_i(\pi) \geq V_i(\pi_i', \pi_{-i}) - \epsilon$. When $\epsilon = 0$, we call π a Nash equilibrium.

Our objective in this paper is to provide an algorithm for learning ϵ-approximate Nash equilibria.

Potential structure. The hardness of Nash equilibrium computation in the general-sum regime [20] motivates us to focus on a tractable subclass such as potential games. For player $i \in \mathcal{N}$, let $\Pi_i \coloneqq \{\pi_i = (\pi_{i,h})_{h=1}^H \mid \pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)\}$ be the set of Markov (state-feedback) policies, and let $\Pi \coloneqq \Pi_1 \times \cdots \times \Pi_N$ denote the induced Markov product policies. Overloading the notation of (1), we write, for $\pi \in \Pi$,

$$V_i(\pi) \coloneqq \mathbb{E}_{\pi,\, s_1 \sim \mu}\Big[\textstyle\sum_{h=1}^{H} r_{i,h}(s_h, a_h)\Big], \tag{2}$$

where the expectation is over joint state-action trajectories of the underlying fully observed Markov game with rewards $r_{i,h}$, transitions $\mathbb{P}$, and initial distribution μ, when following policy the joint π.

Assumption 1. *There exist functions $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $h \in [H]$, such that for all $i \in \mathcal{N}$, $s_{-i} \in \mathcal{S}_{-i}$, $a_{-i} \in \mathcal{A}_{-i}$, $s_i, s_i' \in \mathcal{S}_i$, and $a_i, a_i' \in \mathcal{A}_i$,*

$$\begin{aligned} &r_{i,h}(s_i', s_{-i}, a_i', a_{-i}) - r_{i,h}(s_i, s_{-i}, a_i, a_{-i}) \\ &\qquad = \phi_h(s_i', s_{-i}, a_i', a_{-i}) - \phi_h(s_i, s_{-i}, a_i, a_{-i}). \end{aligned}$$

That is, at every step the stage game admits a potential ϕ_h. This stagewise potential condition is common in the decoupled Markov game setting (Definition 4 of [34]) and is typically easier to verify than a global potential condition on value functions. Moreover, under decoupled transitions it implies the global Markov potential game condition, namely the existence of $\Phi : \Pi \to \mathbb{R}$ with $V_i(\pi) - V_i(\pi_i', \pi_{-i}) = \Phi(\pi) - \Phi(\pi_i', \pi_{-i})$ for all $i \in \mathcal{N}$, $\pi \in \Pi$, and $\pi_i' \in \Pi_i$, which underlies a large body of work, see [23, 38, 34] and others. Assumption 1 is satisfied in the fully cooperative case $r_1 = \cdots = r_N$, as well as in practical settings such as demand-response markets [39], and other mixed cooperative/competitive scenarios for which [39] provides several sufficient conditions.

Before turning to the main objective of Nash equilibrium learning, the next section first establishes structural properties of the considered games which will then guide our algorithm design in Section 4.

3 Approximating POMGs via Markov near-potential games

In this section, we show that any decoupled POMG can be approximated by a Markov game, which we term the *superstate Markov game*, obtained via a finite-window approximation based on truncated histories. This extends related approaches from the single-agent setting [11, 10, 12] to the multi-agent case. We establish that the resulting superstate Markov game approximately preserves Nash equilibria of the original POMG (Proposition 1) and admits a near-potential structure (Proposition 2).

Superstate Markov game. Let $m \geq 1$. The superstate Markov game $\mathcal{G}^m$ has state space $\mathcal{H}^{\leq m}$ and action space $\mathcal{A}$. Concretely, at step h, the state is given by $w = (w_i)_{i \in \mathcal{N}}$, where we use the notation $w_i = (o_{i,h-m+1}, a_{i,h-m+1}, \ldots, a_{i,h-1}, o_{i,h}) \in \mathcal{H}_i^{\min(h,m)}$ for the most recent m-step local history of player i. To define the transition dynamics of the superstate Markov game, we introduce a belief over latent states induced by truncated histories. In general, a truncated history does not determine the posterior distribution of the underlying state. We therefore define, for each player $i \in \mathcal{N}$ and step $h \in [H]$, the belief $b_{i,h}^m(s_i \mid w_i)$: the probability that player i is in state s_i at step h after

observing the window w_i, starting from the initial distribution[1] μ_i. Formally, writing $s_{i,h} = s_i$, this is obtained by marginalizing over all local state trajectories $s_{i,h-m+1}, \dots, s_{i,h-1}$ consistent with w_i,

$$b^m_{i,h}(s_i \mid w_i) = \frac{1}{Z_i(w_i)} \sum_{s_{i,h-m+1},\dots,s_{i,h-1} \in \mathcal{S}_i} \mu_i(s_{i,h-m+1})\, \mathbb{O}_{i,\,h-m+1}(o_{i,h-m+1} \mid s_{i,h-m+1}) \cdot \prod_{k=h-m+1}^{h-1} \mathbb{P}_{i,\,k}(s_{i,k+1} \mid s_{i,k}, a_{i,k})\; \mathbb{O}_{i,\,k+1}(o_{i,k+1} \mid s_{i,k+1}),$$

where the normalization factor $Z_i(w_i)$ is the sum over $s_i \in \mathcal{S}_i$ of the unnormalized probabilities of reaching s_i. Intuitively, this expression averages over all latent state sequences that could have generated the observed window w_i, each weighted by its likelihood under the model. For $h \in [H-1]$, $w, w' \in \mathcal{H}^{\le m}$, $a \in \mathcal{A}$, and $o \in \mathcal{O}$ with $|w'| = n$ and $w' = (w \circ (a,o))_{n-m+2:n+1}$,

$$\mathbb{P}^m_h(w' \mid w, a) \;\coloneqq\; \sum_{s,s' \in \mathcal{S}} \Big(\prod_{i \in \mathcal{N}} b^m_{i,h}(s_i \mid w_i) \Big) \mathbb{P}_h(s' \mid s, a)\, \mathbb{O}_{h+1}(o \mid s'). \tag{3}$$

Otherwise, if w' cannot be obtained by concatenating w with any action-observation pair, we set $\mathbb{P}^m_h(w' \mid w, a) = 0$. Similarly, we define rewards $r^m_{i,h}(w,a) \coloneqq \sum_{s \in \mathcal{S}} r_{i,h}(s,a) \prod_{j \in \mathcal{N}} b^m_{j,h}(s_j \mid w_j)$. Moreover, for policy $\pi \in \Pi^m$, its value in $\mathcal{G}^m$ is defined as

$$V^m_i(\pi) \coloneqq \mathbb{E}_{\pi,\, s_1 \sim \mu} \left[\sum_{h=1}^{H} r^m_{i,h}(w_h, a_h) \right]$$

where the expectation is taken over the randomness of the policy and the superstate transitions $\mathbb{P}^m$.

By construction, $\mathcal{G}^m$ is a Markov game with a finite state space. Moreover, due to the decoupled structure of the underlying dynamics and observations, its transition kernel factorizes across players. Importantly, $\mathcal{G}^m$ is a conceptual object used for analysis: it is defined via a finite-window truncation of histories, which does not correspond to the true evolution of observations in the POMG. In particular, trajectories are generated by the underlying POMG, where observations depend on the full history, and therefore do not follow the superstate transition kernel $\mathbb{P}^m$.

We note that for $m = H$, we obtain a Markov game with state space corresponding to the full action-observation histories. This model is equivalent to the POMG in the sense that any joint policy over histories induces the same trajectory distribution and value in both models.

3.1 Finite-window policies and filter stability

Having defined the superstate Markov game $\mathcal{G}^m$, we aim to establish that its Nash equilibria over finite-window policies correspond to approximate Nash equilibria of the original POMG. Such a guarantee would justify using $\mathcal{G}^m$ as a surrogate game in the analysis of learning algorithms.

In order to argue about finite-window Nash equilibria, we first introduce the respective policy class.

Finite-window policies. For any $m \in [H]$, and each player $i \in \mathcal{N}$, let

$$\Pi^m_i \coloneqq \big\{ \pi_i = (\pi_{i,h})_{h \in [H]} \mid \pi_{i,h} : \mathcal{H}_i^{\min(h,m)} \to \Delta(\mathcal{A}_i) \text{ for all } h \in [H] \big\}.$$

We further define joint finite-window policies given by product policies $\pi = (\pi_1, \dots, \pi_N) \in \Pi^m_1 \times \cdots \times \Pi^m_N$, meaning that at step h, given local histories $\tau = (\tau_1, \dots, \tau_N) \in \mathcal{H}^h$, $a \in \mathcal{A}$ is chosen with probability $\pi'(a \mid \tau_{h-m+1\,:\,h}) = \prod_{i \in \mathcal{N}} \pi'_i(a_i \mid \tau_{i,h-m+1\,:\,h})$. We note that Π^m can be seen as a subclass of Π^H by associating to each $\pi \in \Pi^m$ its extension to Π^H that depends only on the most recent m steps of history. Thus, for $i \in \mathcal{N}$ and $\pi \in \Pi^m$, the value $V_i(\pi)$ is defined as in (1).

Without further assumptions, relevant information about the state may reside arbitrarily far in the past, and it is not clear whether finite-window approximation can yield meaningful guarantees. We therefore introduce a standard condition that ensures sufficient decay of past information.

[1] We use μ_i as a prior on the state at the start of the window; any prior over $\mathcal{S}_i$ could be used instead.

Filter stability. Filter stability formalizes the idea that the influence of initial beliefs decays exponentially over time, and has been studied in prior work on hidden Markov models [40] and POMDPs [11, 13].

For $h \in [H]$ and joint history $\tau = (\tau_1, \dots, \tau_N) \in \mathcal{H}^h$, let $b_h(\cdot \mid \tau) \in \Delta(\mathcal{S})$ denote the posterior distribution over joint states $\mathcal{S}$. By decoupledness of the dynamics and observations, it factorizes into the playerwise beliefs, $b_h(s \mid \tau) = \prod_{i \in \mathcal{N}} b_{i,h}(s_i \mid \tau_i)$. For $\nu, \nu' \in \Delta(\mathcal{S})$, define the total variation distance $\|\nu - \nu'\|_{TV} \coloneqq \frac{1}{2} \sum_{s \in \mathcal{S}} |\nu(s) - \nu'(s)|$.

Assumption 2. *There exists $\rho > 0$ such that for all $h \in [H]$, $\tau, \tau' \in \mathcal{H}^h$, $a \in \mathcal{A}$, and $o \in \mathcal{O}$,*

$$\|b_h(\cdot \mid \tau \circ (a, o)) - b_h(\cdot \mid \tau' \circ (a, o))\|_{TV} \leq (1 - \rho) \|b_h(\cdot \mid \tau) - b_h(\cdot \mid \tau')\|_{TV}$$

where $\circ$ denotes concatenation.

Filter stability requires (a) state transitions to be sufficiently mixing, and (b) observations to be sufficiently noisy. We refer to [41] for a sufficient condition that formalizes these requirements in terms of Dobrushin coefficients of the transition and observation kernel. Although stated for the joint belief, Assumption 2 hides no exponential dependence on the number of players; see the remark following Lemma 3 in Appendix C.

3.2 Finite-window Nash equilibrium approximation and near-potential structure

Based on the filter stability assumption above, we aim to show that Nash equilibria of the superstate Markov game over the finite-window policy class Π^m correspond to approximate Nash equilibria of the POMG over the full-history class Π^H. This approximation result will allow us to analyze equilibria in the tractable superstate setting while retaining guarantees for the original POMG. In addition, we establish that the superstate Markov game admits a near-potential structure, which will enable us to prove convergence of our learning algorithm.

For the single-agent case, it is known that under filter stability, the optimal value of a superstate MDP approximates the optimal value of the corresponding POMDP, see Theorem 2 of [13]. We extend this guarantee to the multi-agent setting by establishing value approximation results for joint policies, as well as for best-response values. Notably, this result, as well as Proposition 1 below, hold for general POMGs and do not rely on the decoupled dynamics assumption.

Lemma 1. *Let $\pi \in \Pi^m$ and $\pi' \in \Pi^H$ such that for all $h \geq m$ and all $w \in \mathcal{H}^m$, $\tau \in \mathcal{H}^h$ with $w = \tau_{h-m+1:h}$, it holds that $\pi_h(\cdot \mid w) = \pi'_h(\cdot \mid \tau)$. Let*

$$\epsilon_\rho^m \coloneqq 4H^2(1 - \rho)^{m-1}. \tag{4}$$

Under Assumption 2, for any $i \in \mathcal{N}$, we have

$$|V_i^m(\pi) - V_i(\pi')| \leq \epsilon_\rho^m.$$

Moreover, for π and π' as above, it holds that

$$\left| \max_{\hat{\pi}_i \in \Pi_i^m} V_i^m(\hat{\pi}_i, \pi_{-i}) - \max_{\hat{\pi}'_i \in \Pi_i^H} V_i(\hat{\pi}'_i, \pi_{-i}) \right| \leq \epsilon_\rho^m.$$

We are then able to connect equilibria of the superstate Markov game to those of the original POMG.

Proposition 1. *Let $\epsilon > 0$, and let $\pi \in \Pi^m$ be an ϵ-approximate Nash equilibrium of the superstate Markov game $\mathcal{G}^m$. Under Assumption 2, π is an $(\epsilon + 2\epsilon_\rho^m)$-approximate Nash equilibrium of $\mathcal{P}$.*

The proofs of Lemma 1 and Proposition 1 are provided in Appendix C.

The following result shows that $\mathcal{G}^m$ inherits a *near-potential* structure [36], suggesting that efficient learning in the superstate Markov game is possible.

Proposition 2. *Under Assumptions 1 and 2, there exists a function $\Psi : \Pi^m \to \mathbb{R}$ such that for all $\pi \in \Pi^m$, $i \in \mathcal{N}$, and $\pi'_i \in \Pi_i^m$, and ϵ_ρ^m as in (4), we have*

$$|(V_i^m(\pi) - V_i^m(\pi'_i, \pi_{-i})) - (\Psi(\pi) - \Psi(\pi'_i, \pi_{-i}))| \leq 2\epsilon_\rho^m.$$

The proof is deferred to Appendix C. At a high level, the decoupled structure allows state visitation distributions to factor across players, enabling the construction of a potential function for the POMG, which then transfers approximately to the finite-window setting.

Building upon these structural properties, we now turn to the question of learning in POMGs.

4 Independent Nash equilibrium learning algorithm

In this section, we design a learning algorithm for decoupled POMGs with potential structure. We begin by stating the learning objective and protocol, then we outline our approach. An implementation of the resulting algorithm, along with simulation results in two decoupled potential POMG environments, is provided in Appendix F.

Learning Objective. Given $\epsilon > 0$, our objective is to efficiently learn a finite-window policy $\pi \in \Pi^m$ such that π is an ϵ-approximate Nash equilibrium of the POMG, that is, for all $i \in \mathcal{N}$ and all deviations with full history dependence $\pi_i' \in \Pi_i^H$, it holds that $V_i(\pi) \geq V_i(\pi_i', \pi_{-i}) - \epsilon$.

Independent learning protocol. We consider an independent learning setting [42, 43, 24] in which players interact with the game over a fixed number of episodes. Within each episode, every player follows a fixed policy $\pi_i \in \Pi^m$, observes only its own sequence of actions, observations, and rewards, and then updates its policy individually. Players do not observe other agents' actions, observations, rewards, or policies. Moreover, no information is shared among players or with any central entity. We do assume, however, that all players agree in advance to follow the same algorithm, which places the resulting dynamics in the self-play regime [44].

Our algorithm (Algorithm 1) closely follows the method of [34] for fully observable Markov games, applied to the superstate game. The main challenges arise in the analysis, which we outline in the proof overviews in Section 5.

Algorithm 1 Learning Finite-Window Policy in POMG (at player i)

Input: iteration number K, episode number T, window length m, stepsizes $\eta^{(k)}$ for $k \in [K]$, exploration probability $\zeta \in (0, 1]$.
Initialize $\pi_{i,h}^{(0)} = 1/|\mathcal{A}_i|$ for all $h \in [H]$.
for $k = 0, 1, \ldots, K-1$ **do**
  Collect trajectory $\tau_i = ((a_{i,h}^{(t)}, o_{i,h}^{(t)}, r_{i,h}^{(t)})_{h=1}^{H})_{t=1}^{T}$ by following $\pi_i^{\zeta,(k)} := \zeta\mathcal{U}(\mathcal{A}_i) + (1-\zeta)\pi_i^{(k)}$, i.e., $a_{i,h}^{(t)} \sim \mathcal{U}(\mathcal{A}_i)$ w.p. ζ and $a_{i,h}^{(t)} \sim \pi_{i,h}^{(k)}(\cdot \mid w_{i,h}^{(t)})$ otherwise, for all $h \in [H]$ and $t \in [T]$.
  Estimate $\hat{\mathbb{P}}_{i,h}^{m,\pi_i^{(k)}}$ and $\hat{r}_i^{m,\pi_{-i}^{(k)}}$ based on empirical frequencies (see Appendix D for details).
  Let $\hat{Q}_{i,H+1}^{(k)}(w_i, a_i) := 0$ for all $w_i \in \mathcal{H}_i^{\leq m}, a_i \in \mathcal{A}_i$.
  for $h = H, \ldots, 1$ **do**
    for $w_i \in \mathcal{H}_i^{\leq m}, a_i \in \mathcal{A}_i$ **do**
      $\hat{Q}_{i,h}^{(k)}(w_i, a_i) := \hat{r}_{i,h}^{m,\pi_{-i}^{(k)}}(w_i, a_i) + \sum_{w_i', a_i'} \hat{\mathbb{P}}_{i,h}^{m,\pi_i^{(k)}}(w_i' \mid w_i, a_i)\pi_{i,h+1}^{(k)}(a_i' \mid w_i')\hat{Q}_{i,h+1}^{(k)}(w_i', a_i')$.
      $\pi_{i,h}^{(k+1)}(a_i \mid w_i) := (1-\eta^{(k)})\pi_{i,h}^{(k)}(a_i \mid w_i) + \eta^{(k)}\mathbf{1}\left\{a_i \in \arg\max_{a_i'} \hat{Q}_{i,h}^{(k)}(w_i, a_i')\right\}$.
    end for
  end for
end for

4.1 Independent soft policy iteration

A key challenge in independent learning in Markov games, even under decoupled transitions, is that each player's value function depends on the joint policy of all players through the coupled rewards. However, when fixing any π_{-i} in the superstate Markov game $\mathcal{G}^m$, player i faces an MDP $\mathcal{M}^{\pi_{-i}}$ with state space $\mathcal{S}_i$, action space $\mathcal{A}_i$, transition kernel $\mathbb{P}_i^m$ as defined in (3), and reward

$$r_{i,h}^{m,\pi_{-i}}(w_i, a_i) := \mathbb{E}_{w_{-i} \sim d_{-i,h}^{m,\pi}, a_{-i} \sim \pi_{-i}(\cdot \mid w_{-i})}\left[r_{i,h}^m(w_i, w_{-i}, a_i, a_{-i})\right],$$

where $d_{i,h}^{m,\pi}(w_{i,h}) := P_\pi(w_{i,h} = w_i)$, and $d_{-i,h}^{m,\pi}(w_{-i,h}) := \prod_{j \in \mathcal{N} \setminus \{i\}} d_{j,h}^{m,\pi}(w_j)$. Note that here P_π is taken over the randomness of policy π and transition kernel $\mathbb{P}^m$. We further denote the Q-function of $\mathcal{M}^{\pi_{-i}}$ by $\overline{Q}_{i,h}^{m,\pi} : \mathcal{H}_i^{\leq m} \times \mathcal{A}_i \to \mathbb{R}$.

Observe that if a policy $\pi^\star \in \Pi^m$ is such that for all $i \in \mathcal{N}$, $\pi_i^\star$ achieves the optimal value in $\mathcal{M}^{\pi^\star_{-i}}$, then $\pi^\star$ is a Nash equilibrium of $\mathcal{G}^m$. This motivates having each player perform policy iteration with respect to $\overline{Q}_{i,h}^{m,\pi}$ in the hope of converging to a joint policy that satisfies this playerwise optimality in $\mathcal{M}^{\pi_{-i}}$. However, in a game setting, such simultaneous updates may lead to cyclic patterns that fail to converge [45]. In contrast, the smoothed version with appropriately chosen stepsize $0 < \eta^{(k)} < 1$, as implemented in Line 10 of Algorithm 1, has been shown to converge to an approximate Nash equilibrium in Markov potential games [34].

It remains to describe how to obtain $\overline{Q}_{i,h}^{m,\pi^{(k)}}$. As the Q-function of an MDP, namely $\mathcal{M}^{\pi_{-i}^{(k)}}$, $\overline{Q}_{i,h}^{m,\pi^{(k)}}$ satisfies the Bellman expectation equation, which allows us to compute an estimate $\hat{Q}_{i,h}^{(k)}$ via backward iteration (see Line 9 of Algorithm 1). This computation relies on having estimates $\hat{\mathbb{P}}_{i,h}^{m,\pi_i^{(k)}}$ and $\hat{r}_{i,h}^{m,\pi_{-i}^{(k)}}$ of the transition kernel of $\mathcal{M}^{\pi_{-i}^{(k)}}$. We next describe how to obtain these estimates for the superstate game from interaction with the POMG at each iteration.

4.2 Sampling & model estimation

As we do not assume access to transition kernels or a generative model, players sample action-observation sequences over T episodes by simultaneously interacting with the POMG and receiving their respective observations and rewards. At iteration k and step h, in order to ensure exploration of actions, player $i \in \mathcal{N}$ takes an action uniformly at random with probability ζ, and otherwise follows its current policy $\pi_{i,h}^{(k)}$. Based on the observed sequences $\tau_i = \{\{a_{i,h}^{(t)}, o_{i,h}^{(t)}, r_{i,h}^{(t)}\}_{h=1}^{H}\}_{t=1}^{T}$, we apply standard model estimation techniques inspired by single-agent methods [46, 47] as detailed below.

Transition estimation. For each player $i \in \mathcal{N}$, we estimate the superstate transition probabilities $\mathbb{P}_{i,h}^m(\cdot \mid w_i, a_i)$ by empirical frequencies of m-step window transitions. Concretely, for each pair $(w_i, a_i) \in \mathcal{H}_i^{\leq m} \times \mathcal{A}_i$, we collect all time steps at which the m-step history window equals w_i and action a_i is played by policy $\pi_i^{(k)}$, and record the resulting next window $w_i' \in \mathcal{H}_i^{\leq m}$. The estimate $\hat{\mathbb{P}}_{i,h}^{m,\pi_i^{(k)}}(w_i' \mid w_i, a_i)$ is given by the relative frequency of observing w_i' among these occurrences. If a pair (w_i, a_i) is not observed, the estimate is set to 0.

Reward estimation. Similarly, we estimate the reward function $r_{i,h}^{m,\pi_{-i}^{(k)}}$ of $\mathcal{M}^{\pi_{-i}^{(k)}}$ by empirical averaging. For each $(w_i, a_i) \in \mathcal{H}_i^{\leq m} \times \mathcal{A}_i$, we average the rewards observed at time steps where the history window equals w_i and action a_i is taken. If (w_i, a_i) is not observed, the estimate is set to 0.

For formal definitions of the estimates $\hat{\mathbb{P}}_i^m$ and $\hat{r}_{i,h}^{m,\pi_{-i}^{(k)}}$, we refer to Appendix D. A key challenge towards establishing guarantees for Algorithm 1 lies in deriving confidence bounds for this sampling procedure. We provide further insight on this in the proof overview of Lemma 2 in the next section.

5 Convergence to Nash equilibrium

In this section, we present our main result, Theorem 1, which provides the convergence guarantee for Algorithm 1 towards an approximate Nash equilibrium.

Our analysis relies on establishing confidence bounds for the estimated superstate model $\hat{\mathbb{P}}_i^m$ and $\hat{r}_{i,h}^{m,\pi_{-i}}$ (see Lemma 2 below). This in turn requires that all m-step history windows are well explored. A sufficient condition is that, at each superstate, all actions are selected and all observations are realized with probabilities bounded away from zero. For actions, this is ensured by the ζ-greedy exploration strategy in Line 4 of Algorithm 1. For observations, we impose the following assumption.

Assumption 3. *There exists $\beta > 0$ such that for all $i \in \mathcal{N}$, $s_i \in \mathcal{S}_i$ and $o_i \in \mathcal{O}_i$, $\mathbb{O}_i(o_i \mid s_i) \geq \beta$.*

We note that Assumption 3 can be enforced, for example, by having each player inject a small amount of uniform noise into its observations, ensuring that every observation occurs with probability bounded away from zero. Next, based on this assumption, we state the estimation confidence bounds.

Lemma 2. *Let Assumptions 2 and 3 hold, and let $\pi \in \Pi^m$. Suppose each player $i \in \mathcal{N}$ collects a trajectory according to Line 4 in Algorithm 1 with exploration probability $\zeta \in (0,1]$, and estimates $\hat{\mathbb{P}}_i^{m,\pi_i}$ and $\hat{r}_i^{m,\pi_{-i}}$ according to (7) and (8), respectively. Let $i \in \mathcal{N}$, $h \in [H]$, $a_i \in \mathcal{A}_i$, and $w_i, w_i' \in \mathcal{H}_i^{\leq m}$. Then for any $0 < \epsilon \leq 1$,*

$$P\left(|\mathbb{P}_{i,h}^m(w_i' \mid w_i, a_i) - \hat{\mathbb{P}}_{i,h}^{m,\pi_i}(w_i' \mid w_i, a_i)| \geq \epsilon + (1-\rho)^{m-1}\right) \leq 4\exp\left(-\frac{\epsilon^2\zeta^{2m}\beta^{2m}T}{32|\mathcal{A}_i|^{2m}}\right),$$

$$P\left(|r_{i,h}^{m,\pi_{-i}}(w_i, a_i) - \hat{r}_{i,h}^{m,\pi_{-i}}(w_i, a_i)| \geq \epsilon + 4NH\zeta + N^2H(1-\rho)^{m-1}\right) \leq 4\exp\left(-\frac{\epsilon^2\zeta^{2m}\beta^{2m}T}{32|\mathcal{A}_i|^{2m}}\right).$$

Proof overview. The main challenge we face is a mismatch between superstate Markov game and the underlying POMG process: action-observation sequences are collected from the POMG kernel with full history dependence $\mathbb{P}^H$, rather than the superstate Markov game kernel $\mathbb{P}^m$. While this issue arises already in POMDPs [12–14], in the game setting the mismatch additionally enters through the other agents: player i's marginalized reward $r_{i,h}^{m,\pi_{-i}}$ averages over the other agents' superstate distributions, which are induced by $\mathbb{P}^m$, whereas in the collected data they follow the POMG kernel $\mathbb{P}^H$. Thus, the approximation errors in all other agents' dynamics propagate into player i's reward estimate. We handle this by decomposing the estimation error of $\hat{\mathbb{P}}_{i,h}^{m,\pi_i}$ and $\hat{r}_{i,h}^{m,\pi_{-i}}$ into a statistical term arising from randomness of the policy and environment, and a bias term due to the finite-window approximation (Lemma 2). The first term is controlled by a concentration argument across the independently sampled episodes, while the bias term is bounded under Assumption 2 by analyzing how mismatch propagates through other agents' states (Lemma 7). Compared to the corresponding result for Markov games (see [34], Lemma 4), we highlight the additional bias terms $(1-\rho)^{m-1}$ and $N^2H(1-\rho)^{m-1}$, that stem from the superstate approximation error.

For the full proof of Lemma 2 and its supporting results, we refer to Appendix E.

It is standard for sample-based single- and multi-agent reinforcement learning methods to require certain assumptions on the state visitation distribution [48, 23, 24]. In the fully observable Markov game case, [34] impose a uniform lower bound on the visitation of the latent states (Assumption 1 in [34]). The analogous uniform lower bound on the superstate occupancy would be exponentially small in m, since the occupancy is spread over $|\mathcal{H}_i^m|$ windows, and is thus incompatible with the choice of m in Theorem 1. We therefore only require that the occupancy of the *iterates* covers that of the best response, a weaker condition that is standard for policy optimization methods with local, occupancy-weighted updates such as policy iteration and policy gradient [49, 48], including in finite-memory POMDPs [12]. In the following, since transitions are decoupled, $d_{i,h}^{m,\pi}$ depends on π only through π_i, and we write $d_{i,h}^{m,\pi_i}$ accordingly.

Assumption 4 (Superstate concentrability)**.** *For iteration $k \in [K]$ of Algorithm 1, let*

$$\kappa^{(k)} := \max_{i\in\mathcal{N},\, h\in[H],\, w_i\in\mathcal{H}_i^{\leq m}} \frac{d_{i,h}^{m,\pi_i^{\star,k}}(w_i)}{d_{i,h}^{m,\pi_i^{(k+1)}}(w_i)}, \qquad \textit{where } \pi_i^{\star,k} \in \arg\max_{\pi_i'\in\Pi_i^m} V_i(\pi_i', \pi_{-i}^{(k)}).$$

We assume that there exists $\kappa \geq 1$ with $\kappa \geq \frac{\sum_{k\in[K]}\eta^{(k)}}{\sum_{k\in[K]}\eta^{(k)}/\kappa^{(k)}}$.

Each $\kappa^{(k)}$ compares, for every player, the best response occupancy $d_{i,h}^{m,\pi_i^{\star,k}}(w_i)$ against the occupancy of policy $d_{i,h}^{m,\pi_i^{(k+1)}}(w_i)$. Rather than bounding each $\kappa^{(k)}$ individually, we only require κ to upper bound their $\eta^{(k)}$-weighted mean. This suffices since iteration k contributes an improvement proportional to $\eta^{(k)}/\kappa^{(k)}$, so that only $\sum_{k\in[K]}\eta^{(k)}/\kappa^{(k)} \geq \kappa^{-1}\sum_{k\in[K]}\eta^{(k)}$ enters the telescoping argument in the proof of Lemma 10.

Next, we state our guarantee for Algorithm 1 in learning an approximate Nash equilibrium.

Theorem 1. *Let Assumptions 1, 2, 3, and 4 hold. For $\epsilon > 0$, $\delta > 0$, suppose each player $i \in \mathcal{N}$ runs Algorithm 1 simultaneously with stepsizes $\eta^{(k)} = \Theta(1/\sqrt{N^2H^3k})$, exploration $\zeta = \epsilon$, and*

$$m \geq \min\left\{H,\ c_1 \cdot \rho^{-1}\log\left(\frac{\kappa NHAO}{\beta\epsilon}\right)\right\},$$

$$T \geq c_2 \cdot \frac{\kappa^2 A^{2m+2} H^6 O^2 N^2}{\beta^{2m} \epsilon^{2m+2}} \log(1/\delta),$$
$$K \geq c_3 \cdot \frac{\kappa^2 H^5 N^4 A^2}{\epsilon^2},$$

where $c_1, c_2, c_3 \in \mathbb{N}$ are constants independent of the problem parameters, and $A \coloneqq \max_{i \in \mathcal{N}} |\mathcal{A}_i|$, $O \coloneqq \max_{i \in \mathcal{N}} |\mathcal{O}_i|$. Then, with probability $1 - \delta$, there exists $k \in [K]$ such that the joint policy $\pi^{(k)}$ is an ϵ-approximate Nash equilibrium of the POMG $\mathcal{P}$. The sample and computational complexity of Algorithm 1 are at most

$$\mathcal{O}\left(\left(\frac{\kappa N H A O}{\beta \epsilon}\right)^{\mathcal{O}(\rho^{-1} \log(A/(\beta\epsilon)))} \log(1/\delta)\right). \tag{5}$$

Proof overview. Although the superstate game need not admit an exact potential, it satisfies a near-potential property (Proposition 2). This ensures that each iteration of soft policy iteration yields an approximate improvement, up to an error on the order of $H^2(1-\rho)^{m-1}$ (Lemma 10). Choosing m appropriately controls the accumulated error and yields convergence. For the full proof, we refer to Appendix E.

(Quasi-)efficiency. We highlight that the sample and computational complexity (5) is quasi-polynomial whenever the concentrability constant κ is, and in particular does *not* scale exponentially in the number of players N. By decoupledness, $d_{i,h}^{m,\pi}$ depends on π only through π_i, so that κ is a single-player quantity carrying no dependence on N. Thus, in our decoupled potential game setting, Algorithm 1 breaks the curse of multi-agency.

Window size tradeoff. The choice of the window size m induces a tradeoff between approximation accuracy and statistical efficiency. Larger values of m reduce the approximation error of the value in the superstate Markov game relative to the respective value in the original POMG (see Propositions 1 and 2), but increase the sampling complexity required to accurately estimate the superstate model (see Lemma 2). Choosing m on a logarithmic scale as in Theorem 1 balances these effects.

Comparison with prior work. Compared to [19], our framework relies on a different structural assumption, filter stability rather than observability, and focuses on decoupled dynamics with potential structure. This restriction enables an *independent learning* approach that does not rely on computationally intractable oracles. In contrast to [30], our setting does not require information sharing. Moreover, unlike [19, 30] our method does not suffer from the curse of multi-agency.

6 Conclusion

We studied Nash equilibrium learning in partially observable Markov games under decoupled dynamics and potential structure. We proposed an independent learning algorithm that converges to an approximate Nash equilibrium with quasi-polynomial sample and computational complexity, under a boundedness assumption on a superstate concentrability coefficient. Our analysis is based on a superstate Markov game representation, which we show approximately preserves equilibria and admits a near-potential structure.

Limitations and future work. The setting of POMGs with decoupled dynamics and potential structure forms a rich but still restricted subclass of POMGs. Understanding the boundary between tractable and intractable regimes remains an important direction for future work. Another limitation is Assumption 4, which bounds the superstate concentrability coefficients along the trajectory of Algorithm 1; identifying concrete classes of POMGs in which κ is small is an interesting direction. To the best of our knowledge, this is the first work on provable independent learning in POMGs. Our results open up new possibilities for independent learning in partially observable multi-agent systems and provide a starting point for further extensions to other game classes, such as two-player zero-sum games, and to broader equilibrium notions such as correlated equilibria in general-sum games.

References

[1] Mikko Lauri, David Hsu, and Joni Pajarinen. Partially observable Markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1):21–40, 2022.

[2] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online POMDP planning for autonomous driving in a crowd. In *International Conference on Robotics and Automation*, pages 454–460. IEEE, 2015.

[3] Yuchen Xiao, Weihao Tan, Joshua Hoffman, Tian Xia, and Christopher Amato. Asynchronous multi-agent deep reinforcement learning under partial observability. *The International Journal of Robotics Research*, 44(8):1257–1286, 2025.

[4] Jian Yao, Zeyu Zhang, Li Xia, Jun Yang, and Qianchuan Zhao. Solving imperfect information Poker games using Monte Carlo search and POMDP models. In *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 1060–1065. IEEE, 2020.

[5] Martin L Puterman. Markov decision processes: Discrete stochastic dynamic programming, 1994.

[6] Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

[7] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 2016.

[8] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 349–362, 2023.

[9] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

[10] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable POMDPs, without computationally intractable oracles. *Advances in Neural Information Processing Systems*, 2022.

[11] Ali Devran Kara and Serdar Yüksel. Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4):2066–2093, 2023.

[12] Semih Cayci, Niao He, and R Srikant. Finite-time analysis of natural actor-critic for POMDPs. *SIAM Journal on Mathematics of Data Science*, 6(4):869–896, 2024.

[13] Ameya Anjarlekar, S Rasoul Etesami, and R Srikant. Scalable policy-based RL algorithms for POMDPs. *Advances in Neural Information Processing Systems*, 2025.

[14] Philip Jordan and Maryam Kamgarpour. Model-based learning of near-optimal finite-window policies in POMDPs. *arXiv preprint arXiv:2604.01024*, 2026.

[15] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 136–143. IEEE, 2004.

[16] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

[17] Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7187–7194, 2020.

[18] Runyu Lu, Yuanheng Zhu, and Dongbin Zhao. Divergence-regularized discounted aggregation: Equilibrium finding in multiplayer partially observable stochastic games. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable Markov games. *Advances in Neural Information Processing Systems*, 2022.

[20] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

[21] Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for Markov potential games. In *International Conference on Learning Representations*, 2018.

[22] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022.

[23] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2022.

[24] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

[25] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. Independent and decentralized learning in Markov potential games. *IEEE Transactions on Automatic Control*, 2025.

[26] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research*, 27(4): 819–840, 2002.

[27] Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research*, 55:443–497, 2016.

[28] Jilles Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable MDPs. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2018.

[29] Weichao Mao, Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Decentralized learning of finite-memory policies in Dec-POMDPs. *IFAC-PapersOnLine*, 56(2):2601–2607, 2023.

[30] Xiangyu Liu and Kaiqing Zhang. Partially observable multi-agent rl with (quasi-) efficiency: The blessing of information sharing. In *International Conference on Machine Learning*, pages 22370–22419. PMLR, 2023.

[31] Eitan Altman, Konstantin Avratchenkov, Nicolas Bonneau, Mérouane Debbah, Rachid El-Azouzi, and Daniel Sadoc Menasché. Constrained stochastic games in wireless networks. In *IEEE GLOBECOM 2007-IEEE Global Telecommunications Conference*, pages 315–320. IEEE, 2007.

[32] Eitan Altman, Konstantin Avrachenkov, Ishai Menache, Gregory Miller, Balakrishna J Prabhu, and Adam Shwartz. Dynamic discrete power control in cellular networks. *IEEE Transactions on Automatic Control*, 54(10):2328–2340, 2009.

[33] S Rasoul Etesami, Walid Saad, Narayan B Mandayam, and H Vincent Poor. Stochastic games for the smart grid energy management with prospect prosumers. *IEEE Transactions on Automatic Control*, 63(8):2327–2342, 2018.

[34] Runyu Zhang, Yuyang Zhang, Rohit Konda, Bryce Ferguson, Jason Marden, and Na Li. Markov games with decoupled dynamics: Price of anarchy and sample complexity. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 8100–8107. IEEE, 2023.

[35] S Rasoul Etesami. Learning stationary nash equilibrium policies in n-player stochastic games with independent chains. *SIAM Journal on Control and Optimization*, 62(2):799–825, 2024.

[36] Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov α-potential games. *IEEE Transactions on Automatic Control*, 2025.

[37] Xin Guo, Xinyu Li, and Yufei Zhang. An α-potential game framework for n-player dynamic games. *SIAM Journal on Control and Optimization*, 63(4):2964–3005, 2025.

[38] 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. *Advances in Neural Information Processing Systems*, 2022.

[39] Dheeraj Narasimha, Kiyeob Lee, Dileep Kalathil, and Srinivas Shakkottai. Multi-agent learning via Markov potential games in marketplaces for distributed energy resources. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6350–6357. IEEE, 2022.

[40] Ramon van Handel. Hidden Markov models. *Unpublished lecture notes*, 2008.

[41] Ali Kara and Serdar Yuksel. Near optimality of finite memory feedback policies in partially observed Markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022.

[42] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[43] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

[44] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

[45] Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in Markov games. *Advances in Neural Information Processing Systems*, 2005.

[46] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert Kappen. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning*, 2012.

[47] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

[48] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[49] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Supplementary Material

Table of Contents

A Overview of Notation **15**

B Discussion of decoupled dynamics assumption **16**

B.1 Illustrative examples 16

C Proofs for Section 3 . . . . . **16**

C.1 Finite window approximation results 16

C.2 Nash equilibrium approximation 19

C.3 Near-potential structure 19

C.4 Auxiliary lemma 21

D Algorithm details **22**

E Proofs for Section 5 . . . . . **22**

E.1 Model estimation guarantees 22

F Numerical experiments **30**

F.1 Dynamic transmission game 30

F.2 Cooperative gridworld 31

A Overview of Notation

Table 1 gives an overview of the notation used throughout this paper. All notations are introduced before their first use as well.

Table 1: Overview of notation

Symbol	Description
	POMG
$\mathcal{P} = (\mathcal{N}, H, \mathcal{S}, \mathcal{A}, \mathcal{O}, r, \mathbb{P}, \mathbb{O}, \mu)$	the partially observable Markov game
$\mathcal{N} = [N]$	set of N players
H	planning horizon
$\mathcal{S} = \prod_i \mathcal{S}_i$	joint state space
$\mathcal{A} = \prod_i \mathcal{A}_i$	joint action space; $A \coloneqq \max_i \lvert\mathcal{A}_i\rvert$
$\mathcal{O} = \prod_i \mathcal{O}_i$	joint observation space; $O \coloneqq \max_i \lvert\mathcal{O}_i\rvert$
$r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0,1]$	reward function for player i at step h
$\mathbb{P}_{i,h}(\cdot \mid s_i, a_i)$	transition kernel for player i at step h
$\mathbb{O}_{i,h}(\cdot \mid s_i)$	observation kernel for player i at step h
$\mu = \prod_i \mu_i$	initial state distribution over $\mathcal{S}$
	Histories and Policies
$\mathcal{H}_i^h = (\mathcal{O}_i \times \mathcal{A}_i)^{h-1} \times \mathcal{O}_i$	set of h-step local histories for player i
$\mathcal{H}_i^{\leq h}$	$\bigcup_{h'=0}^{h} \mathcal{H}_i^{h'}$
$\tau_i = (o_{i,1}, a_{i,1}, \ldots, a_{i,h-1}, o_{i,h})$	local history of length $\lvert\tau_i\rvert = h$; $\circ$ denotes concatenation
$\tau_{i,k:l}$	sub-history from step k to step l (inclusive)
Π_i^m	m-step finite-window policies for player i
$V_i(\pi)$	value of player i under joint policy π in $\mathcal{P}$
	Superstate Markov Game $\mathcal{G}^m$
$w_i \in \mathcal{H}_i^{\leq m}$	local superstate (truncated history window) of player i
$b_{i,h}^m(s_i \mid w_i)$	player i's belief over $\mathcal{S}_i$ induced by window w_i at step h
$b_{i,h}(s_i \mid \tau_i)$	player i's belief over $\mathcal{S}_i$ given full history τ_i
$\mathbb{P}_h^m(w' \mid w, a)$	superstate transition kernel of $\mathcal{G}^m$
$r_{i,h}^m(w, a)$	reward of player i in $\mathcal{G}^m$
$V_i^m(\pi)$	value of player i under π in $\mathcal{G}^m$
$\overline{Q}_{i,h}^{m,\pi}(w_i, a_i)$	marginalized Q-function in $\mathcal{G}^m$ under policy π
$d_{i,h}^{m,\pi}(w_i)$	superstate visitation distribution at step h under π
	Potential Structure
$\lambda_{i,h}^{\pi_i}(s_i, a_i)$	state-action occupancy measure of player i
$\Phi : \Pi \to \mathbb{R}$	potential function of the underlying Markov game
$\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$	stage-wise potential function at step h
$\Psi : \Pi^m \to \mathbb{R}$	near-potential function for $\mathcal{G}^m$ (see Proposition 2)
	Assumptions and Key Parameters
$\rho > 0$	filter stability contraction factor (see Assumption 2)
$\beta > 0$	minimum observation probability (see Assumption 3)
$\kappa \geq 1$	superstate concentrability constant (see Assumption 4)
$\zeta \in (0, 1]$	exploration probability of Algorithm 1
$\epsilon_\rho^m \coloneqq 4H^2(1-\rho)^{m-1}$	finite-window approximation error (see Lemma 1)
	Algorithm 1
T	number of episodes per iteration
K	number of iterations
$\eta^{(k)}$	stepsize at iteration k
$\hat{\mathbb{P}}_{i,h}^{m,\pi_i},\ \hat{r}_{i,h}^{m,\pi_{-i}}$	empirical estimates of superstate transitions and rewards
$\hat{Q}_{i,h}^{(k)}$	estimated Q-function at iteration k

B Discussion of decoupled dynamics assumption

B.1 Illustrative examples

We provide several representative application domains in which the assumption of decoupled transition dynamics arises naturally, while interactions are captured through the reward structure.

Wireless networks. Consider a wireless communication network in which N users aim to transmit packets over a shared channel. Each user maintains a local queue of packets and, at each time step, selects a transmission power level. The local state corresponds to the queue length, which evolves based on the transmission success and exogenous packet arrivals. Crucially, under standard modeling assumptions, the queue dynamics depend only on the user's own state and action, and are thus conditionally independent across users.

In contrast, rewards depend on the joint action profile: simultaneous high-power transmissions lead to increased interference and congestion, reducing transmission success probabilities. This induces a coupling in the reward function, often leading to a (statewise) congestion (and thus potential) structure. Such models have been studied in the networking literature (see, e.g., [31, 32]). In practice, the queue length may only be partially observed due to delays, noise, or aggregation effects, leading naturally to a partially observable setting.

Distributed energy markets. In distributed energy systems [33], multiple prosumers (e.g., households equipped with photovoltaic panels and storage) interact through a shared energy market. Each prosumer's local state is given by its battery level, which evolves according to locally generated and consumed energy, as well as charging and discharging decisions. Assuming participants are geographically separated with independent local conditions, these state dynamics may be modelled as decoupled across agents.

However, the reward, typically representing negative energy cost, depends on the aggregate supply and demand in the market, and hence on the joint actions of all agents. This induces strategic coupling through prices or clearing mechanisms, while maintaining independent state evolution.

Multi-robot systems. Decoupled transition dynamics also naturally arise in multi-robot systems and autonomous driving. Each agent (robot or vehicle) evolves according to its own local dynamics, which depend primarily on its state and control inputs. Coupling between agents typically enters through the reward function, which encodes coordination objectives such as collision avoidance, or task allocation. In many practical settings, agents have only partial observations of the local state due to limited sensing. For instance, [3] implements reinforcement learning methods under partial observability in real-world multi-robot tasks. Here, agents learn localized policies based on their individual state observations.

C Proofs for Section 3

In this section, we provide the proofs for our structural results on POMGs presented in Section 3.

C.1 Finite window approximation results

In this subsection, we establish several approximation results under filter stability that lead us towards proving Proposition 1.

First, we show that beliefs formed from joint histories corresponding to the same superstate must be close to each other in terms of total variation distance.

Lemma 3. *For $m, h, h' \in \mathbb{N}$ with $h, h' \geq m$, let $\tau \in \mathcal{H}^h, \tau' \in \mathcal{H}^{h'}$ such that $\tau_{h-m+1:h} = \tau'_{h'-m+1:h'}$. Then, under Assumption 2, we have*

$$\|b_h(\cdot \mid \tau) - b_{h'}(\cdot \mid \tau')\|_{TV} \leq (1-\rho)^{m-1}.$$

Proof. Since the POMG can be viewed as an POMDP with action space $\mathcal{A}$ and observation space $\mathcal{O}$, the result follows equivalently to the respective single-agent result, see Lemma 1 in [13]. An

m-step window consists of one observation followed by $m-1$ appended action-observation pairs, so Assumption 2 is applied $m-1$ times, and hence the desired bound follows. □

Remark. Although Assumption 2 is stated for the joint belief, it hides no exponential dependence on the number of players. It enters our analysis only through the belief bound of Lemma 3. If filter stability is instead assumed playerwise with parameter ρ_i for each $i \in \mathcal{N}$, setting $\rho \coloneqq \min_{i\in\mathcal{N}} \rho_i$, using decoupledness and the fact that $\|\prod_i P_i - \prod_i Q_i\|_{TV} \le \sum_i \|P_i - Q_i\|_{TV}$ for product measures, a linear N-dependence is introduced which gets absorbed in the big-$\mathcal{O}$ notation of Theorem 1.

Next, we bound the error in terms of superstate rewards and transition kernels resulting from finite-window approximation. Recall that $\circ$ denotes concatenation of histories.

Lemma 4. *Under Assumption 2, let $h \ge m$ and $w \in \mathcal{H}^m$, $\tau \in \mathcal{H}^h$ with $w = \tau_{h-m+1:h}$. Then, for any $i \in \mathcal{N}$, $h \in [H]$, and $a \in \mathcal{A}$,*

$$\left|r^m_{i,h}(w,a) - r^H_{i,h}(\tau,a)\right| \le 2(1-\rho)^{m-1},$$
$$\sum_{o\in\mathcal{O}} \left|\mathbb{P}^m_h(w_{2:m}\circ(a,o) \mid w,a) - \mathbb{P}^H_h(\tau_{h-m+1:h}\circ(a,o) \mid \tau,a)\right| \le 2(1-\rho)^{m-1}.$$

Proof. For the first bound, note that

$$\left|r^m_{i,h}(w,a) - r^H_{i,h}(\tau,a)\right| \le \sum_{s\in\mathcal{S}} r_{i,h}(s,a)\left|b^m_h(s \mid w) - b_h(s \mid \tau)\right| \le 2(1-\rho)^{m-1}$$

where in the last step we have used Lemma 3 and the fact that rewards are bounded by 1.

For the second bound, denoting $w'_o \coloneqq w_{2:m}\circ(a,o)$ and $\tau'_o \coloneqq \tau\circ(a,o)$, we have

$$\begin{aligned}
&\sum_{o\in\mathcal{O}} \left|\mathbb{P}^m_h(w'_o \mid w,a) - \mathbb{P}^H_h(\tau'_o \mid \tau,a)\right| \\
&\quad= \sum_{o\in\mathcal{O}}\sum_{s,s'\in\mathcal{S}} \mathbb{O}_{h+1}(o \mid s')\mathbb{P}_h(s' \mid s,a)\left|b^m_h(s \mid w) - b_h(s \mid \tau)\right| \\
&\quad\le \sum_{s\in\mathcal{S}} \left|b^m_h(s \mid w) - b_h(s \mid \tau)\right| \sum_{o\in\mathcal{O}}\sum_{s'\in\mathcal{S}} \mathbb{O}_{h+1}(o \mid s')\mathbb{P}_h(s' \mid s,a) \\
&\quad\le 2(1-\rho)^{m-1}.
\end{aligned}$$

□

Using the above bound, we obtain the following approximation result for value functions and best responses, restated here for convenience.

Lemma 1. *Let $\pi \in \Pi^m$ and $\pi' \in \Pi^H$ such that for all $h \ge m$ and all $w \in \mathcal{H}^m$, $\tau \in \mathcal{H}^h$ with $w = \tau_{h-m+1:h}$, it holds that $\pi_h(\cdot \mid w) = \pi'_h(\cdot \mid \tau)$. Let*

$$\epsilon^m_\rho \coloneqq 4H^2(1-\rho)^{m-1}. \tag{4}$$

Under Assumption 2, for any $i \in \mathcal{N}$, we have

$$|V^m_i(\pi) - V_i(\pi')| \le \epsilon^m_\rho.$$

Moreover, for π and π' as above, it holds that

$$\left|\max_{\hat\pi_i\in\Pi^m_i} V^m_i(\hat\pi_i,\pi_{-i}) - \max_{\hat\pi'_i\in\Pi^H_i} V_i(\hat\pi'_i,\pi_{-i})\right| \le \epsilon^m_\rho.$$

Proof. The proof of the first bound proceeds by showing a stronger inequality by backward induction on h. For this, we define for any $w \in \mathcal{H}^{\le m}$ and $\tau \in \mathcal{H}^h$ the step h value functions

$$V_{i,h}(\pi;\tau) \coloneqq \mathbb{E}_{\pi,\, s_1\sim\mu}\left[\sum_{h'=h}^{H} r^m_{i,h'}(s_{h'},a_{h'}) \mid (o_1,a_1,\ldots,a_{h-1},o_h) = \tau\right]$$
$$V^m_{i,h}(\pi;w) \coloneqq \mathbb{E}_{\pi,\, s_1\sim\mu}\left[\sum_{h'=h}^{H} r^m_{i,h'}(w_{h'},a_{h'}) \mid w_h = w\right].$$

The inequality we aim to show for all $h \in [H]$ is

$$\left|V^m_{i,h}(\pi; w) - V_{i,h}(\pi; \tau)\right| \leq 4(1-\rho)^{m-1} H(H+1-h). \tag{6}$$

For $H+1$, (6) trivially holds. Suppose (6) holds for some $h \in [H]$. The Bellman expectation equations are given by

$$V_{i,h}(\pi'; \tau) = \sum_{a \in \mathcal{A}} \pi'_h(a \mid \tau) \left[r^H_{i,h}(\tau, a) + \sum_{o \in \mathcal{O}} \mathbb{P}^H_h(\tau \circ (a,o) \mid \tau, a) V_{i,h+1}(\pi'; \tau \circ (a,o)) \right],$$

$$V^m_{i,h}(\pi; w) = \sum_{a \in \mathcal{A}} \pi_h(a \mid w) \left[r^m_{i,h}(w, a) + \sum_{o \in \mathcal{O}} \mathbb{P}^m_h(w_{2:m} \circ (a,o) \mid w, a) V^m_{i,h+1}(\pi; w_{2:m} \circ (a,o)) \right].$$

Hence, we can write the difference as

$$\begin{aligned}
&\left|V^m_{i,h}(\pi; w) - V_{i,h}(\pi'; \tau)\right| \\
&\quad \leq \sum_{a \in \mathcal{A}} \pi_h(a \mid w) \Bigg[\left| r^m_{i,h}(w, a) - r^H_{i,h}(\tau, a) \right| \\
&\qquad + \sum_{o \in \mathcal{O}} \mathbb{P}^m_h(w_{2:m} \circ (a,o) \mid w, a) \left[V^m_{i,h+1}(\pi; w_{2:m} \circ (a,o)) - V_{i,h+1}(\pi'; \tau \circ (a,o)) \right] \\
&\qquad + \sum_{o \in \mathcal{O}} \left[\mathbb{P}^m_h(w_{2:m} \circ (a,o) \mid w, a) - \mathbb{P}^H_h(\tau \circ (a,o) \mid \tau, a) \right] V_{i,h+1}(\pi'; \tau \circ (a,o)) \Bigg] \\
&\quad \overset{(a)}{\leq} \sum_{a \in \mathcal{A}} \pi_h(a \mid w) \left[2(1-\rho)^{m-1} + \max_{o \in \mathcal{O}} \left| V^m_{i,h+1}(\pi; w_{2:m} \circ (a,o)) - V_{i,h+1}(\pi'; \tau \circ (a,o)) \right| + 2(1-\rho)^{m-1} H \right] \\
&\quad \leq 4H(1-\rho)^{m-1} + \max_{a \in \mathcal{A}, o \in \mathcal{O}} \left| V^m_{i,h+1}(\pi; w_{2:m} \circ (a,o)) - V_{i,h+1}(\pi'; \tau \circ (a,o)) \right| \\
&\quad \overset{(b)}{\leq} 4H(1-\rho)^{m-1} + 4(1-\rho)^{m-1} H(H-h) \\
&\quad \leq 4(1-\rho)^{m-1} H(H+1-h)
\end{aligned}$$

where in (a) we use Lemma 4, and in (b) we apply the induction hypothesis.

To show the second bound, define the best response functions for any $i \in \mathcal{N}$, $h \in [H]$, $w \in \mathcal{H}^{\leq m}$, and $\tau \in \mathcal{H}^h$,

$$\begin{aligned}
V^\dagger_{i,h}(\pi'_{-i}; \tau) &\coloneqq \max_{\hat{\pi}'_i \in \Pi^H_i} V_{i,h}(\hat{\pi}'_i, \pi_{-i}; \tau), \\
V^{m,\dagger}_{i,h}(\pi_{-i}; w) &\coloneqq \max_{\hat{\pi}_i \in \Pi^m_i} V^m_{i,h}(\hat{\pi}_i, \pi_{-i}; w).
\end{aligned}$$

By Bellman optimality, we have

$$\begin{aligned}
V^\dagger_{i,h}(\pi'_{-i}; \tau) = \max_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi'_{-i}(a_{-i} \mid \tau_{-i}) \Bigg[& r^H_{i,h}(\tau, a_i, a_{-i}) \\
& + \sum_{o \in \mathcal{O}} \mathbb{P}^H_h(\tau \circ (a_i, a_{-i}, o) \mid \tau, a_i, a_{-i}) V^\dagger_{i,h+1}(\pi'_{-i}; \tau \circ (a_i, a_{-i}, o)) \Bigg], \\
V^{m,\dagger}_{i,h}(\pi_{-i}; w) = \max_{a_i \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi_{-i}(a_{-i} \mid w_{-i}) \Bigg[& r^m_{i,h}(w, a_i, a_{-i}) \\
& + \sum_{o \in \mathcal{O}} \mathbb{P}^m_h(w_{2:m} \circ (a_i, a_{-i}, o) \mid w, a_i, a_{-i}) V^{m,\dagger}_{i,h+1}(\pi_{-i}; w_{2:m} \circ (a_i, a_{-i}, o)) \Bigg].
\end{aligned}$$

Therefore, using the same steps as above, we can decompose the difference as above and apply Lemma 4 to obtain

$$\begin{aligned}
&\left|V_{i,h}^{m,\dagger}(\pi_{-i};w) - V_{i,h}^{\dagger}(\pi'_{-i};\tau)\right| \\
&\quad\leq \max_{a_i\in\mathcal{A}_i}\sum_{a_{-i}\in\mathcal{A}_{-i}}\pi_{-i}(a_{-i}\mid w_{-i})\Bigg[4H(1-\rho)^{m-1} \\
&\qquad + \max_{o\in\mathcal{O}}\left|V_{i,h+1}^{m,\dagger}(\pi_{-i};w_{2:m}\circ(a_i,a_{-i},o)) - V_{i,h+1}^{\dagger}(\pi'_{-i};\tau\circ(a_i,a_{-i},o))\right|\Bigg] \\
&\quad\leq 4H(1-\rho)^{m-1} + \max_{a\in\mathcal{A},o\in\mathcal{O}}\left|V_{i,h+1}^{m,\dagger}(\pi_{-i};w_{2:m}\circ(a_i,a_{-i},o)) - V_{i,h+1}^{\dagger}(\pi'_{-i};\tau\circ(a_i,a_{-i},o))\right|.
\end{aligned}$$

Then, by an induction argument over h similar to above, we can show that for any $h\in[H]$,

$$\left|V_{i,h}^{m,\dagger}(\pi_{-i};w) - V_{i,h}^{\dagger}(\pi'_{-i};\tau)\right| \leq 4(1-\rho)^{m-1}H(H+1-h),$$

which implies the desired bound. □

C.2 Nash equilibrium approximation

As a direct application of Lemma 1, we obtain the following approximation guarantee for Nash equilibria for m-step history-dependent policies.

Proposition 1. *Let $\epsilon>0$, and let $\pi\in\Pi^m$ be an ϵ-approximate Nash equilibrium of the superstate Markov game $\mathcal{G}^m$. Under Assumption 2, π is an $(\epsilon+2\epsilon_\rho^m)$-approximate Nash equilibrium of $\mathcal{P}$.*

Proof. With best response function defined as in the proof of Lemma 1, we have

$$\begin{aligned}
V_i^{\dagger}(\pi_{-i}) - V_i(\pi) &\leq \left[V_i^{\dagger}(\pi_{-i}) - V_i^{m,\dagger}(\pi_{-i})\right] + \left[V_i^{m,\dagger}(\pi_{-i}) - V_i^{m}(\pi)\right] + \left[V_i^{m}(\pi) - V_i(\pi)\right] \\
&\overset{(a)}{\leq} \left|V_i^{\dagger}(\pi_{-i}) - V_i^{m,\dagger}(\pi_{-i})\right| + \epsilon + \left|V_i^{m}(\pi) - V_i(\pi)\right| \\
&\overset{(b)}{\leq} \epsilon + 2\epsilon_\rho^m
\end{aligned}$$

where (a) is due to π being an ϵ-approximate Nash equilibrium, and (b) follows from the two respective bounds in Lemma 1. □

C.3 Near-potential structure

Having shown our Nash equilibrium approximation guarantee, we next establish near-potential structure of the superstate Markov game $\mathcal{G}^m$ in terms of the value functions V_i^m, for $i\in\mathcal{N}$.

Proposition 2. *Under Assumptions 1 and 2, there exists a function $\Psi:\Pi^m\to\mathbb{R}$ such that for all $\pi\in\Pi^m$, $i\in\mathcal{N}$, and $\pi'_i\in\Pi_i^m$, and ϵ_ρ^m as in (4), we have*

$$|(V_i^m(\pi) - V_i^m(\pi'_i,\pi_{-i})) - (\Psi(\pi) - \Psi(\pi'_i,\pi_{-i}))| \leq 2\epsilon_\rho^m.$$

Proof. The proof is divided into two parts. First, using decoupledness of transitions, we show that under Assumption 1 on the underlying latent state Markov game, there exists an exact potential function for the POMG. As a second step, based on this, we then prove the $(2\epsilon_\rho^m)$-approximate potential structure for the superstate Markov game.

- **POMG is potential:** For any m-step policy $\pi_i\in\Pi_i^m$ and $h\in[H]$, let $\lambda_{i,h}^{\pi_i}\in\Delta(\mathcal{S}_i\times\mathcal{A}_i)$ denote the induced playerwise state-action visitation probability at step h, defined as

  $$\lambda_{i,h}^{\pi_i}(s_i,a_i) \coloneqq P_{\pi_i}\left(s_{i,h}=s_i, a_{i,h}=a_i\right)$$

  where P_{π_i} refers to the distribution over player i's (unobserved) state trajectories when following π_i in the underlying POMG. Analogously, for a joint policy $\pi\in\Pi^m$, let $\lambda_h^{\pi}(s,a) \coloneqq P_\pi(s_h=s,a_h=a)\in\Delta(\mathcal{S}\times\mathcal{A})$ denote the induced joint state-action

occupancy at step h. Since the initial distribution μ, the transitions $\mathbb{P}$, the observations $\mathbb{O}$, and the product policy π all factor across players, each player's trajectory evolves independently, so the joint occupancy factorizes into the playerwise marginals, i.e.,

$$\lambda_h^{\pi}(s,a) = \prod_{j\in\mathcal{N}} \lambda_{j,h}^{\pi_j}(a_i, s_i)$$

for any $a \in \mathcal{A}$, $s \in \mathcal{S}$, and $h \in [H]$.

By Assumption 1, there exist $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each $h \in [H]$ such that for all $i \in \mathcal{N}$, $h \in [H]$, $s \in \mathcal{S}$, $s_i' \in \mathcal{S}_i$, $a \in \mathcal{A}$, and $a_i' \in \mathcal{A}_i$, we have

$$\begin{aligned} &r_{i,h}(s_{i,h}', s_{-i,h}, a_{i,h}', a_{-i,h}) - r_{i,h}(s_{i,h}, s_{-i,h}, a_{i,h}, a_{-i,h}) \\ &\qquad = \phi_h(s_{i,h}', s_{-i,h}, a_{i,h}', a_{-i,h}) - \phi_h(s_{i,h}, s_{-i,h}, a_{i,h}, a_{-i,h}). \end{aligned}$$

Define the function $\Phi : \Pi^m \to \mathbb{R}$ as

$$\Phi(\pi) \coloneqq \mathbb{E}_{\pi, s_1\sim\mu}\left[\sum_{h=1}^{H} \phi_h(s_h, a_h)\right],$$

where the expectation is taken over the distribution of trajectories induced by the POMG $\mathcal{P}$ under joint policy $\pi \in \Pi^m$, including randomness from state transitions according to $\mathbb{P}$ and observations emissions according to $\mathbb{O}$.

Further define for any $i \in \mathcal{N}$, $\pi_{-i} \in \Pi_{-i}^m$, $h \in [H]$, $s_i \in \mathcal{S}_i$, and $a_i \in \mathcal{A}_i$,

$$\begin{aligned} r_{i,h}^{\pi_{-i}}(s_i, a_i) &\coloneqq \sum_{s_{-i}, a_{-i}} r_{i,h}(s_i, s_{-i}, a_i, a_{-i}) \prod_{j\in\mathcal{N}\setminus\{i\}} \lambda_{j,h}^{\pi_j}(s_j, a_j), \\ \phi_h^{\pi_{-i}}(s_i, a_i) &\coloneqq \sum_{s_{-i}, a_{-i}} \phi_h(s_i, s_{-i}, a_i, a_{-i}) \prod_{j\in\mathcal{N}\setminus\{i\}} \lambda_{j,h}^{\pi_j}(s_j, a_j). \end{aligned}$$

Then, for any m-step policy $\pi \in \Pi^m$ and deviation $\pi_i' \in \Pi_i^m$, with the POMG value function defined as in (1), we can write

$$\begin{aligned} &V_i(\pi) - V_i(\pi_i', \pi_{-i}) \\ &\overset{(a)}{=} \sum_{h\in[H]} r_{i,h}^{\top}\lambda_h^{\pi} - \sum_{h\in[H]} r_{i,h}^{\top}\lambda_h^{(\pi_i', \pi_{-i})} \\ &\overset{(b)}{=} \sum_{h\in[H]} r_{i,h}^{\top}\left(\bigtimes_{j\in\mathcal{N}} \lambda_{j,h}^{\pi_j}\right) - \sum_{h\in[H]} r_{i,h}^{\top}\left(\lambda_{i,h}^{\pi_i'} \times \left(\bigtimes_{j\in\mathcal{N}\setminus\{i\}} \lambda_{j,h}^{\pi_j}\right)\right) \\ &= \sum_{h\in[H]} \left(r_{i,h}^{\pi_{-i}}\right)^{\top} \lambda_{i,h}^{\pi_i} - \left(r_{i,h}^{\pi_{-i}}\right)^{\top} \lambda_{i,h}^{\pi_i'} \\ &\overset{(c)}{=} \frac{1}{2} \sum_{h\in[H]} \sum_{s_i, s_i', a_i, a_i'} \left(\lambda_{i,h}^{\pi_i}(s_i, a_i)\lambda_{i,h}^{\pi_i'}(s_i', a_i') - \lambda_{i,h}^{\pi_i}(s_i', a_i')\lambda_{i,h}^{\pi_i'}(s_i, a_i)\right)\left(r_{i,h}^{\pi_{-i}}(s_i, a_i) - r_{i,h}^{\pi_{-i}}(s_i', a_i')\right) \\ &= \frac{1}{2} \sum_{h\in[H]} \sum_{s_i, s_i', a_i, a_i'} \left(\lambda_{i,h}^{\pi_i}(s_i, a_i)\lambda_{i,h}^{\pi_i'}(s_i', a_i') - \lambda_{i,h}^{\pi_i}(s_i', a_i')\lambda_{i,h}^{\pi_i'}(s_i, a_i)\right)\left(\phi_h^{\pi_{-i}}(s_i, a_i) - \phi_h^{\pi_{-i}}(s_i', a_i')\right) \\ &= \sum_{h\in[H]} \left(\phi_h^{\pi_{-i}}\right)^{\top} \lambda_{i,h}^{\pi_i} - \left(\phi_h^{\pi_{-i}}\right)^{\top} \lambda_{i,h}^{\pi_i'} \\ &= \sum_{h\in[H]} \phi_h^{\top}\left(\bigtimes_{j\in\mathcal{N}} \lambda_{j,h}^{\pi_j}\right) - \phi_h^{\top}\left(\lambda_{i,h}^{\pi_i'} \times \left(\bigtimes_{j\in\mathcal{N}\setminus\{i\}} \lambda_{j,h}^{\pi_j}\right)\right) \\ &= \Phi(\pi) - \Phi(\pi_i', \pi_{-i}) \end{aligned}$$

where (a) writes each value as a linear form in the joint state-action occupancy, which holds by linearity of expectation since each $r_{i,h}$ depends only on (s_h, a_h); (b) factorizes the

joint occupancy into the playerwise marginals, $\lambda_h^\pi = \bigtimes_{j\in\mathcal{N}} \lambda_{j,h}^{\pi_j}$, by decoupledness of the initial distribution, transitions, observations, and policies; and (c) uses the identity proven in Lemma 5. We have established potential structure with potential function Φ over the policy class Π^m for the POMG $\mathcal{P}$. Next, we use this fact, together with our finite-window approximation guarantees relating the value functions V_i and V_i^m, to show near-potential structure for the superstate Markov game.

- **Superstate Markov game is near-potential:** For any $\pi \in \Pi^m$, $i \in \mathcal{N}$, and $\pi_i' \in \Pi_i^m$,

$$\begin{aligned}(V_i^m(\pi) - V_i^m(\pi_i', \pi_{-i})) - (\Phi(\pi) - \Phi(\pi_i', \pi_{-i}))& \\ = \underbrace{(V_i^m(\pi) - V_i(\pi)) + (V_i(\pi_i', \pi_{-i}) - V_i^m(\pi_i', \pi_{-i}))}_{(a)}& \\ + \underbrace{(V_i(\pi) - V_i(\pi_i', \pi_{-i})) - (\Phi(\pi) - \Phi(\pi_i', \pi_{-i}))}_{(b)}.&\end{aligned}$$

  Note that since Φ is a potential function of the POMG $\mathcal{P}$, we have $(b) = 0$. Moreover, using Lemma 1, we bound $|(a)| \le 2\epsilon_\rho^m$, which concludes the proof.

□

C.4 Auxiliary lemma

In the above proof of Proposition 2, we invoked the following algebraic identity, which expresses the difference of two linear forms over distributions as a symmetric pairwise sum. Here we provide its proof.

Lemma 5. *For $k \in \mathbb{N}$ and $\mathcal{I} = \{1, \dots, k\}$, let $p, q \in \Delta(\mathcal{I})$, and let $x : \mathcal{I} \to \mathbb{R}$ be arbitrary. Then*

$$\sum_{i\in\mathcal{I}} p(i)x(i) - \sum_{i\in\mathcal{I}} q(i)x(i) = \frac{1}{2}\sum_{i,j\in\mathcal{I}} \big(p(i)q(j) - p(j)q(i)\big)\big(x(i) - x(j)\big).$$

Proof. Expanding the right-hand side yields

$$\frac{1}{2}\sum_{i,j} p(i)q(j)x(i) - \frac{1}{2}\sum_{i,j} p(i)q(j)x(j) - \frac{1}{2}\sum_{i,j} p(j)q(i)x(i) + \frac{1}{2}\sum_{i,j} p(j)q(i)x(j).$$

We first group terms according to whether they multiply $x(i)$ or $x(j)$:

$$\sum_{i,j} \frac{1}{2}\big(p(i)q(j) - p(j)q(i)\big)x(i) + \sum_{i,j} \frac{1}{2}\big(p(j)q(i) - p(i)q(j)\big)x(j).$$

For the first sum, fixing i and summing over j yields

$$\sum_j \frac{1}{2}\big(p(i)q(j) - p(j)q(i)\big) = \frac{1}{2}\Big(p(i)\sum_j q(j) - q(i)\sum_j p(j)\Big) = \frac{1}{2}\big(p(i) - q(i)\big),$$

where we used $\sum_j p(j) = \sum_j q(j) = 1$. Hence

$$\sum_{i,j} \frac{1}{2}\big(p(i)q(j) - p(j)q(i)\big)x(i) = \sum_i \frac{1}{2}\big(p(i) - q(i)\big)x(i).$$

Applying the same argument to the second sum (with indices relabeled) gives

$$\sum_{i,j} \frac{1}{2}\big(p(j)q(i) - p(i)q(j)\big)x(j) = \sum_i \frac{1}{2}\big(q(i) - p(i)\big)x(i).$$

Combining the two expressions and canceling terms yields

$$\sum_i (p(i) - q(i))x(i) = \sum_i p(i)x(i) - \sum_i q(i)x(i),$$

which completes the proof. □

D Algorithm details

Below, we provide further details omitted in the main part by formally specifying how to compute empirical frequencies for estimating superstate transition probabilities and rewards based on the sampled action-observation sequences. Denote by $\tau_i^{(t)} = (a_{i,h}^{(t)}, o_{i,h}^{(t)}, r_{i,h}^{(t)})_{h=1}^{H}$ for $t \in [T]$ the t-th sequence sampled at some iteration $k \in [K]$ of Algorithm 1 when players follow the joint policy $\pi^{(k)} \in \Pi^m$.

Transition probabilities. At each player $i \in \mathcal{N}$, at step $h \in [H]$, and for all $w_i, w_i' \in \mathcal{H}_i^{\leq m}$ and $a_i \in \mathcal{A}_i$, take the empirical average of transitions from w_i to w_i', that is,

$$\hat{\mathbb{P}}_{i,h}^{m,\pi_i^{(k)}}(w_i' \mid w_i, a_i) \coloneqq \begin{cases} \frac{\sum_{t=1}^{T} \mathbf{1}\{\tau^{(t)}_{i,h-|w_i'|+2\,:\,h+1} = w_i' \wedge \tau^{(t)}_{i,h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\}}{\sum_{t=1}^{T} \mathbf{1}\{\tau^{(t)}_{i,h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\}}, \\ \quad \text{if } \sum_{t=1}^{T} \mathbf{1}\{\tau^{(t)}_{h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\} \geq 1 \text{ and } |w_i'| = \min(m, |w_i|+1), \\ 0, \text{ otherwise.} \end{cases} \tag{7}$$

Rewards. Similarly, at each player $i \in \mathcal{N}$, at step $h \in [H]$, and for all $w_i \in \mathcal{H}_i^{\leq m}$ and $a_i \in \mathcal{A}_i$, set

$$\hat{r}_{i,h}^{m,\pi_{-i}^{(k)}}(w_i, a_i) \coloneqq \begin{cases} \frac{\sum_{t=1}^{T} r_{i,h}^{(t)} \cdot \mathbf{1}\{\tau^{(t)}_{i,h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\}}{\sum_{t=1}^{T} \mathbf{1}\{\tau^{(t)}_{h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\}}, \\ \quad \text{if } \sum_{t=1}^{T} \mathbf{1}\{\tau^{(t)}_{i,h-|w_i|+1\,:\,h} = w_i \wedge a_{i,h}^{(t)} = a_i\} \geq 1, \\ 0, \text{ otherwise.} \end{cases} \tag{8}$$

E Proofs for Section 5

In this section, we provide the proof our main result, Theorem 1, on the convergence of Algorithm 1 to an ϵ-approximate Nash equilibrium.

E.1 Model estimation guarantees

We begin by deriving confidence bounds for the transition and reward estimates computed at each iteration of Algorithm 1. Note that these differ from the Markov game case, since we additionally need to account for the bias introduced by the superstate Markov game approximation.

Lemma 2. *Let Assumptions 2 and 3 hold, and let $\pi \in \Pi^m$. Suppose each player $i \in \mathcal{N}$ collects a trajectory according to Line 4 in Algorithm 1 with exploration probability $\zeta \in (0, 1]$, and estimates $\hat{\mathbb{P}}_i^{m,\pi_i}$ and $\hat{r}_i^{m,\pi_{-i}}$ according to (7) and (8), respectively. Let $i \in \mathcal{N}$, $h \in [H]$, $a_i \in \mathcal{A}_i$, and $w_i, w_i' \in \mathcal{H}_i^{\leq m}$. Then for any $0 < \epsilon \leq 1$,*

$$P\left(|\mathbb{P}_{i,h}^{m}(w_i' \mid w_i, a_i) - \hat{\mathbb{P}}_{i,h}^{m,\pi_i}(w_i' \mid w_i, a_i)| \geq \epsilon + (1-\rho)^{m-1}\right) \leq 4\exp\left(-\frac{\epsilon^2 \zeta^{2m} \beta^{2m} T}{32|\mathcal{A}_i|^{2m}}\right),$$

$$P\left(|r_{i,h}^{m,\pi_{-i}}(w_i, a_i) - \hat{r}_{i,h}^{m,\pi_{-i}}(w_i, a_i)| \geq \epsilon + 4NH\zeta + N^2H(1-\rho)^{m-1}\right) \leq 4\exp\left(-\frac{\epsilon^2 \zeta^{2m} \beta^{2m} T}{32|\mathcal{A}_i|^{2m}}\right).$$

Proof. We divide the proof into two parts for proving each of the two bounds.

- **Bound for estimation of transitions:** For some $\pi_i \in \Pi_i^H$, let $\widetilde{\mathbb{P}}_{i,h}^{\pi_i}(\cdot \mid w_i, a_i) \in \Delta(\mathcal{H}_i^{\leq m})$ denote the distribution over m-step windows observed at step h marginalized over the full

history, i.e.,

$$\widetilde{\mathbb{P}}^{\pi_i}_{i,h}(w_i' \mid w_i, a_i) \coloneqq \sum_{\tau_i=(o_{i,1},a_{i,1},\ldots,a_{i,h},o_{i,h+1})\in\mathcal{H}_i^{h+1}} P_{\pi_i}\left(\mathcal{T}_{i,h}=\tau_i\right) \cdot \mathbf{1}\left\{\tau_{i,h-m+2:h+1}=w_i' \wedge \tau_{i,h-m+1:h}=w_i \wedge a_{i,h+1}=a_i\right\}$$

where $P_{\pi_i}\left(\mathcal{T}_{i,h}=\tau_i\right)$ is the probability that the random history $\mathcal{T}_{i,h}$, induced by the POMG dynamics and policy π_i, is equal to τ_i at step h for player i.

Note that

$$\mathbb{E}_{\pi_i,s_{i,1}\sim\mu_i}\left[\hat{\mathbb{P}}^{m,\pi_i}_{i,h}(w_i' \mid w_i, a_i)\right] = \widetilde{\mathbb{P}}^{\pi_i^\zeta}_{i,h}(w_i' \mid w_i, a_i),$$

where π_i^ζ is the ζ-greedy policy from which we sample in Line 4 of Algorithm 1. Being a ratio of two empirical counts over the T sampled trajectories, $\hat{\mathbb{P}}^{m,\pi_i}_{i,h}(w_i' \mid w_i, a_i)$ is itself not a sum of independent terms. Lemma 4 of [34] does not require this: it applies to empirical transition estimates of exactly this form, and is obtained by first rearranging the ratio into a quantity to which Hoeffding's inequality applies. Invoking it, for any $0 < \epsilon \leq 1$, we have

$$P\left(\left|\widetilde{\mathbb{P}}^{\pi_i^\zeta}_{i,h}(w_i' \mid w_i, a_i) - \hat{\mathbb{P}}^{m,\pi_i}_{i,h}(w_i' \mid w_i, a_i)\right| \geq \epsilon\right) \leq 4\exp\left(-\frac{\epsilon^2\zeta^{2m}\beta^{2m}T}{32A_i^{2m}}\right).$$

where the lower bound $c > 0$ on state visitation probabilities required by this Lemma in our case is given by $c = \left((\beta\zeta)/A_i\right)^m$. This is due to Assumption 3 and the fact that we sample trajectories in a ζ-greedy manner. At any step $h \in [H]$, any m-step window is visited with probability at least $\left((\beta\zeta)/A_i\right)^m$.

Moreover, by Lemma 4, we have the deterministic bound

$$\left|\widetilde{\mathbb{P}}^{\pi_i^\zeta}_{i,h}(w_i' \mid w_i, a_i) - \mathbb{P}^{m}_{i,h}(w_i' \mid w_i, a_i)\right| \leq 2(1-\rho)^{m-1}$$

which together with the above concludes the proof of the first bound.

- **Bound for reward estimation:** Similar to above, we define

$$\widetilde{r}^{\pi}_{i,h}(w_i, a_i) \coloneqq \sum_{\tau_i=(o_{i,1},a_{i,1},\ldots,a_{i,h-1},o_{i,h})\in\mathcal{H}_i^{h}} P_{\pi}\left(\mathcal{T}_{i,h}=\tau_i\right) \cdot \mathbf{1}\left\{\tau_{i,h-m+1:h}=w_i \wedge a_{i,h+1}=a_i\right\} r^{H,\pi_{-i}}_{i,h}(\tau_i, a_i),$$

such that we have

$$\mathbb{E}_{\pi,s_1\sim\mu}\left[\hat{r}^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right] = \widetilde{r}^{\pi_i^\zeta}_{i,h}(w_i, a_i),$$

Similar as for transition probabilities, we obtain from Lemma 4 of [34] the bound

$$P\left(\left|\widetilde{r}^{\pi_i^\zeta}_{i,h}(w_i, a_i) - \hat{r}^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right| \geq \epsilon\right) \leq 4\exp\left(-\frac{\epsilon^2\zeta^{2m}\beta^{2m}T}{32A_i^{2m}}\right).$$

Moreover, we decompose the bias as follows

$$\begin{aligned}&\left|\widetilde{r}^{\pi_i^\zeta}_{i,h}(w_i, a_i) - r^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right| \\ &\quad\leq \underbrace{\left|\widetilde{r}^{\pi_i^\zeta}_{i,h}(w_i, a_i) - \widetilde{r}^{\pi}_{i,h}(w_i, a_i)\right|}_{(a)} + \underbrace{\left|\widetilde{r}^{\pi}_{i,h}(w_i, a_i) - r^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right|}_{(b)}.\end{aligned}$$

Using Lemma 7 below, for the second term, we obtain

$$\begin{aligned}(b) &\leq \left|\widetilde{r}^{\pi}_{i,h}(w_i, a_i) - r^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right| \\ &\leq \sum_{\tau_i=(o_{i,1},a_{i,1},\ldots,a_{i,h-1},o_{i,h})\in\mathcal{H}_i^{h}} P\left(\mathcal{T}_{i,h}=\tau_i\right)\mathbf{1}\left\{\tau_{i,h-m+1:h}=w_i \wedge a_{i,h+1}=a_i\right\} \\ &\qquad\cdot\left|r^{H,\pi_{-i}}_{i,h}(\tau_i, a_i) - r^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right| \\ &\lesssim N^2H(1-\rho)^{m-1},\end{aligned}$$

and the first term is bounded by Lemma 6,

$$(a) \le 4H(N-1)\zeta.$$

Together with our probabilistic bound on $|\widetilde{r}^{\pi}_{i,h}(w_i, a_i) - \hat{r}^{m,\pi_{-i}}_{i,h}(w_i, a_i)|$, this concludes the proof.

□

Lemma 6. *Let $\pi \in \Pi^m$, and for $\zeta > 0$ define $\pi^\zeta \in \Pi^m$ such that*

$$\pi^\zeta_i(a_i \mid w_i) \coloneqq \frac{\zeta}{A_i} + (1-\zeta)\pi_i(a_i \mid w_i).$$

Let also

$$r^{m,\pi}_{i,h}(w_i, a_i) \coloneqq \mathbb{E}_{w_{-i}\sim d^{m,\pi}_{-i,h},\, a_{-i}\sim\pi_{-i}(\cdot|w_{-i})}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big].$$

Assume $r^m_{i,h} \in [0,1]$. Then, for any $w_i \in \mathcal{H}^{\le m}_i$ and $a_i \in \mathcal{A}_i$,

$$\big|r^{m,\pi}_{i,h}(w_i, a_i) - r^{m,\pi^\zeta}_{i,h}(w_i, a_i)\big| \le 4H(N-1)\zeta.$$

Proof. Fix w_i, a_i. We write

$$\begin{aligned}
&\big|r^{m,\pi}_{i,h}(w_i, a_i) - r^{m,\pi^\zeta}_{i,h}(w_i, a_i)\big| \\
&= \left|\mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi}_{-i,h}\\ a_{-i}\sim\pi_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big] - \mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi^\zeta}_{-i,h}\\ a_{-i}\sim\pi^\zeta_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big]\right| \\
&\le \underbrace{\left|\mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi}_{-i,h}\\ a_{-i}\sim\pi_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big] - \mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi}_{-i,h}\\ a_{-i}\sim\pi^\zeta_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big]\right|}_{(a)} \\
&\quad + \underbrace{\left|\mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi}_{-i,h}\\ a_{-i}\sim\pi^\zeta_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big] - \mathbb{E}_{\substack{w_{-i}\sim d^{m,\pi^\zeta}_{-i,h}\\ a_{-i}\sim\pi^\zeta_{-i}(\cdot|w_{-i})}}\big[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})\big]\right|}_{(b)}.
\end{aligned}$$

For term (a), since $r^m_{i,h} \in [0,1]$, for any fixed w_{-i},

$$\begin{aligned}
&\Big|\mathbb{E}_{a_{-i}\sim\pi_{-i}(\cdot|w_{-i})}[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})] - \mathbb{E}_{a_{-i}\sim\pi^\zeta_{-i}(\cdot|w_{-i})}[r^m_{i,h}(w_i, w_{-i}, a_i, a_{-i})]\Big| \\
&\qquad \le 2\|\pi_{-i}(\cdot \mid w_{-i}) - \pi^\zeta_{-i}(\cdot \mid w_{-i})\|_{TV} \\
&\qquad \le 2(N-1)\zeta,
\end{aligned}$$

and hence $(a) \le 2(N-1)\zeta$.

For the second term, we have $(b) \le \|d^{m,\pi}_{-i,h} - d^{m,\pi^\zeta}_{-i,h}\|_{TV}$. By decoupledness of dynamics, $d^{m,\pi}_{-i,h}$ factorizes across players, and hence by induction on $h \in [H]$ similar to the proof of Lemma 1, we obtain a bound of

$$(b) \le 2\sum_{j\in\mathcal{N}\setminus\{i\}} \|d^{m,\pi}_{j,h} - d^{m,\pi^\zeta}_{j,h}\|_{TV} \le 2H(N-1)\zeta.$$

Combining the bounds on (a) and (b) gives the claimed inequality. □

Lemma 7. *Suppose Assumption 2 holds. Let $\pi \in \Pi^m$, and for any $i \in \mathcal{N}$, $h \in [H]$, $w_i \in \mathcal{H}_i^{\leq m}$, $\tau_i \in \mathcal{H}_i^h$, and $a_i \in \mathcal{A}_i$, recall that marginal rewards are defined as*

$$r_{i,h}^{m,\pi_{-i}}(w_i, a_i) \coloneqq \mathbb{E}_{w_{-i} \sim d_{-i,h}^{m,\pi}, a_{-i} \sim \pi_{-i}(\cdot|w_{-i})} \left[r_{i,h}^m(w_i, w_{-i}, a_i, a_{-i}) \right],$$
$$r_{i,h}^{H,\pi_{-i}}(\tau_i, a_i) \coloneqq \mathbb{E}_{\tau_{-i} \sim d_{-i,h}^{H,\pi}, a_{-i} \sim \pi_{-i}(\cdot|\tau_{-i})} \left[r_{i,h}^H(\tau_i, \tau_{-i}, a_i, a_{-i}) \right].$$

Then, for any $w_i \in \mathcal{H}_i^{\leq m}$ and $\tau_i \in \mathcal{H}_i^H$ with $\tau_{i,h-m+1:h} = w_i$, it holds that

$$\left| r_{i,h}^{m,\pi_{-i}}(w_i, a_i) - r_{i,h}^{H,\pi_{-i}}(\tau_i, a_i) \right| \lesssim N^2 H (1-\rho)^{m-1}.$$

Proof. Let $h \in [H]$. We decompose the difference as follows,

$$\begin{aligned}
&\left| r_{i,h}^{m,\pi_{-i}}(w_i, a_i) - r_{i,h}^{H,\pi_{-i}}(\tau_i, a_i) \right| \\
&= \Bigg| \sum_{w_{-i} \in \mathcal{H}_{-i}^{\leq m}, a_{-i} \in \mathcal{A}_{-i}} d_{-i,h}^{m,\pi}(w_{-i}) \cdot \pi_{-i}(a_{-i} \mid w_{-i}) \cdot r_{i,h}^m(w_i, w_{-i}, a_i, a_{-i}) \\
&\qquad - \sum_{\tau_{-i} \in \mathcal{H}_{-i}^H, a_{-i} \in \mathcal{A}_{-i}} d_{-i,h}^{H,\pi}(\tau_{-i}) \cdot \pi_{-i}(a_{-i} \mid \tau_{-i}) \cdot r_{i,h}^H(\tau_i, \tau_{-i}, a_i, a_{-i}) \Bigg| \\
&\overset{(a)}{\leq} \Bigg| \sum_{w_{-i} \in \mathcal{H}_{-i}^{\leq m}, a_{-i} \in \mathcal{A}_{-i}} d_{-i,h}^{m,\pi}(w_{-i}) \cdot \pi_{-i}(a_{-i} \mid w_{-i}) \cdot r_{i,h}^m(w_i, w_{-i}, a_i, a_{-i}) \\
&\qquad - \sum_{\tau_{-i} \in \mathcal{H}_{-i}^H, a_{-i} \in \mathcal{A}_{-i}} d_{-i,h}^{H,\pi}(\tau_{-i}) \cdot \pi_{-i}(a_{-i} \mid \tau_{-i}) \cdot r_{i,h}^m(\tau_{i,h-m+1:h}, \tau_{-i,h-m+1:h}, a_i, a_{-i}) \Bigg| \\
&\qquad + 2N(1-\rho)^{m-1} \\
&\overset{(b)}{\leq} \sum_{w_{-i} \in \mathcal{H}_{-i}^{\leq m}} \left| d_{-i,h}^{m,\pi}(w_{-i}) - \sum_{\tau_{-i,1:h-m} \in \mathcal{H}_{-i}^H} d_{-i,h}^{H,\pi}(\tau_{-i,1:h-m} \circ w_{-i}) \right| + 2N(1-\rho)^{m-1} \\
&\overset{(c)}{\leq} \sum_{j \in \mathcal{N} \setminus \{i\}} \sum_{w_j \in \mathcal{H}_j^{\leq m}} \left| d_{j,h}^{m,\pi}(w_j) - \sum_{\tau_{j,1:h-m} \in \mathcal{H}_j^H} d_{j,h}^{H,\pi}(\tau_{j,1:h-m} \circ w_j) \right| + 2N(1-\rho)^{m-1} \\
&\overset{(d)}{\leq} N \cdot HN(1-\rho)^{m-1} + 2N(1-\rho)^{m-1},
\end{aligned}$$

where

- (a) uses the reward approximation bound from Lemma 4,
- (b) is by boundedness of rewards and rearraging the summations,
- (c) holds since by decoupledness of state transitions, we can factor state visitation distributions over players, and
- (d) follows from bounding the error in state visitation distributions induced by $\mathbb{P}_i^m$ vs. $\mathbb{P}_i^H$ inductively over the H steps (similar to the proofs in Lemma 1), and using the transition probability approximation bound from Lemma 4.

□

Next, we provide an approximation bound for the Q function obtained from backward dynamic programming in terms of reward and transition probability estimation errors.

Lemma 8. *Suppose the bounds on transition and reward estimates from Lemma 2 hold, namely, we have*

$$\left|\mathbb{P}^m_{i,h}(w_i' \mid w_i, a_i) - \hat{\mathbb{P}}^{m,\pi_i}_{i,h}(w_i' \mid w_i, a_i)\right| \leq \epsilon + (1-\rho)^{m-1} =: \epsilon_r,$$
$$\left|r^{m,\pi_{-i}}_{i,h}(w_i, a_i) - \hat{r}^{m,\pi_{-i}}_{i,h}(w_i, a_i)\right| \leq \epsilon + 4NH\zeta + N^2H(1-\rho)^{m-1} =: \epsilon_{\mathbb{P}}.$$

Then, for all $i \in \mathcal{N}$, $h \in [H]$, $k \in [K]$, $w_i \in \mathcal{H}_i^{\leq m}$, and $a_i \in \mathcal{A}_i$, it holds that

$$\left|\hat{Q}^{m,\pi}_{i,h}(w_i, a_i) - \overline{Q}^{m,\pi}_{i,h}(w_i, a_i)\right| \lesssim H^3 N \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i| \cdot \left(\epsilon N + (1-\rho)^{m-1}\right).$$

Proof. The proof proceeds by showing a stronger inequality by backward induction on h, namely that

$$\left|\hat{Q}^{m,\pi}_{i,h}(w_i, a_i) - \overline{Q}^{m,\pi}_{i,h}(w_i, a_i)\right| \leq \epsilon_r(H+1-h) + \epsilon_{\mathbb{P}} H(H+1-h) \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i|. \tag{9}$$

For $H+1$, (9) trivially holds. Suppose (9) holds for some $h \in [H]$. We can write the error as

$$\begin{aligned}
\left|\left(\hat{Q}^{m,\pi}_{i,h} - \overline{Q}^{m,\pi}_{i,h}\right)(w_i, a_i)\right| \leq{}& \left|\left(\hat{r}^{m,\pi_{-i}}_{i,h} - r^{m,\pi_{-i}}_{i,h}\right)(w_i, a_i)\right| \\
&+ \sum_{w_i', a_i'} \mathbb{P}_{i,h}(w_i' \mid w_i, a_i) \cdot \pi_{i,h+1}(a_i' \mid w_i') \cdot \left(\hat{Q}^{m,\pi}_{i,h+1} - \overline{Q}^{m,\pi}_{i,h+1}\right)(w_i', a_i') \\
&+ \sum_{w_i', a_i'} \left(\hat{\mathbb{P}}_{i,h} - \mathbb{P}_{i,h}\right)(w_i' \mid w_i, a_i) \cdot \pi_{i,h+1}(a_i' \mid w_i') \cdot \hat{Q}^{m,\pi}_{i,h+1}(w_i', a_i') \\
\overset{(a)}{\leq}{}& \epsilon_r + \left\|\hat{Q}^{m,\pi}_{i,h+1} - \overline{Q}^{m,\pi}_{i,h+1}\right\|_\infty + \epsilon_{\mathbb{P}} H \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i| \\
\overset{(b)}{\leq}{}& \epsilon_r + (\epsilon_r(H-h) + \epsilon_{\mathbb{P}} H(H-h) \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i|) + \epsilon_{\mathbb{P}} H \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i| \\
\leq{}& \epsilon_r(H+1-h) + \epsilon_{\mathbb{P}} H(H+1-h) \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i|
\end{aligned}$$

For (a) we use the fact that for most m-step windows, the transition probability error is 0. Namely, for all $w_i' \in \mathcal{H}_i^{\leq m}$ for which the first $m-1$ actions and observation do not coincide with the last $m-1$ actions and observations of w_i, we have $\hat{\mathbb{P}}_{i,h}(w_i' \mid w_i, a_i) = \mathbb{P}_{i,h}(w_i' \mid w_i, a_i) = 0$. This observation is crucial, as it saves us an factor exponential in m one would obtain from summing over the entire space $\mathcal{H}_i^{\leq m}$. In (b) we apply the induction hypothesis. The final bound follows after plugging in ϵ_r and $\epsilon_{\mathbb{P}}$. □

Based on the marginalized Q-function, we define the advantage functions in $\mathcal{G}^m$ and in the Markov game underlying $\mathcal{P}$, respectively. For any $i \in \mathcal{N}$, $h \in [H]$, $\pi \in \Pi^m$, $w_i \in \mathcal{H}_i^{\leq m}$, $s_i \in \mathcal{S}_i$, and $a_i \in \mathcal{A}_i$, let

$$A^{m,\pi}_{i,h}(w_i, a_i) \coloneqq \overline{Q}^{m,\pi}_{i,h}(w_i, a_i) - \sum_{a_i' \in \mathcal{A}_i} \pi_{i,h}(a_i' \mid w_i) \overline{Q}^{m,\pi}_{i,h}(w_i, a_i'),$$
$$A^{\pi}_{i,h}(s_i, a_i) \coloneqq \overline{Q}^{\pi}_{i,h}(s_i, a_i) - \sum_{a_i' \in \mathcal{A}_i} \pi_{i,h}(a_i' \mid s_i) \overline{Q}^{\pi}_{i,h}(s_i, a_i').$$

Note for later use that the gain $\max_{a_i} A^{m,\pi}_{i,h}(w_i, a_i)$ is nonnegative for every $\pi \in \Pi^m$, $i \in \mathcal{N}$, $h \in [H]$, and $w_i \in \mathcal{H}_i^{\leq m}$, since $\sum_{a_i} \pi_{i,h}(a_i \mid w_i) A^{m,\pi}_{i,h}(w_i, a_i) = 0$ by definition.

The following lemma is the analogue of the gradient domination bound of Lemma 8 of [34] for the superstate game $\mathcal{G}^m$, transferred to the POMG $\mathcal{P}$ via the finite-window approximation of Lemma 1. It is applied in the proof of Lemma 10 below (see step (d)).

Lemma 9 (Approximate gradient domination)**.** *Let Assumption 2 hold, let $\pi \in \Pi^m$ and $i \in \mathcal{N}$, and let $\pi_i^\star \in \arg\max_{\pi_i' \in \Pi_i^m} V_i(\pi_i', \pi_{-i})$. Then*

$$\max_{\pi_i' \in \Pi_i^m} V_i(\pi_i', \pi_{-i}) - V_i(\pi) \leq \sum_{h=1}^{H} \sum_{w_i \in \mathcal{H}_i^{\leq m}} d^{m,\pi_i^\star}_{i,h}(w_i) \max_{a_i} A^{m,\pi}_{i,h}(w_i, a_i) + 2\epsilon^m_\rho. \tag{10}$$

Proof. We first establish the corresponding bound within the superstate game $\mathcal{G}^m$. Since state transitions are decoupled, neither the kernel $\mathbb{P}_i^m$ nor the marginalized reward $r_{i,h}^{m,\pi_{-i}}$ of the MDP $\mathcal{M}^{\pi_{-i}}$ depends on player i's policy, so that $V_i^m(\pi_i', \pi_{-i})$ and $V_i^m(\pi)$ are the values of π_i' and π_i in the same MDP $\mathcal{M}^{\pi_{-i}}$, whose Q-function under π is $\overline{Q}_{i,h}^{m,\pi}$. Applying the performance difference lemma in $\mathcal{M}^{\pi_{-i}}$ therefore gives, for any $\pi_i' \in \Pi_i^m$,

$$\begin{aligned} V_i^m(\pi_i', \pi_{-i}) - V_i^m(\pi) &= \sum_{h=1}^{H} \sum_{w_i \in \mathcal{H}_i^{\leq m}} d_{i,h}^{m,\pi_i'}(w_i) \sum_{a_i \in \mathcal{A}_i} \pi_{i,h}'(a_i \mid w_i)\, A_{i,h}^{m,\pi}(w_i, a_i) \\ &\leq \sum_{h=1}^{H} \sum_{w_i \in \mathcal{H}_i^{\leq m}} d_{i,h}^{m,\pi_i'}(w_i)\, \max_{a_i} A_{i,h}^{m,\pi}(w_i, a_i), \end{aligned} \tag{11}$$

since $\pi_{i,h}'(\cdot \mid w_i)$ is a probability distribution over $\mathcal{A}_i$, and where we used that $d_{i,h}^{m,(\pi_i', \pi_{-i})} = d_{i,h}^{m,\pi_i'}$, as $d_{i,h}^{m,\pi}$ depends on π only through π_i by decoupledness.

It remains to transfer this bound from $\mathcal{G}^m$ to the POMG $\mathcal{P}$. Both $(\pi_i^\star, \pi_{-i})$ and π lie in Π^m, so applying the first bound of Lemma 1 to each of them yields

$$V_i(\pi_i^\star, \pi_{-i}) - V_i(\pi) \;\leq\; V_i^m(\pi_i^\star, \pi_{-i}) - V_i^m(\pi) + 2\epsilon_\rho^m.$$

Bounding the right-hand side by (11) at $\pi_i' = \pi_i^\star$, and using $V_i(\pi_i^\star, \pi_{-i}) = \max_{\pi_i' \in \Pi_i^m} V_i(\pi_i', \pi_{-i})$ gives (10). □

Lemma 10. *Suppose Assumptions 1, 2, and 4 hold, and that for some $0 < \epsilon \leq 1$ and for all $i \in \mathcal{N}$, $h \in [H]$, and $k \in [K]$, our Q-function estimates satisfies*

$$\left| \hat{Q}_{i,h}^{m,\pi^{(k)}}(w_i, a_i) - Q_{i,h}^{m,\pi^{(k)}}(w_i, a_i) \right| \lesssim H^3 N \cdot |\mathcal{A}_i| \cdot |\mathcal{O}_i| \cdot \left(\epsilon N + (1-\rho)^{m-1}\right) \coloneqq \epsilon_Q.$$

Then, choosing stepsize $\eta^{(k)} = \frac{1}{\sqrt{4N^2H^3k}}$, there exists $k \in [K]$ such that for all $i \in \mathcal{N}$,

$$\max_{\pi_i' \in \Pi_i^m} V(\pi_i', \pi_{-i}^{(k)}) - V(\pi^{(k)}) \lesssim \kappa \left(\frac{\sqrt{N^2H^3}\,(HN + \log(K))}{\sqrt{K}} + NH\epsilon_Q + N^2 H^{5/2} \epsilon_\rho^m \sqrt{K} \right). \tag{12}$$

Proof. Fix some $k \in [K]$ and define for each $i \in \mathcal{N}$ the joint policy

$$\widetilde{\pi}^i := (\pi_1^{(k)}, \ldots, \pi_{i-1}^{(k)}, \pi_i^{(k+1)}, \ldots, \pi_N^{(k+1)}) \in \Pi^m.$$

Then, since by Proposition 2, the superstate game $\mathcal{G}^m$ is a Markov $2\epsilon_\rho^m$-potential game, we can decompose

$$\Psi(\pi^{(k+1)}) - \Psi(\pi^{(k)}) \tag{13}$$

$$= \sum_{i=1}^{N} \Psi(\widetilde{\pi}^i) - \Psi(\widetilde{\pi}^{i+1}) \tag{14}$$

$$\geq -2N\epsilon_\rho^m + \sum_{i=1}^{N} V_i^m(\widetilde{\pi}^i) - V_i^m(\widetilde{\pi}^{i+1}) \tag{15}$$

$$\overset{(a)}{=} -2N\epsilon_\rho^m + \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{w_i, a_i} d_{i,h}^{m,\widetilde{\pi}^i}(w_i) \left(\pi_i^{(k+1)}(a_i \mid w_i) - \pi_i^{(k)}(a_i \mid w_i)\right) Q_i^{\widetilde{\pi}^{i+1}}(w_i, a_i) \tag{16}$$

$$\overset{(b)}{\geq} \eta^{(k)} \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{w_i} d_{i,h}^{m,\widetilde{\pi}^i}(w_i)\, \max_{a_i} A_{i,h}^{m,\pi^{(k)}}(w_i, a_i) - 2\eta^{(k)} NH\epsilon_Q - 4N^2H^3(\eta^{(k)})^2 - 2N\epsilon_\rho^m \tag{17}$$

$$\overset{(c)}{\geq} \frac{\eta^{(k)}}{\kappa^{(k)}} \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{w_i} d_{i,h}^{m,\pi_i^{\star,k}}(w_i) \max_{a_i} A_{i,h}^{m,\pi^{(k)}}(w_i, a_i) - 2\eta^{(k)} NH\epsilon_Q - 4N^2H^3(\eta^{(k)})^2 - 2N\epsilon_\rho^m \tag{18}$$

$$\overset{(d)}{\geq} \frac{\eta^{(k)}}{\kappa^{(k)}} \sum_{i=1}^{N} \left(\max_{\pi_i' \in \Pi_i^m} V(\pi_i', \pi_{-i}^{(k)}) - V(\pi^{(k)}) \right) - 2\eta^{(k)} NH\epsilon_Q - 4N^2H^3(\eta^{(k)})^2 - 4NH\epsilon_\rho^m \tag{19}$$

where (a) is by the performance difference lemma (see Lemma 6 of [34]), and (b) follows similar as in [34]. Step (c) compares the superstate visitation of the current iterate with that of the best response. This is where the coefficient $\kappa^{(k)}$ of Assumption 4 enters: Since the i-th component of $\widetilde{\pi}^i$ is $\pi_i^{(k+1)}$ and $d_{i,h}^{m,\pi}$ depends on π only through π_i, we have $d_{i,h}^{m,\widetilde{\pi}^i} = d_{i,h}^{m,\pi_i^{(k+1)}}$, and the definition of $\kappa^{(k)}$ gives $d_{i,h}^{m,\pi_i^{\star,k}}(w_i) \leq \kappa^{(k)} d_{i,h}^{m,\widetilde{\pi}^i}(w_i)$ for every w_i. As the gains $\max_{a_i} A_{i,h}^{m,\pi^{(k)}}(w_i, a_i)$ are nonnegative, this yields, for every $i \in \mathcal{N}$ and $h \in [H]$,

$$\sum_{w_i \in \mathcal{H}_i^{\leq m}} d_{i,h}^{m,\widetilde{\pi}^i}(w_i) \max_{a_i} A_{i,h}^{m,\pi^{(k)}}(w_i, a_i) \;\geq\; \frac{1}{\kappa^{(k)}} \sum_{w_i \in \mathcal{H}_i^{\leq m}} d_{i,h}^{m,\pi_i^{\star,k}}(w_i) \max_{a_i} A_{i,h}^{m,\pi^{(k)}}(w_i, a_i),$$

and summing over $h \in [H]$ gives step (c).

Step (d) follows from Lemma 9, applied at $\pi = \pi^{(k)}$ with $\pi_i^\star = \pi_i^{\star,k}$ the best response of Assumption 4. Multiplying (10) by $\eta^{(k)}/\kappa^{(k)} \leq 1$, summing over $i \in \mathcal{N}$, and using $H \geq 1$ shows that the additional error is absorbed by replacing $-2N\epsilon_\rho^m$ with $-4NH\epsilon_\rho^m$ in the last line of (13).

The final bound (12) follows from a telescoping argument applied to (13), the fact that $\Psi(\pi) \leq HN$ for all $\pi \in \Pi^m$, and our choice of stepsize. Concretely, writing $G^{(k)} := \sum_{i \in \mathcal{N}} \left(\max_{\pi_i' \in \Pi_i^m} V(\pi_i', \pi_{-i}^{(k)}) - V(\pi^{(k)}) \right) \geq 0$ and summing (13) over $k \in [K]$ gives

$$\left(\min_{k \in [K]} G^{(k)} \right) \sum_{k=1}^{K} \frac{\eta^{(k)}}{\kappa^{(k)}} \;\leq\; HN + 2NH\epsilon_Q \sum_{k=1}^{K} \eta^{(k)} + 4N^2H^3 \sum_{k=1}^{K} \left(\eta^{(k)}\right)^2 + 4NH\epsilon_\rho^m K.$$

Dividing by $\sum_{k=1}^K \eta^{(k)}/\kappa^{(k)} \geq \kappa^{-1} \sum_{k=1}^K \eta^{(k)}$, which holds by Assumption 4, together with $\sum_{k=1}^K \eta^{(k)} \gtrsim \sqrt{K/(N^2H^3)}$, and using $\sum_{k=1}^K (\eta^{(k)})^2 \lesssim \log(K)/(N^2H^3)$ then yields (12). □

We are now ready to prove the main theorem.

Proof of Theorem 1. If the minimum in the requirement on m is attained by H, then $m = H$ and, as noted in Section 3, $\mathcal{G}^H$ is equivalent to $\mathcal{P}$, so that the finite-window approximation is exact and every term involving ϵ_ρ^m below vanishes. We may therefore assume throughout that $m \geq c_1 \cdot \rho^{-1} \log\left(\frac{\kappa NHAO}{\beta\epsilon}\right)$. By Lemma 10, there exists $k \in [K]$ such that for all $i \in \mathcal{N}$,

$$\max_{\pi_i' \in \Pi_i^m} V(\pi_i', \pi_{-i}^{(k)}) - V(\pi^{(k)}) \lesssim \underbrace{\frac{\kappa\sqrt{N^2H^3}\,(HN + \log(K))}{\sqrt{K}}}_{(a)} + \underbrace{\kappa NH\epsilon_Q}_{(b)} + \underbrace{\kappa N^2 H^{5/2} \epsilon_\rho^m \sqrt{K}}_{(c)}.$$

We next show that for appropriately chosen iteration number K, trajectory length T, and window size m, the terms (a), (b), and (c) are all upper bounded by $\epsilon/3$.

With our choices of $\zeta = \epsilon$, $T \gtrsim \frac{\kappa^2 A^{2m+2} O^2 N^2 H^6}{\beta^{2m} \epsilon^{2m+2}} \log(1/\delta)$ and $m \gtrsim \rho^{-1} \log\left(\frac{\kappa HAON}{\beta\epsilon}\right)$, applying Lemma 2 and Lemma 8, yields that with probability at least $1 - \delta$, we have $\epsilon_Q \lesssim \frac{\epsilon}{\kappa NH}$ and hence $(b) \lesssim \epsilon/3$.

Moreover, by choosing $K \gtrsim \frac{\kappa^2 N^4 H^5 A^2}{\epsilon^2}$, we ensure that both $(a) \lesssim \epsilon/3$, as well as $(c) \lesssim \epsilon/3$. Note that, in contrast to a uniform lower bound on the superstate visitation, κ does not depend on m, so that the requirement on m imposed by (c), namely $\epsilon_\rho^m \lesssim \epsilon^2/(\kappa^2 N^4 H^5 A)$, is met by the logarithmic choice of m above for any $\rho > 0$.

The overall sample complexity of Algorithm 1 is given by

$$TK = \mathcal{O}\left(\frac{\kappa^4 A^2 O^2 N^6 H^{11}}{\epsilon^2}\left(\frac{A}{\beta\epsilon}\right)^{2m}\log(1/\delta)\right).$$

Moreover, we have

$$\left(\frac{A}{\beta\epsilon}\right)^{2m} = \exp\left(\frac{2}{\rho}\log\frac{A}{\beta\epsilon}\cdot\log\left(\frac{\kappa HAON}{\beta\epsilon}\right)\right) = \left(\frac{\kappa HAON}{\beta\epsilon}\right)^{\frac{2}{\rho}\log(A/(\beta\epsilon))},$$

from which we can conclude the bound

$$TK = \mathcal{O}\left(\frac{\kappa^4 A^2 O^2 N^6 H^{11}}{\epsilon^2}\left(\frac{\kappa HAON}{\beta\epsilon}\right)^{\frac{2}{\rho}\log(A/(\beta\epsilon))}\log(1/\delta)\right).$$

□

F Numerical experiments

In this section we report simulation results for Algorithm 1 in two decoupled POMG environments related to the applications discussed in Appendix B.[2] The first is a dynamic transmission game over a shared channel, in which the action and observation spaces are deliberately small so that exact Nash gaps against full-history deviations can be computed. The second is a cooperative gridworld navigation task with a congestion penalty, in which exact Nash gaps are out of reach and we instead report the exact expected episode return. In both environments we run Algorithm 1 for different window lengths m.

F.1 Dynamic transmission game

Setting. We consider a minimal instance of the wireless network example of Appendix B: two users share a communication channel and each maintains a local packet queue, with rewards coupled through the congestion of the shared channel. For $N = 2$ players over horizon $H = 6$, the POMG is as follows.

- *States.* $s_i \in \mathcal{S}_i = \{0, 1\}$, indicating whether player i's queue is empty or holds a packet. The initial distribution μ_i is uniform on $\mathcal{S}_i$.
- *Actions.* $a_i \in \mathcal{A}_i = \{\mathsf{wait}, \mathsf{transmit}\}$.
- *Rewards.* Writing T for transmit, we set

$$r_{i,h}(s, a) \coloneqq c_0 + c_1 \mathbf{1}\{s_i = 1,\, a_i = \mathsf{T}\} - c_2 \mathbf{1}\{a_i = \mathsf{T}\} - c_3 \mathbf{1}\{a_1 = a_2 = \mathsf{T}\}. \tag{20}$$

  This models a benefit c_1 for transmitting a queued packet, a fixed transmission cost c_2, and an interference penalty c_3 incurred when both users transmit at once. We use $c_0 = 0.5$, $c_1 = 0.3$, $c_2 = 0.15$, and $c_3 = 0.35$, so that $r_{i,h} \in [0, 1]$.
- *Transitions.* Player i's queue evolves independently of the other players: an empty queue receives a packet with probability λ; a held packet is delivered successfully with probability q if the player transmits, and otherwise expires with probability p_{e} while the player waits. We use $\lambda = 0.125$, $q = 0.8$, and $p_{\mathrm{e}} = 0.06$, so that packets are long-lived and the local state stays correlated over several steps.
- *Observations.* Player i observes its true queue state, $o_i = s_i$, with probability $p_{\mathrm{o}} = 0.1$, and an observation drawn uniformly from $\mathcal{O}_i = \{0, 1\}$ otherwise.

Assumptions. Transitions and observations are decoupled by construction, Assumption 3 holds with $\beta = (1 - p_{\mathrm{o}})/2 = 0.45$, and Assumption 1 is verified below. For filter stability (Assumption 2), we numerically maximize the one-step contraction ratio of the belief update over pairs of beliefs and all action-observation pairs, which is tractable since $|\mathcal{S}_i| = 2$. The maximum, $1 - \rho \approx 0.996$, coincides with the product of the largest observation likelihood ratio $(1+p_{\mathrm{o}})/(1-p_{\mathrm{o}})$ and the largest Dobrushin coefficient of the transition kernel, $\max\{|1-\lambda-p_{\mathrm{e}}|,\, |\lambda-(1-q)|\}$. Hence Assumption 2 holds with $\rho \approx 0.004$.

Potential structure. Each player's reward consists of a term depending only on its own state and action, $c_1 \mathbf{1}\{s_i = 1,\, a_i = \mathsf{T}\} - c_2 \mathbf{1}\{a_i = \mathsf{T}\}$, and the interference cost $c_3 \mathbf{1}\{a_1 = a_2 = \mathsf{T}\}$, which is shared by both players. Adding up the players' own terms and counting the shared cost once gives the potential

$$\phi_h(s, a) = c_1\big(\mathbf{1}\{s_1 = 1,\, a_1 = \mathsf{T}\} + \mathbf{1}\{s_2 = 1,\, a_2 = \mathsf{T}\}\big) - c_2\big(\mathbf{1}\{a_1 = \mathsf{T}\} + \mathbf{1}\{a_2 = \mathsf{T}\}\big) - c_3 \mathbf{1}\{a_1 = a_2 = \mathsf{T}\}.$$

Indeed, writing j for the player other than i, a direct computation from (20) gives

$$r_{i,h}(s, a) - \phi_h(s, a) = c_0 - c_1 \mathbf{1}\{s_j = 1,\, a_j = \mathsf{T}\} + c_2 \mathbf{1}\{a_j = \mathsf{T}\},$$

which depends neither on s_i nor on a_i. This is exactly the stage potential condition of Assumption 1. Note that the game is not of identical interest, as players disagree on who should be the one to transmit.

[2] The code for all experiments is available at `https://github.com/philip-jordan/NE-POMG`.

Evaluation. We report the exact Nash gap of the iterates $\pi^{(k)}$ against *full-history* deviations,

$$\max_{i\in\mathcal{N}}\Big[\max_{\pi_i'\in\Pi_i^H} V_i(\pi_i',\pi_{-i}^{(k)}) - V_i(\pi^{(k)})\Big],$$

which is the quantity our learning objective in Section 4 controls. This is feasible here because, by decoupledness, fixing π_{-i} leaves player i facing a POMDP over its own two-state chain in which the other player enters only through the mean rewards $r_{i,h}^{H,\pi_{-i}}$, and the resulting best-response value can be computed by backward induction with exact belief propagation over player i's history tree, which has $(|\mathcal{A}_i|\cdot|\mathcal{O}_i|)^H = 4^6 = 4096$ leaves.

Experiment. We run Algorithm 1 for window lengths $m \in \{1,\dots,6\}$ with $K = 500$ iterations, $T = 10^5$ episodes per iteration, exploration $\zeta = 0.02$, and stepsizes $\eta^{(k)} = \min\{1, \frac{1}{2}(k+1)^{-1/2}\}$, over 8 independent runs. All runs start from the uniform policy and differ only in the random seed used for sampling episodes, i.e., in the realized transitions, observations, and actions.

Choice of parameters. We set the parameters by hand, using a few trial runs. The stepsizes follow the $k^{-1/2}$ decay of Theorem 1. We chose K large enough for all curves to flatten, and T and ζ such that the statistical error and the exploration perturbation are small compared to the finite-window approximation error, so that different window lengths can be distinguished. These values are far smaller than those prescribed by Theorem 1, whose bounds are worst-case.

Results. Figure 1 shows the Nash gap over the first 100 iterations, by which point all curves have flattened; the gaps at $k = 500$ agree with those at $k = 100$ to within the plotted band. For every window length, the gap drops by one to two orders of magnitude within roughly 20 iterations and then settles at an m-dependent floor which decreases monotonically in m. This is the behavior predicted by Proposition 1 and Theorem 1: larger windows shrink the finite-window approximation error and therefore the accuracy at which independent learning saturates. The curves for $m = 5$ and $m = 6$ coincide, since with $H = 6$ a window of five action-observation pairs already covers the entire history available to a player, so that no further increase of m can help.

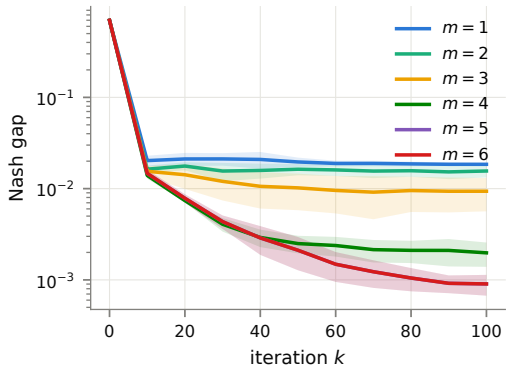
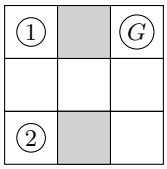


Figure 1: Transmission game: exact Nash gap against full-history deviations over the iterations of Algorithm 1, for window lengths $m \in \{1,\dots,6\}$ ($N = 2$, $H = 6$, $T = 10^5$). Curves show the mean over 8 independent runs and shaded areas one standard deviation. The $m = 5$ curve is hidden underneath the $m = 6$ curve.

F.2 Cooperative gridworld

Setting. Two agents, starting in cells (1) and (2), navigate a 3×3 grid with two walls (shaded) towards a common goal cell (G), while congestion is penalized. This instantiates the multi-robot example of Appendix B: transitions are decoupled, so agents may occupy the same cell, but sharing a cell is penalized. Since the only route between the two halves of the grid is the narrow passage in the center cell, the agents have to coordinate their crossings.



Over horizon $H = 10$, the POMG is as follows.

- *States.* $s_i \in \mathcal{S}_i$ is player i's cell, one of the 7 non-wall cells. Player 1 starts in the top-left cell and player 2 in the bottom-left cell, so μ_i is a point mass. The goal cell G is absorbing.
- *Actions.* $\mathcal{A}_i = \{\mathsf{up}, \mathsf{down}, \mathsf{left}, \mathsf{right}, \mathsf{stay}\}$. A move into a wall or off the grid leaves the agent in place.
- *Rewards.* A player receives $+10$ at the step it first reaches G and 0 at every step thereafter. Before reaching G, it receives -5 if it shares its cell with the other player and -1 otherwise. For readability we state these rewards unnormalized; rescaling them into $[0, 1]$ by a common affine map, as assumed in Section 2, leaves the potential structure and the set of (approximate) Nash equilibria unchanged up to the corresponding rescaling of ϵ.
- *Transitions.* Each player moves according to its own action, but with probability $\mathsf{slip} = 0.1$ the executed action is replaced by one drawn uniformly from $\mathcal{A}_i$.
- *Observations.* Player i observes its true cell with probability $p_{\mathrm{o}} = 0.5$, and a uniformly random non-wall cell otherwise, so that $|\mathcal{O}_i| = 7$.

Assumptions. Transitions and observations are decoupled by construction, Assumption 3 holds with $\beta = (1 - p_{\mathrm{o}})/7 = 1/14$, and Assumption 1 is verified below. Filter stability (Assumption 2), however, does not hold. The reason is the absorbing goal combined with informative observations: a belief concentrated on G remains there forever, whereas observations of cells far from G keep driving any other belief away from G. This experiment evaluates Algorithm 1 beyond the scope of the approximation guarantees provided by our theory.

Potential structure. Write $g_{i,h}$ for the part of player i's reward that depends only on its own trajectory: the goal bonus $+10$ at the step it first reaches G, and the step cost -1 at every step before. The players are coupled only through the additional co-location penalty $-4 \cdot \mathbf{1}\{s_1 = s_2 \neq G\}$, which both players incur, so that $r_{i,h} = g_{i,h} - 4 \cdot \mathbf{1}\{s_1 = s_2 \neq G\}$. As in Appendix F.1, Assumption 1 then holds with potential

$$\phi_h = g_{1,h} + g_{2,h} - 4 \cdot \mathbf{1}\{s_1 = s_2 \neq G\},$$

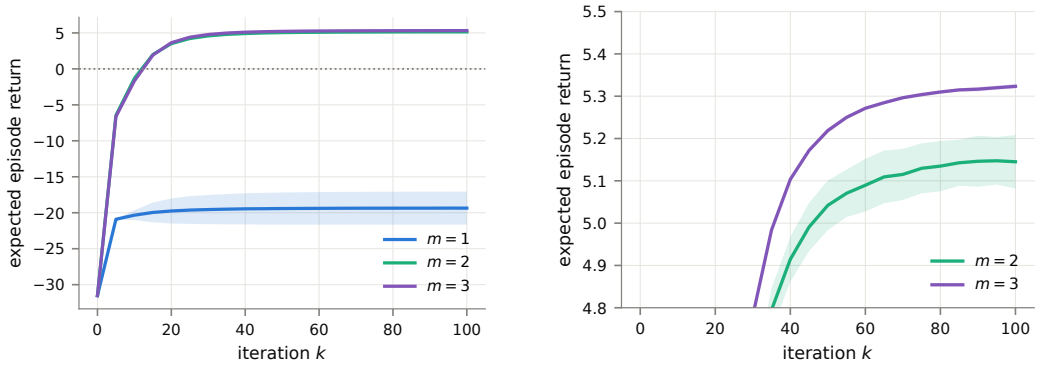
since $r_{i,h} - \phi_h = -g_{j,h}$ for $j \neq i$ does not depend on player i's state or action. The game is not of identical interest: at the central passage each player prefers the other to yield.

Evaluation. Here, computing the exact Nash gap is out of reach, since determining a full-history best response would require a dynamic program over $(|\mathcal{A}_i| \cdot |\mathcal{O}_i|)^H = 35^{10} \approx 2.8 \cdot 10^{15}$ histories. We therefore report the expected episode return of the iterate $\pi^{(k)}$ (without exploration), summed over both players, where the expectation is over the slip in the transitions, the observation noise, and the randomness of the policies.

Experiment. We run Algorithm 1 for window lengths $m \in \{1, 2, 3\}$ with $K = 100$ iterations, $T = 10^6$ episodes per iteration, exploration $\zeta = 0.1$, and stepsizes $\eta^{(k)} = \min\{1, \frac{1}{2}(k + 1)^{-1/2}\}$, over 8 independent runs, which, as in Appendix F.1, differ only in the random seed. As before, the parameters were set by hand: the stepsizes are as in Appendix F.1, ζ, K and T were chosen after a few trial runs. All runs complete within about an hour on a consumer-grade CPU when the independent runs are executed in parallel.

Results. Figure 2 shows the expected episode return over the iterations of Algorithm 1. In the left panel, the shortest window $m = 1$ is insufficient for learning policies that reach the goal: the return improves over the initial uniform policy but stays deeply negative, around -19. For $m = 2$ and $m = 3$, the return instead climbs to a positive value of about $+5$, i.e., the players learn to reach the goal while avoiding congestion in the narrow central passage. The right panel zooms in on these two

window lengths and shows that the longer window $m = 3$ converges to a higher return, about 5.32, than $m = 2$, about 5.15, and does so with a visibly smaller spread across runs.



(a) All window lengths $m \in \{1, 2, 3\}$.

(b) Zoom on $m \in \{2, 3\}$.

Figure 2: Gridworld: exact expected episode return, summed over both players, over the iterations of Algorithm 1 ($N = 2$, $H = 10$, $T = 10^6$). Curves show the mean over 8 independent runs and shaded areas one standard deviation. Left: $m = 1$ fails to reach the goal, while $m \geq 2$ does. Right: among the successful window lengths, the longer history $m = 3$ attains the higher return.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We refer to the **Contributions** paragraph in Section 1. Our main contribution is Theorem 1. Additional insights leading up to this result are stated in Propositions 1 and 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 1 we state and motivate the restrictions of our POMG setting to the potential game case with decoupled dynamics (see also Appendix B). We further include a discussion of our result after Theorem 1, and discuss limitations in the context of future directions in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All our assumptions (see Assumptions 1, 2, 3, 4) and our main results (see Theorem 1, Proposition 1, and 2) are stated in the main part of the paper. We also give proof overviews but defer full proofs to the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our main contributions are theoretical. For the numerical experiments in Appendix F, we fully specify both environments (states, actions, rewards, transitions, observations, and all their parameters), the evaluation metrics, and all parameters of Algorithm 1 (m, K, T, ζ, $\eta^{(k)}$, number of runs). In addition, the code is publicly available.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for all experiments is available at `https://github.com/philip-jordan/NE-POMG` (see the footnote in Appendix F), including scripts and instructions to reproduce the figures. No external data is used.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix F states all parameters of Algorithm 1 for both experiments and explains how they were chosen (see the paragraph *Choice of parameters* and the *Experiment* paragraphs).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All curves in Figures 1 and 2 show the mean over 8 independent runs together with one standard deviation as a shaded band, as stated in the captions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F reports the runtime of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is of theoretical nature, complemented by small-scale simulations, so there are no direct consequences that could violate the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is of theoretical nature, so there is no direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper is of theoretical nature, and hence poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the simulation code, which is released at `https://github.com/philip-jordan/NE-POMG` together with a README documenting the environments, the implementation of Algorithm 1, and how to reproduce the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.