Multi-Player Zero-Sum Markov Games with Networked Separable Interactions

Chanwoo Park\textsuperscript{a}  
MIT  
cpark97@mit.edu

Kaiqing Zhang\textsuperscript{a}  
University of Maryland, College Park  
kaiqing@umd.edu

Asuman Ozdaglar  
MIT  
asuman@mit.edu

Abstract

We study a new class of Markov games, (multi-player) zero-sum Markov Games with Networked separable interactions (zero-sum NMGs), to model the local interaction structure in non-cooperative multi-agent sequential decision-making. We define a zero-sum NMG as a model where the payoffs of the auxiliary games associated with each state are zero-sum and have some separable (i.e., polymatrix) structure across the neighbors over some interaction network. We first identify the necessary and sufficient conditions under which an MG can be presented as a zero-sum NMG, and show that the set of Markov coarse correlated equilibrium (CCE) collapses to the set of Markov Nash equilibrium (NE) in these games, in that the product of per-state marginalization of the former for all players yields the latter. Furthermore, we show that finding approximate Markov stationary CCE in infinite-horizon discounted zero-sum NMGs is PPAD-hard, unless the underlying network has a “star topology”. Then, we propose fictitious-play-type dynamics, the classical learning dynamics in normal-form games, for zero-sum NMGs, and establish convergence guarantees to Markov stationary NE under a star-shaped network structure. Finally, in light of the hardness result, we focus on computing a Markov non-stationary NE and provide finite-iteration guarantees for a series of value-iteration-based algorithms. We also provide numerical experiments to corroborate our theoretical results.

1 Introduction

Nash Equilibrium (NE) has been broadly used as a solution concept in game theory, since the seminal works of [1, 2]. Perhaps equally important, NE is also deeply rooted in the prediction and analysis of learning dynamics in multi-agent strategic environments: it may appear as a natural outcome of many non-equilibrating learning processes of multiple agents interacting with each other [3, 4]. A prominent example of such learning processes is fictitious play (FP) [5, 6], in which myopic agents estimate the opponents’ play using history, and then choose a best-response action (based on their payoff matrix) against this estimate, as if the opponents use it as their stationary strategy. The focus of these studies has initially been on the convergence to NE in zero-sum games (see [5, 6], and also [7, 4]), and in games with aligned objective (identical-interest and potential games, see [8]). Since then, FP has been shown to converge to NE in more important classes of games, including 2\times n games [9, 10], “one-against-all” games [11], and zero-sum polymatrix games [12], justifying the prediction power of NE in learning in normal-form/matrix games.

Some of these results have recently been extended to the stochastic game (also known as the Markov game (MG)) setting, a model for multi-agent sequential decision-making with state transition dynamics, first introduced in [13]. In particular, [14, 15, 16, 17, 18] have studied best-response-type learning dynamics in two-player zero-sum MGs, and [19, 17] have studied that in multi-player

\textsuperscript{a}Alphabetical order.

identical-interest games. Following the same path as studying matrix games, one natural question arises: Are there other classes of MGs beyond two-player zero-sum and identical-interest cases that allow natural learning dynamics, e.g., fictitious play, to justify NE as the long-run emerging outcome?

On the other end of the spectrum, it is well-known that for general-sum normal-form games, the special case of MGs without the state transition dynamics, computing an NE is intractable [20][21]. Relaxed solution concepts as (coarse) correlated equilibrium ((C)CE) have thus been favored when it comes to equilibrium computation for general-sum, multi-player games [22][3]. Encouragingly, when the interactions among players have some networked separable structure, also known as being polymatrix, computing NE may be made tractable even in multi-player settings. This has been instantiated in the seminal works [23][24] in the normal-form game setting, which showed that any CCE collapses to the NE in such games when the payoffs are zero-sum. Thus, any algorithms that can efficiently compute the CCE in such games will lead to the efficient computation of the NE.

In fact, besides being of theoretical interest, multi-player zero-sum games with networked separable interactions also find a range of applications, including security games [24], fashion games [25][26][27], and resource allocation problems [28]. These examples, oftentimes, naturally involve some state transition that captures the dynamics of the evolution of the environment in practice. For example, in the security game, the protection level or immunity of a target increases as a function of the number of past attacks, leading to a smaller probability of an attack on the target being successful. Hence, it is imperative to study such multi-player zero-sum games with state transitions. As eluded in the recent results [29][30], such a transition from stateless to stateful cases may not always yield straightforward and expected results, e.g., computing stationary CCE can be computationally intractable in stochastic games, in stark contrast to the normal-form case where CCE can be efficiently computed. This naturally prompts another question: Are there other types of (multi-player) MGs that may circumvent the computational hardness of computing NE/CCE?

In an effort to address these two questions, we introduce a new class of Markov games – (multi-player) zero-sum Markov games with networked separable interactions (zero-sum NMGs). We summarize our contributions as follows, and defer a more detailed literature review to Appendix [A].

Contributions. First, we introduce a new class of non-cooperative Markov games: (multi-player) zero-sum MGs with Networked separable interactions (zero-sum NMGs), wherein the payoffs of the auxiliary-games associated with each state, i.e., the sum of instantaneous reward and expectation of any estimated state-value functions, possess the multi-player zero-sum and networked separable (i.e., polymatrix) structure as in [31][28][23][24] for normal-form games, a strict generalization of the latter. We also provide structural results on the reward and transition dynamics of the game, as well as examples of this class of games. Specifically, for a Markov game to qualify as a zero-sum NMG, if and only if its reward function has the zero-sum polymatrix structure, and its transition dynamics is an ensemble of multiple single-controller transition dynamics that are sampled randomly at each state (see Remark [5] for more details). This transition dynamics covers the common ones in the MG literature, including the single-controller and turn-based dynamics. Second, we show that Markov CCE and Markov NE collapse in that the product of per-state marginal distributions of the former yields the latter, making it sufficient to focus on the former in equilibrium computation. We then show the PPAD-hardness [32] of computing the Markov stationary equilibrium, a natural solution concept in infinite-horizon discounted MGs, unless the underlying network has a star-topology. This is in contrast to the normal-form case where CCE is always computationally tractable. Third, we study the fictitious-play property [8] of zero-sum NMGs, showing that the fictitious-play dynamics [16][19] converges to the Markov stationary NE, for zero-sum NMGs with a star-shaped network structure. Finally, in light of the hardness of computing stationary equilibria, we develop a series of value-iteration-based algorithms for computing a Markov non-stationary NE of zero-sum NMGs, with finite-iteration guarantees. We also provide numerical experiments to corroborate our theoretical results in Section [C]. We hope our results serve as a starting point for studying this networked separable interaction structure in non-cooperative Markov games.

Notation. For a real number $c$, we use $c \_+$ to denote $\max\{c, 0\}$. For an event $\mathcal{E}$, we use $1(\mathcal{E})$ to denote the indicator function such that $1(\mathcal{E}) = 1$ if $\mathcal{E}$ is true, and $1(\mathcal{E}) = 0$ otherwise. We define multinomial distribution with probability $(\pi_i)_{i \in \mathcal{N}}$ as Multinomial$(\pi_i)_{i \in \mathcal{N}}$. We denote the uniform distribution over a set $\mathcal{S}$ as Unif$(\mathcal{S})$. We denote the Bernoulli distribution with probability $p$ as Bern$(p)$. The sgn function is defined as $\text{sgn}(x) = 2 \times 1(x \geq 0) - 1$. The KL-divergence between two probability distributions $p, q$ is denoted as $\text{KL}(p, q) = E_p[\log(p/q)]$. For a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we denote the set of neighboring nodes of node $i \in \mathcal{N}$ as $\mathcal{E}_i$ (without including $i$). The maximum norm of a matrix $X \in \mathbb{R}^{m \times n}$, denoted as $\|X\|_{\max}$, is defined as $\|X\|_{\max} := \max_{i \in [m], j \in [n]} |X_{i,j}|$. 

2
2 Preliminaries

2.1 Markov games

We define a Markov game as a tuple \((N, S, A, H, (P_h)_{h \in [H]}, (r_{h,i})_{h \in [H], i \in N}, \gamma)\), where \(N = [n]\) is the set of players, \(S\) is the state space with \(|S| = S, A_i\) is the action space for player \(i\) with \(|A_i| = A_i\), and \(A = \prod_{i \in N} A_i\). \(H \leq \infty\) is the length of the horizon, \(P_h : S \times A \rightarrow \Delta(S)\) captures the state transition dynamics at timestep \(h\), \(r_{h,i} \in [0, R]\) is the reward function for player \(i\) at timestep \(h\), bounded by some \(R > 0\), and \(\gamma \in (0, 1]\) is a discount factor. An MG with a finite horizon \((H < \infty)\) is also referred to as an episodic MG, while an MG with an infinite horizon \((H = \infty)\) and \(\gamma < 1\) is referred to as an infinite-horizon \(\gamma\)-discounted MG. When \(H = \infty\), we will consider the transition dynamics and reward functions, denoted by \(P\) and \((r_{i})_{i \in N}\), respectively, to be independent of \(h\).

Hereafter, we may use agent and player interchangeably.

**Policy.** Consider the stochastic Markov policy for player \(i\), denoted by \(\pi_i\), as \(\pi_i := \{\pi_{h,i} : S \rightarrow \Delta(A_i)\}_{h \in [H]}\). A joint Markov policy is a policy \(\pi := \{\pi_h : S \rightarrow \Delta(A)\}_{h \in [H]}\), where \(\pi_h : S \rightarrow \Delta(A)\) decides the joint action of all players that can be potentially correlated. A joint Markov policy is a product Markov policy if \(\pi_h : S \rightarrow \prod_{i \in N} \Delta(A_i)\) for all \(h \in [H]\), and is denoted as \(\pi = \pi_1 \times \pi_2 \times \cdots \times \pi_n\). When the policy is independent of \(h\), the policy is called a stationary policy. We let \(a_h\) denote the joint action of all agents at timestep \(h\). Unless otherwise noted, we will work with Markov policies throughout. We denote \(\pi(s) \in \Delta(A)\) as the joint policy at state \(s \in S\).

**Value function.** For player \(i\), the value function under joint policy \(\pi\), at timestep \(h\) and state \(s_h\) is defined as \(V^\pi_{h,i}(s_h) := \mathbb{E}_\pi^h \left[ \sum_{h'=h}^H \gamma^{h-h'} r_{h',i}(s_{h'},a_{h'}) \mid s_h \right]\), which denotes the expected cumulative reward for player \(i\) at step \(h\) if all players adhere to policy \(\pi\). We also define \(V^\pi_{h,i}(\rho) := \mathbb{E}_{\rho \sim \pi}^h [V^\pi_{h,i}(s_h)]\) for some state distribution \(\rho \in \Delta(S)\). We denote the \(Q\)-function for the \(i\)-th player under policy \(\pi\), at step \(h\) and state \(s_h\) as \(Q^\pi_{h,i}(s_h,a_h) := \mathbb{E}_\pi^h \left[ \sum_{h'=h}^H \gamma^{h-h'} r_{h',i}(s_{h'},a_{h'}) \mid s_h, a_h \right]\), which determines the expected cumulative reward for the \(i\)-th player at step \(h\), when starting from the state-action pair \((s_h, a_h)\). For the infinite-horizon discounted setting, we also use \(V^\pi_{i}\) and \(Q^\pi_{i}\) to denote \(V^\pi_{i}\) and \(Q^\pi_{i}\) for short, respectively.

**Approximate equilibrium.** Define an \(\epsilon\)-approximate Markov perfect Nash equilibrium as a product policy \(\pi\), which satisfies \(\max_{i \in N} \max_{\mu_i \in \Delta(A_i)} (\pi_{h,i}(\mu_i) - V^\mu_{h,i}(\rho)) \leq \epsilon\) for all \(\rho \in \Delta(S)\) and \(h \in [H]\), where \(\pi_{h,i}\) represents the marginalized policy of all players except player \(i\). Define an \(\epsilon\)-approximate Markov coarse correlated equilibrium as a joint policy \(\pi\), which satisfies \(\max_{i \in N} \max_{\mu_i \in \Delta(A_i)} (\pi_{h,i}(\mu_i) - V^\mu_{h,i}(\rho)) \leq \epsilon\) for all \(\rho \in \Delta(S)\) and \(h \in [H]\). In the infinite-horizon setting, they can be equivalently defined as satisfying \(\max_{i \in N} \max_{\mu_i \in \Delta(A_i)} (V^\mu_i(s) - V^\pi_i(s)) \leq \epsilon\). If the above conditions only hold for certain \(\rho\) and \(h\), we refer to them as Markov non-perfect NE and CCE, respectively. Unless otherwise noted, we hereafter focus on Markov perfect equilibria, and sometimes refer to them simply as Markov equilibria when it is clear from the context. In the infinite-horizon setting, if additionally, the policy is stationary, then they are referred to as a Markov stationary NE and CCE, respectively.

2.2 Multi-player zero-sum games with networked separable interactions

As a generalization of two-player zero-sum matrix games, (multi-player) zero-sum polymatrix games have been introduced in [31, 28, 23, 24]. A polymatrix game, also known as a separable network game, is defined by a tuple \((G = (N, E_r), A = \prod_{i \in N} A_i, (r_{i,j})_{(i,j) \in E_r})\). Here, \(G\) is an undirected connected graph where \(N = [n]\) denotes the set of players and \(E_r \subseteq N \times N\) denotes the set of edges describing the rewards’ networked structures, where the graph neighborhoods represent the interactions among players. For each edge, a two-player game is defined for players \(i\) and \(j\), with action sets \(A_i, A_j\), and reward functions \(r_{i,j} : A_i \times A_j \rightarrow \mathbb{R}\), and similarly for \(r_{j,i}\). The reward for player \(i\) for a given joint action \(\alpha = (a_i)_{i \in N} \in \prod_{i \in N} A_i\) is calculated as the sum of the rewards for all edges involving player \(i\), that is, \(r_i(\alpha) = \sum_{(i,j) \in E_r} r_{i,j}(a_i, a_j)\). To be consistent with our terminology later, hereafter, we also refer to such games as (multi-player) games with Networked separable interactions (NGs).

In a zero-sum polymatrix game (i.e., a (multi-player) zero-sum Game with Networked separable interactions (zero-sum NG)), the sum of rewards for all players at any joint action \(\alpha = (a_i)_{i \in N} \in \prod_{i \in N} A_i\) equals zero, i.e., \(\sum_{i \in N} r_i(\alpha) = 0\). One can define the policy of agent \(i\), i.e., \(\pi_i \in \Delta(A_i)\), so that the agent takes actions by sampling \(a_i \sim \pi_i(\cdot)\). Note that \(\pi_i\) can be viewed as the reduced case of the policy defined in Section 2.1 when \(S = 0\) and \(H = 1\). The expected reward for player \(i\)
We now introduce our model of multi-player zero-sum MGs with networked separable interactions.

with respect to a set which at least covers the polymatrix structure of the reduced case regarding only reward functions. As our hope is to exploit the (multi-player) zero-sum MG with networked separable interactions

A finite-horizon MG is called a

3 Multi-Player (Zero-Sum) MGs with Networked Separable Interactions

We now introduce our model of multi-player zero-sum MGs with networked separable interactions.

3.1 Definitions

Definition 1. An infinite-horizon \( \gamma \)-discounted MG is called a (multi-player) MG with Networked separable interactions (NMG) characterized by a tuple \((G = (N, E_Q), S, A, P, (r_i)_{i \in N}, \gamma)\) if for any function \( V : S \rightarrow \Re \), defining \( Q^V_i(s, a) := r_i(s, a) + \gamma \sum_{s' \in S} P(s' \mid s, a)V(s') \), there exist a set of functions \((Q^V_i)_{(i,j) \in E_Q}\) and an undirected connected graph \( G = (N, E_Q) \) such that \( Q^V_i(s, a) = \sum_{j \in E_Q,i} Q^V_j(s_i, a_i, a_j) \) holds for every \( i \in N, s \in S, a \in A, \) where \( E_Q,i \) denotes the neighbors of player \( i \) induced by the edge set \( E_Q \) (without including \( i \)). When it is clear from the context, we represent the NMG tuple simply as \( G = (N, E_Q) \).

A finite-horizon MG is called a (multi-player) MG with Networked separable interactions if for any set of functions \( V := \{V_h\}_{h \in [H+1]} \) where \( V_h : S \rightarrow \Re \), defining \( Q^V_h(s, a) := r_{h,i}(s, a) + \gamma \sum_{s' \in S} P_h(s' \mid s, a)V_{h+1}(s') \), there exist a set of functions \((Q^V_i)_{(i,j) \in E_Q,h \in [H]}\) such that \( Q^V_i(s, a) = \sum_{j \in E_Q,i} Q^V_j(s_i, a_i, a_j) \) holds for every \( i \in N, s \in S, h \in [H], a \in A \).

A (multi-player) zero-sum MG with Networked separable interactions (zero-sum NMG) if additionally \((G = (N, E_Q), A = \prod_{i \in N} A_i, (r_{i,j}(s))_{(i,j) \in E_Q})\) forms a zero-sum NG for all \( s \in S \) in the infinite-horizon \( \gamma \)-discounted case, or \((G = (N, E_Q), A = \prod_{i \in N} A_i, \{r_{i,j}(s) := Q^{\theta}_{i,j}(s)\}_{(i,j) \in E_Q})\) forms a zero-sum NG for all \( s \in S \) in the finite-horizon case.

Regarding the assumption that the above conditions hold under any (set of) functions \( V \), one may understand this as a structural requirement to inherit the polymatrix structure in the Markov game case. It is natural since \( \{Q^V_i\}_{i \in N} \) would play the role of the payoff matrix in the normal-form case, when value-(iteration) based algorithms are used to solve the MG. As our hope is to exploit the networked structure in the payoff matrices to develop efficient algorithms for solving such MGs, if we do not know a priori which value function estimate \( V \) will be encountered in the algorithm update, the networked structure may easily break if we do not assume them to hold for all possible \( V \). Moreover, such a definition easily encompasses the normal-form case, by preserving the polymatrix structure of the reward functions (when substituting \( V \) to be a zero function). Some alternative definition (see Remark 2) may not necessarily preserve the polymatrix structure of even the reward functions in a consistent way (see Section B.3 for a concrete example). We thus focus on Definition 1 which at least covers the polymatrix structure of the reduced case regarding only reward functions.

Indeed, such a networked structure in Markov games may be fragile. We now propose both sufficient and necessary conditions for the reward function’s structure and the transition dynamics of the MG, to be an NMG. Here we focus on the infinite-horizon discounted setting for a simpler exposition. For finite-horizon cases, a similar statement holds, which is deferred to Appendix B. We also defer the full statement and proof of the following result to Appendix B. We first introduce the definition of decomposability and the set \( N_C \), which will be used in establishing the conditions. For a graph \( G = (N, E) \), we define \( N_C := \{i \mid (i,j) \in E \text{ for all } j \in N\} \), which may be an empty set if no such node \( i \) exists.

Definition 2 (Decomposability). A non-negative function \( f : X^{[D]} \rightarrow \Re^+ \cup \{0\} \) is decomposable with respect to a set \( D \neq \emptyset \) if there exists a set of non-negative functions \((f_i)_{i \in D}\) with \( f_i : X \rightarrow \Re^+ \cup \{0\} \), such that \( f(x) = \sum_{i \in D} f_i(x_i) \) holds for any \( x \in X^{[D]} \). A non-negative function \( f : X^{[D]} \rightarrow \Re^+ \cup \{0\} \) is decomposable with respect to a set \( D = \emptyset \), if there exists a non-negative constant \( f_o \) such that \( f(x) = f_o \) holds for any \( x \in X^{[D]} \).
Proposition 1. For a given graph $G = (\mathcal{N}, \mathcal{E}_Q)$, an MG $(\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, (r_i)_{i \in \mathcal{N}}, \gamma)$ with more than two players is an NG with respect to $G$ if and only if: (1) $r_i(s, a_i, \cdot)$ is decomposable with respect to $\mathcal{E}_Q$, for each $i \in \mathcal{N}$, $s \in \mathcal{S}$, $a_i \in \mathcal{A}_i$, i.e., $r_i(s, a) = \sum_{j \in \mathcal{E}_Q} r_{ij}(s, a_i, a_j)$ for a set of functions \{\(r_{ij}(s, a_i, \cdot)\)\}_{j \in \mathcal{E}_Q}, and (2) the transition dynamics $\mathbb{P}(s' | s, a)$ is decomposable with respect to the set $\mathcal{N}_C$ of this $G$, i.e., $\mathbb{P}(s' | s, a) = \sum_{j \in \mathcal{E}_Q} \mathbb{P}(s' | s, a)_{ij}$ for a set of functions \{\(\mathbb{P}(s' | s, a)_{ij}\)\}_{j \in \mathcal{E}_Q}$ if $\mathcal{N}_C \neq \emptyset$, or $\mathbb{P}(s' | s, a) = \mathbb{P}_0(s' | s)$ for some constant function (of $a$) $\mathbb{P}_0(s' | s)$ if $\mathcal{N}_C = \emptyset$.

Moreover, an MG qualifies as a zero-sum NG if and only if it satisfies an additional condition: the NG, characterized by $(G, \mathcal{A}, (r_{ij}(s)(i,j)_{j \in \mathcal{E}_Q}))$, is a zero-sum NG for all $s \in \mathcal{S}$. In the case of two players, every (zero-sum) Markov game becomes a (zero-sum) NG.

Remark 1 (Stronger sufficient condition). We note that for an MG $(\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, (r_i)_{i \in \mathcal{N}}, \gamma)$, if for every agent $i$, $r_i(s, a_i, \cdot)$ is decomposable with respect to $\mathcal{E}_Q \subseteq \mathcal{E}_Q$ and $\mathbb{P}(s' | s, \cdot)$ is decomposable with respect to some $\mathcal{N}_P \subseteq \mathcal{N}_C$, then one can still prove the if part, i.e., there exists some $G = (\mathcal{N}, \mathcal{E}_Q)$ such that the game is an NG with respect to this $G$. See Figure 1 for the illustration. This is because by our definition, being decomposable with respect to a subset implies being decomposable with respect to a larger set, as one can choose the functions $f_i$ for the $i$ in the complement of the subset to be simply zero. We chose to state as in Proposition 1 just for the purpose of presenting both the if and only if conditions in a concise and unified way.

Remark 2 (An alternative NG definition). Another reasonable definition of NG may be as follows: for any policy $\pi$, there exist a set of functions $(Q_i^\pi)_{i \in \mathcal{N}}$, and an undirected connected graph $G = (\mathcal{N}, \mathcal{E}_Q)$ such that $Q_i^\pi(s, a) = \sum_{j \in \mathcal{E}_Q} Q_{ij}^\pi(s, a_i, a_j)$ holds for every $i \in \mathcal{N}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$. Note that such a definition can be useful in developing policy-based algorithms (while Definition 1 is more amenable to developing value-based algorithms), e.g., policy iteration, policy gradient, actor-critic methods, where the $Q$-value under certain policy $\pi$ will appear in the updates and may need to preserve certain decomposability structure, for any policy $\pi$ encountered in the algorithm updates. However, in this case, we cannot always guarantee the decomposability of $\mathbb{P}(s' | s, \cdot)$ or $r_i(s, a_i, \cdot)$. For example, if we assume that $r_i(s, a) = 0$ for every $i \in \mathcal{N}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, then $Q_i^\pi(s, a) = \sum_{j \in \mathcal{E}_Q} \mathbb{P}(s' | s, a)_{ij}$ and thus $Q_i^\pi(s, a_i, a_j)$ is decomposable regardless of $\mathbb{P}(s' | s, a)$.

However, interestingly, we can show that the decomposability of the transition dynamics and the reward function as in Proposition 1 can still be guaranteed, as long as some degenerate cases as above do not occur. In particular, if there exist no $i \in \mathcal{N}$ and $s \in \mathcal{S}$ such that $Q_i^\pi(s, a)$ is a constant function of $a$ for any $\pi$, then the results in Proposition 1 and hence after still hold. We defer a detailed discussion on this alternative definition to Appendix B.

Remark 3 (Implication of decomposable transition dynamics). For an MG to be an NG, by Proposition 1, the transition dynamics should be decomposable, i.e., $\mathbb{P}(\cdot | s, a) = \sum_{j \in \mathcal{E}_Q} \mathbb{P}_j(\cdot | s, a_j)$ or $\mathbb{P}(\cdot | s, a) = \mathbb{P}_o(s' | s)$ if $\mathcal{N}_C \neq \emptyset$. We first focus on the discussion of the former case. Define $w_j(s, a_j) := \sum_{s' \in \mathcal{S}} \mathbb{P}_j(s' | s, a_j)$. If we fix the value of $s$ and $a_j$, then $w_j(s, a_j)$ has to be the same for different values of $a_j$ due to the fact $\sum_{s' \in \mathcal{S}} \mathbb{P}_j(s' | s, a_j) = 1$. Also note that by definition, $w_j(s, a_j)$ does not depend on the choice of this fixed $a_j$. Therefore, such a $w_j(s, a_j)$ can be written as $w_j(s)$, where $w_j(s) = \sum_{s' \in \mathcal{S}} \mathbb{P}_j(s' | s, a_j)$ for all $a_j \in \mathcal{A}_j$. We can thus rewrite $\mathbb{P}_j(\cdot | s, a_j)$ using some actual probability distributions $\mathbb{P}_j|_{s, a_j}$ such that if $w_j(s) \neq 0$, then we rewrite $\mathbb{P}_j$ as $\mathbb{P}_j(s' | s, a_j) = w_j(s) \mathbb{P}_j|_{s, a_j}(s' | s, a_j)$ and if $w_j(s) = 0$, we rewrite $\mathbb{P}_j$ as $\mathbb{P}_j(s' | s, a_j) = w_j|_{s, a_j}(s) \mathbb{P}_j|_{s, a_j}(s' | s, a_j)$ for an arbitrary probability distribution $\mathbb{P}_j|_{s, a_j}$. Notice that $\sum_{s' \in \mathcal{S}} \mathbb{P}_j|_{s, a_j}(s' | s, a_j) = 1$ for any $j \in \mathcal{N}_C$. Then, the decomposable transition dynamics can be represented as $\mathbb{P}(\cdot | s, a) = \sum_{j \in \mathcal{E}_Q} w_j(s) \mathbb{P}_j|_{s, a_j}(\cdot | s, a_j)$, i.e., an ensemble of the transition dynamics that is only controlled by single controllers. The model’s transition dynamics thus act according to the following two steps: (1) sampling the controller according to the distribution Multinomial$(\{w_j(s)|_{s, a_j}\}_{j \in \mathcal{E}_Q})$, and (2) transitioning the state following the sampled controller’s dynamics. Such a model has also been investigated under the name of transition dynamics with additive structures in [33]. Note that our model is more general and thus covers the single-controller MG setting [34], where there is only one agent controlling the transition dynamics at all states. It also covers the setting of turn-based MGs [34], where in each round, depending on the current state $s$, the transition dynamics is by turns affected by only one of the agents. This can be captured by the proper choice of $(\{w_j(s)|_{s, a_j}\}_{j \in \mathcal{E}_Q})$ that takes value 1 only for one agent at each state $s$ (while takes value 0 for all other non-controller agents at each state $s$). Additionally, the second case where $\mathbb{P}(\cdot | s, a) = \mathbb{P}_o(s' | s)$ corresponds to the one with no ensemble of controller agents.

Proposition 2 (Decomposition of $(Q_i^V)_{i \in \mathcal{N}}$). For an infinite-horizon $\gamma$-discounted NG with $G = (\mathcal{N}, \mathcal{E}_Q)$ such that $\mathcal{N}_C \neq \emptyset$, if we know that $\mathbb{P}(s' | s, a) = \sum_{j \in \mathcal{E}_Q} F_j(s' | s, a)$, and
We introduce this canonical decomposition since the representation of a CCE leads to a NE, which makes it computationally tractable to find the NE [23, 24]. We now provide below a counterpart in the Markov game setting, and provide a more detailed statement of the result in Appendix B.

**Proposition 3.** Given an ε-approximate Markov CCE of an infinite-horizon γ-discounted zero-sum NMG, marginalizing it at each state results in an \((\frac{n+1}{1-\gamma})\)-ε-approximate Markov NE of the zero-sum NMG. The same argument also holds for the finite-horizon episodic setting with \((1-\gamma)^{-1}\) being replaced by \(H\).
This result holds for both stationary and non-stationary $\epsilon$-approximate Markov CCEs. We defer the proof of Proposition 3 to Appendix B. This proposition suggests that if we can have some algorithms to find an approximate Markov CCE for a zero-sum NGM, we can obtain an approximate Markov NE by marginalizing the approximate CCE at each state. We also emphasize that the Markovian property of the equilibrium policies is important for the result to hold. As a result, the learning algorithm in [30], which learns an approximate Markov non-stationary CCE with polynomial time and samples, may thus be used to find an approximate Markov non-stationary NE in zero-sum NGMs. However, as the focus of [30] was the more challenging setting of model-free learning, the complexity therein has a high dependence on the problem parameters, and the algorithm can only find non-perfect equilibria. When it comes to (perfect) equilibrium computation, one may exploit the multi-player zero-sum structure of zero-sum NGMs, and develop more natural and faster algorithms to find a Markov non-stationary NE. Moreover, when it comes to stationary equilibrium computation, even Markov CCE is not tractable in general-sum cases [30, 29]. Hereafter, we will focus on approaching zero-sum NGMs from these perspectives.

4 Hardness for Stationary CCE Computation

Given the results in Section 3.3, it seems tempting and sufficient to compute the Markov CCE of the zero-sum NGM. Indeed, computing CCE (and thus NE) in zero-sum polymatrix games is known to be tractable [23, 24]. It is thus natural to ask: Is finding Markov CCE computationally tractable? Next, we answer the question with different answers for finding stationary CCE (in infinite-horizon $\gamma$-discounted setting) and non-stationary CCE (in finite-horizon episodic setting), respectively.

For two-player infinite-horizon $\gamma$-discounted zero-sum NGMs, significant progress in computing/learning the (Markov) stationary NE has been made recently [35, 36, 37, 38, 39, 40, 41, 42]. On the other hand, for multi-player general-sum NGMs, recent results in [30, 29] showed that computing (Markov) stationary CCE can be PPAD-hard and thus believed to be computationally intractable. We next show that this hardness persists in most non-degenerate cases even if one enforces the zero-sum and networked interaction structures in the multi-player case. We state the formal result as follows, whose detailed proof is available in Section C.

Theorem 1. There is a constant $\epsilon > 0$ for which computing an $\epsilon$-approximate Markov perfect stationary CCE in infinite-horizon $\frac{1}{2}$-discounted zero-sum NGMs, whose underlying network structure contains either a triangle or a 3-path subgraph, is PPAD-hard. Moreover, given the PCP for PPAD conjecture [43], there is a constant $\epsilon > 0$ such that computing even an $\epsilon$-approximate Markov non-perfect stationary CCE in such zero-sum NGMs is PPAD-hard.

Proof Sketch of Theorem 1. Due to space constraints, we focus on the case with three players, and the underlying network structure has a triangle subgraph. Proof for the 3-path case is similar and can be found in Section C. We will show that for any general-sum two-player turn-based NGM (A), the problem of computing its Markov stationary CCE, which is inherently a PPAD-hard problem [30], can be reduced to computing the Markov stationary CCE of a three-player zero-sum NGM with a triangle structure networked separable interactions (B). Consider an MG (A) with two players, players 1 and 2, and reward functions $r_1(s, a_1, a_2)$ and $r_2(s, a_2, a_1)$, where $a_i$ is the action of the $i$-th player and $r_i$ is the reward function of the $i$-th player. The transition dynamics is given by $P(s' | s, a_1, a_2)$. In even rounds, player 2’s action space is limited to Noop2, and in odd rounds, player 1’s action space is limited to Noop1, where Noop is an abbreviation of “no-operation”, i.e., the player does not affect the transition dynamics or the reward in that round. We denote player 1’s action space in even rounds as $A_{1,\text{even}}$ and player 2’s action space in odd rounds as $A_{2,\text{odd}}$, respectively.

Now, we construct a three-player zero-sum NGM, with a triangle network structure. We set the reward function as $r_i(s, a) = \sum_{j \neq i} r_{i,j}(s, a_i, a_j)$ and $r_{i,j}(s, a_i, a_j) = -r_{j,i}(s, a_j, a_i)$. The reward functions are designed so that $r_{i,j} = -r_{j,i}$ for all $i, j, r_{1,2} + r_{1,3} = r_1$, and $r_{2,1} + r_{3,2} = r_2$, where $r_1, r_2$ are the reward functions in game (A), by introducing a dummy player, player 3. In even rounds, player 2’s action space is limited to Noop2, and in odd rounds, player 1’s action space is limited to Noop1. Player 3’s action space is always limited to Noop3 in all rounds. The transition dynamics is defined as $P(s' | s, a_1, a_2, a_3) = P(s' | s, a_1, a_2)$, since $a_3$ is always chosen from Noop3. In other words, player 3’s action does not affect the rewards of the other two players, nor the transition dynamics, and players 1 and 2 will receive the reward as in the two-player turn-based NG. Also, note that due to the turn-based structure of the game (A), the transition dynamics satisfy the decomposable condition in our Proposition 1, and it is thus a zero-sum NGM. In fact, turn-based dynamics can be represented as an ensemble of single controller dynamics, as we have discussed in Section 5.
In this section, we study the fictitious-play property of multi-player zero-sum games with networked separable interactions, for both the matrix and Markov game settings. Following the convention in \cite{FV18}, we refer to the games in which fictitious-play dynamics converge to the NE as the games that have the fictitious-play property. We defer the matrix game case results to Section \ref{section:matrix} where we have also established convergence of the well-known variant of FP, smooth FP \cite{CF15}, in zero-sum NGs.

Echoing the computational intractability of computing CCE of zero-sum NMG unless the underlying network structure is star-shaped in the infinite-horizon discounted setting (c.f. Theorem \ref{thm:intractability}), we now consider the FP property in such games. Note that by Proposition \ref{prop:intractability}, $\mathcal{E}_Q$ is a star-shape if and only if the reward structure is a star shape and $\mathcal{N}_C = \{1\}$, where player 1 is the center of the star (Figure \ref{fig:star}), or there are only two players in zero-sum NMG. There is already existing literature for the latter case \cite{Klopfenstein93a}, \cite{Klopfenstein93b}, so we focus on the former case, which is a single-controller case where player 1 controls the transition dynamics, i.e., $\mathbb{P}(s' | s, a) = \mathbb{P}_1(s' | s, a_1)$ for some $\mathbb{P}_1$. We now introduce the fictitious-play dynamics for such zero-sum NMGs.

Each player $i$ first initializes her beliefs of other players’ policies as uniform distributions, and also initializes her belief of the $Q$-value estimates with arbitrary values. Then, at iteration $k$, player $i$ takes the best-response action based on her belief of other players’ policies ($\pi_i^{(k)}(s^{(k)})$), and their $Q$ beliefs $Q_i^{(k)}(s^{(k)}, a)$:

$$a_i^{(k)} = \arg\max_{a_i \in A_i} Q_i^{(k)}(s^{(k)}, a_i, \pi_i^{(k)}(s^{(k))).$$

Then, player $i$ implements the action $a_i^{(k)}$, observes other players’ actions $a_{-i}^{(k)}$, and updates her beliefs as follows: for each player $i \in \mathcal{N}$, she updates her belief of the opponents’ policies as

$$\pi_i^{(k+1)}(s) = \pi_i^{(k)}(s) + 1(s = s^{(k)}) \alpha N(s) \epsilon_{a_{-i}^{(k)}} - \pi_i^{(k)}(s))$$

for all $s \in \mathcal{S}$, with stepsize $\alpha N(s) \geq 0$ where $N(s)$ is the visitation count for the state $s$; then if $i = 1$, this player 1 updates the belief of $Q_{1,j}$ for all $j \in \mathcal{N}/\{1\}$ and her own $Q_1(s, a)$ for all $s \in \mathcal{S}$ as
We defer the proof to Section D.2 due to space constraints. Note that to illustrate the idea, we only present the result for the model-based case, i.e., when the transition dynamics \( P \) is known. With this result, it is direct to extend to the model-free and learning case, where \( P \) is not known [16][19][17], still using the tool of stochastic approximation [44]. See Section D.2 for more details.

**Remark 4** (Challenges for analyzing general cases). One might ask why we had to focus on a star-shaped structure. First, for general networked structures, even in the matrix-game case, it is known that the NE values of a zero-sum NG may not be unique [24]. Hence, suppose one performs Nash-value iteration, i.e., solving for the NE of the stage game and conducting backward induction, this value iteration process does not converge in general as the number of backward steps increases, since the solution at each stage is not even unique, and there may not exist a unique fixed point. This is in stark contrast to the max and min operators in the value iteration updates for single-player and two-player zero-sum cases, respectively. By exploiting a star-shaped structure, we managed to reformulate a minimax optimization problem when solving each stage game, which makes the corresponding value iteration operator contracting, and thus iterating it infinitely converges to the unique fixed point. Second, suppose there exists some other network structure (other than star-shaped ones) that also leads to a contracting value iteration operator, then for a fixed constant \( \gamma \), the fixed point (which corresponds to the Markov stationary CCE/NE of the zero-sum NMG) becomes unique and can be computed efficiently, which contradicts our hardness result in Theorem 1. Indeed, it was the exclusion of a star-shaped structure in Theorem 1 that inspired us to consider this structure in proving the convergence of FP dynamics. That being said, we note that having a contracting value iteration operator is only a sufficient condition for the FP dynamics to converge. It would be interesting to explore other structures that enjoy the FP property for reasons beyond this contraction property. We leave this as an immediate future work.

Next, we present another positive result in light of the hardness in Theorem 1 regarding the computation of non-stationary equilibria in multi-player zero-sum NMGs.
6 Non-Stationary NE Computation

We now focus on computing an (approximate) Markov non-stationary equilibrium in zero-sum NGMs. In particular, we show that when relaxing the stationarity requirement, not only CCE, but NE, can be computed efficiently. Before introducing our algorithm, we first recall the folklore result that approximating Markov non-stationary NE in infinite-horizon discounted settings can be achieved by finding approximate Markov NE in finite-horizon settings, with a large enough horizon length (c.f. Proposition 10). Hence, we will focus on the finite-horizon setting from now on.

Before delving into the details of our algorithm, we introduce the notation \( Q_{h,i}(s) \) and \( Q_h(s) \) for \( h \in [H], i \in \mathcal{N}, s \in \mathcal{S} \) as follows:

\[
Q_{h,i}(s) := (Q_{h,i,1}(s), \ldots, Q_{h,i,i-1}(s), 0, Q_{h,i,i+1}(s), \ldots, Q_{h,i,n}(s)) \in \mathbb{R}^{|A_h| \times |\mathcal{S} |}
\]

\[
Q_h(s) := ((Q_{h,1}(s))^\top, (Q_{h,2}(s))^\top, \ldots, (Q_{h,n}(s))^\top)^\top \in \mathbb{R}^{|\mathcal{S}| \times |A_h| \times |\mathcal{S} |}
\]

Here, \( Q_{h,i,j} \) represents an estimate of the equilibrium value function with canonical decomposition (Proposition 2). Hereafter, we similarly define the notation of \( Q \) Table 2, which summarizes the iteration complexities required to find an \( \epsilon \)-NE for some zero-sum NG \((\mathcal{G}, \mathcal{A}, (Q_{h,i,j}(s))_{(i,j)\in \mathcal{E}_Q})\) for all \( s \in \mathcal{S} \), and \( \epsilon \)-NE for the zero-sum NG. We have the following formal result.

Proposition 4. Suppose that for all \( h \in [H], i \in \mathcal{N}, s \in \mathcal{S}, \) \( \mathcal{NE} \)-ORACLE \((\mathcal{G}, \mathcal{A}, (Q_{h,i,j}(s))_{(i,j)\in \mathcal{E}_Q})\) provides an \( \epsilon_h,s \)-approximate NE for the zero-sum NG \((\mathcal{G}, \mathcal{A}, (Q_{h,i,j}(s))_{(i,j)\in \mathcal{E}_Q})\) in Algorithm 6. Then, the output policy \( \pi \) in Algorithm 6 is an \( \sum_{h \in [H]} \max_{s \in \mathcal{S}} (\epsilon_h,s) \)-approximate NE for the corresponding zero-sum NGM \((\mathcal{G} = (\mathcal{N}, \mathcal{E}_Q), \mathcal{S}, \mathcal{A}, H, (\mathcal{P}_h)_{h \in [H]}, (\tau_{h,i,j}(s))_{(i,j)\in \mathcal{E}_Q, s \in \mathcal{S})\).

The proof of Proposition 4 is deferred to Section 6. In light of Proposition 4 and Table 1, we obtain Table 2, which summarizes the iteration complexities required to find an \( \epsilon \)-NE for zero-sum NGMs, with different \( \mathcal{NE} \)-ORACLE subroutines. Note that the iteration complexities are all polynomial in \( H, n, |\mathcal{S}| \), and inherit the order of dependencies on \( \epsilon \) from Table 1 for the matrix-game case. In particular, Algorithm 6 with the OMWU in [51] yields the fast rate of \( \tilde{O}(1/\epsilon) \) for the last iterate.
Acknowledgement

The authors would like to thank the anonymous reviewers of NeurIPS for their helpful feedback. C.P. acknowledges support from the Xianhong Wu Fellowship, the Korea Foundation for Advanced Studies, and the Siebel Scholarship. K.Z. acknowledges support from the Northrop Grumman – Maryland Seed Grant Program. A.O. acknowledges support from the MIT-Air Force AI Innovation Accelerator Grant and the MIT-DSTA grant 031017-00016.

References


[70] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *ICLR*, 2021.


