2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Independent Learning and Subjectivity in Mean-Field Games

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992399

***Abstract*— Independent learners naively employ single-agent learning algorithms in multi-agent systems, oblivious to the effect of other strategic agents present in their environment. This paper studies partially observed N-player mean-field games from a decentralized learning perspective with two primary objectives: (i) to study the convergence properties of independent learners, and (ii) to identify structural properties that can guide algorithm design. Toward the first objective, we study the learning iterates obtained by independent learners, and find that these iterates converge under mild conditions. We then present a notion of subjective equilibrium suitable for analyzing independent learners. Toward the second objective, we study policy updating processes subject to a so-called ϵ-satisficing condition: agents who are subjectively ϵ-best-responding at a given joint policy do not change their policy. After establishing structural results for such processes, we develop an independent learning algorithm for N-player mean-field games. Exploiting the aforementioned structural results, we give guarantees of convergence to subjective ϵ-equilibrium under self-play.**

I. Introduction

Mean-field games (MFGs) are a recent theoretical framework for studying decentralized systems with a large number of weakly coupled agents [1], [2]. In a MFG, the cost and state dynamics of an agent are influenced by the collective behaviour of others only through a distributional *mean-field* term. Mean-field games can be viewed as limit models of N-player symmetric stochastic games. A number of papers have formally examined the connection between games with finitely many players and the corresponding limit model, including the works of [3], [4], and [5].

Multi-agent reinforcement learning (MARL) is the study of the emergent behaviour in systems of interacting learning agents, with stochastic games serving as the most popular framework for modelling such systems [6]. In recent years, there has been a considerable amount of research in MARL that has aimed to produce algorithms with desirable system-wide performance and convergence properties. These efforts have lead to a number of empirically successful algorithms in settings with a small number of agents, but there are comparatively fewer works that are well-suited to large-scale systems and/or offer formal convergence analyses.

Most theoretical contributions in MARL focus on highly structured classes of stochastic games, such as two-player zero-sum games [7], [8] and N-player stochastic teams and their generalizations [9], [10]. In much of the existing literature on MARL, a great deal of information is assumed to be available to the agents while they learn. Assumptions such as full state observability ([7]–[10]) or action-sharing among all agents (e.g. [11]) are appropriate in some settings but are unrealistic in models of large-scale, decentralized systems.

Independent learners [12] are learning agents that intentionally ignore other strategic agents in their environment and naively employ techniques from single-agent learning to evaluate their performance and select their actions. This approach has two advantages: first, the computational burden at any given agent is small; second, the algorithms are truly decentralized and scalable. As such, independent learners may be well-suited for use in the large-scale systems modelled by MFGs. (Ignoring pertinent information about the system may lead to deficiencies in some independent learners. See [13] and the references therein for examples of the mixed success of independent learners.)

This paper studies independent learners in partially observed N-player mean-field games. Under mild conditions on the game, we find that learning iterates obtained by independent learners converge when all agents use soft, stationary policies. Building on this, we define a notion of subjective equilibrium that is appropriate for analyzing independent learners. We then prove that under two different information structures for the game, subjective ϵ-equilibrium exists for any $\epsilon > 0$. In the context of policy dynamics, we establish a useful structural property, to be called the subjective ϵ-satisficing paths property. Building on this structure, we develop a decentralized independent learning algorithm for N-player mean-field games, and we argue that it drives play to subjective ϵ-equilibrium under self-play. Unlike other learning algorithms for mean-field games, which typically constrain all agents to follow the same policy by considering a representative agent, our algorithm allows for agents to use different policies during learning and to use different policy updating rules when exploring their policy space, and in this sense is truly decentralized.

Due to space constraints, some material is omitted and can be found in the longer version of this paper, [14]. Notably, this includes results for a third information structure, all proofs, and additional exposition.

Notation: For standard Borel spaces A, B we let $\mathcal{P}(A)$ denote the set of probability measures on A and we let $\mathcal{P}(A|B)$ denote the set of transition kernels on A given B.

A. Related Work and Discussion

Learning in MFGs is a nascent but active research area. We now briefly review on some common themes in this young

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

B. Yongacoglu and S. Yüksel are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada, email: {1bmy,yuksel@queensu.ca}. G. Arslan is with the University of Hawaii, email: gurdal@hawaii.edu.

research area; a longer literature review is available in [14].

By and large, existing literature on learning in MFGs focuses on the standard model of MFGs, where the player set is a continuum. We selectively cite [15], [16], [17], [18], [19], [20], [21], [22] and [23] as work in this vein. Taking the set of agents to be infinite effectively removes any strategic interaction between agents; as a result, learning theory in this tradition is single-agent in spirit rather than multi-agent.

For the most part, existing work on learning in MFGs attempts to produce algorithms that compute *mean-field equilibrium*, a symmetric notion of equilibrium different than the equilibrium concepts used here. (For a definition of the symmetric notion of equilibrium considered elsewhere, see [19, Definition 2.1]; our notion of equilibrium resembles that of [19, Definition 5.1].) This is done by studying a generic, representative, agent. While this approach can be justified for the existence analysis, imposing such a generic agent perspective (in particular, requiring all agents to use the same policy) implies coordination, which may not be natural in a learning context.

In contrast, the aim of this paper is to study the convergence behaviour of truly decentralized learning algorithms that may reasonably be used in large-scale systems. Unlike the existing work on learning in MFGs, we study a finite player setting with non-trivial strategic interaction, and we place greater emphasis on the use of heterogenous policies during the learning process. In so doing, our results do not guarantee convergence to the symmetric notion of equilibrium used elsewhere in the theory, but rather to policies that form a kind of subjective equilibrium that is better suited for analyzing independent learners.

II. Model

A. N-Player Mean Field Games

For $N \in \mathbb{N}$, a partially observed N-player mean-field game (MFG) is described by the following list:

$$\mathcal{G} = (\mathcal{N}, X_{\mathrm{loc}}, \mathbf{X}, \mathbb{Y}, \mathbb{U}, \{\varphi^i\}_{i\in\mathcal{N}}, c, \beta, P_{\mathrm{loc}}, \nu_0). \quad (1)$$

The components of $\mathcal{G}$ are such:

- $\mathcal{N}$ is a set of N players/agents;
- X_{loc} is a finite set and $\mathbf{X} = \times_{i\in\mathcal{N}} X_{\mathrm{loc}}$. Elements of X_{loc} (resp. $\mathbf{X}$) are called local states (global states);
- For each $\mathbf{s} \in \mathbf{X}$, we define $\mu(\cdot|\mathbf{s}) \in \mathcal{P}(X_{\mathrm{loc}})$ as:
$$\mu(B|\mathbf{s}) = \frac{1}{N}\sum_{i\in\mathcal{N}} \delta_{s^i}(B), \quad \forall B \subseteq X_{\mathrm{loc}},$$
and we let $\mathrm{Emp}_N := \{\mu(\cdot|\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$. An element $\mu \in \mathrm{Emp}_N$ is called a *mean-field state*;
- For each $i \in \mathcal{N}$, $\varphi^i : \mathbf{X} \to \mathbb{Y}$ is an observation function, where $\mathbb{Y}$ is a finite set. We refer to the pair $(\mathbb{Y}, \{\varphi^i\}_{i\in\mathcal{N}})$ as the observation channel;
- $\mathbb{U}$ is a finite set of actions, and we let $\mathbf{U} := \times_{i\in\mathcal{N}}\mathbb{U}$. An element of $\mathbb{U}$ (resp. $\mathbf{U}$) is called an action (joint action);
- $c : X_{\mathrm{loc}} \times \mathcal{P}(X_{\mathrm{loc}}) \times \mathbb{U} \to \mathbb{R}$ is a stage cost function;
- $\beta \in (0, 1)$ is a discount factor;
- $P_{\mathrm{loc}} \in \mathcal{P}(X_{\mathrm{loc}}|X_{\mathrm{loc}}\times\mathcal{P}(X_{\mathrm{loc}})\times\mathbb{U})$ is a transition kernel governing local state transitions for each player;
- $\nu_0 \in \mathcal{P}(\mathbf{X})$ is a probability distribution for the initial global state variable, $\mathbf{x}_0$.

At time $t \in \mathbb{Z}_{\geq 0}$, player i's local state is denoted x_t^i, while the global state variable is denoted by $\mathbf{x}_t$ and the mean-field state is denoted by $\mu_t := \mu(\cdot|\mathbf{x}_t)$. Player i observes its local observation variable $y_t^i := \varphi^i(\mathbf{x}_t)$ and uses its locally observable history variable, defined below, to select an action $u_t^i \in \mathbb{U}$. The joint action at time t is denoted $\mathbf{u}_t$. Player i then incurs a cost $c_t^i := c(x_t^i, \mu_t, u_t^i)$, and player i's local state variable evolves according to $x_{t+1}^i \sim P_{\mathrm{loc}}(\cdot|x_t^i, \mu_t, u_t^i)$. This process is then repeated at time $t+1$, and so on.

For any $t \in \mathbb{Z}_{\geq 0}$, we let $\mathbf{H}_t := (\mathbf{X}\times\mathbf{U})^t \times \mathbf{X}$ and $H_t := \mathcal{P}(\mathbf{X}) \times (\mathbb{Y}\times\mathbb{U}\times\mathbb{R})^t \times \mathbb{Y}$. The set $\mathbf{H}_t$ is the set of overall system histories of length t, while the set H_t is the set of histories of length t that an individual player may observe. Elements of $\mathbf{H}_t$ are called *system histories of length* t, and we use $\mathbf{h}_t = (\mathbf{x}_0, \mathbf{u}_0, \cdots, \mathbf{u}_{t-1}, \mathbf{x}_t) \in \mathbf{H}_t$, to denote the t^{th} *system history variable*. Similarly, elements of H_t are called *observable histories of length* t, and for $i \in \mathcal{N}$, we use $h_t^i = (\nu_0, y_0^i, u_0^i, c_0^i, \cdots, c_{t-1}^i, y_t^i) \in H_t$ to denote player i's t^{th} *locally observable history variable*.

Definition 1 (Policies): A policy for player $i \in \mathcal{N}$ is a sequence $\pi^i = (\pi_t^i)_{t\geq 0}$ such that $\pi_t^i \in \mathcal{P}(\mathbb{U}|H_t)$ for every $t \geq 0$. We let Γ^i denote the set of all policies for player i.

Definition 2 (Soft Policies): For $i \in \mathcal{N}$, $\xi > 0$, $\pi^i \in \Gamma^i$ is called ξ-*soft* if $\pi^i(a|\tilde{h}_t) \geq \xi$ for all $t \geq 0$ and $\tilde{h}_t \in H_t$. A policy $\pi^i \in \Gamma^i$ is called *soft* if it is ξ-soft for some $\xi > 0$.

Notation: We let $\mathbf{\Gamma} := \times_{i\in\mathcal{N}}\mathbf{\Gamma}^i$ denote the set of *joint policies*. To isolate player i's component in a particular joint policy $\boldsymbol{\pi} \in \mathbf{\Gamma}$, we write $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i})$, where $-i$ is used in the agent index to represent all agents other than i. Similarly, we write the joint policy set as $\mathbf{\Gamma} = \Gamma^i \times \mathbf{\Gamma}^{-i}$, and so on.

Definition 3 (Stationary Policies): Let $i \in \mathcal{N}$. A policy $\pi^i \in \Gamma^i$ is called *stationary* if there exists $f^i \in \mathcal{P}(\mathbb{U}|\mathbb{Y})$ such that for any $t \geq 0$ and any $\tilde{h}_t = (\nu, \tilde{y}_0, \ldots, \tilde{y}_t) \in H_t$, we have $\pi_t^i(\cdot|\tilde{h}_t) = f^i(\cdot|\tilde{y}_t)$. We let Γ_S^i denote the set of stationary policies for player i.

For $i \in \mathcal{N}$, we identify Γ_S^i with the set $\mathcal{P}(\mathbb{U}|\mathbb{Y})$ and treat stationary policies as elements of $\mathcal{P}(\mathbb{U}|\mathbb{Y})$, omitting reference to the (complete) locally observable history.

For each $i \in \mathcal{N}$, we introduce a metric d^i on Γ_S^i, defined for all $\pi^i, \tilde{\pi}^i \in \Gamma_S^i$ as

$$d^i(\pi^i, \tilde{\pi}^i) := \max\{|\pi^i(a^i|y) - \tilde{\pi}^i(a^i|y)| : y \in \mathbb{Y}, a^i \in \mathbb{U}\}.$$

For any joint policy $\boldsymbol{\pi} \in \mathbf{\Gamma}$ and $\nu \in \mathcal{P}(\mathbf{X})$, there exists a unique probability measure on the set $(\mathbf{X}\times\mathbf{U})^\infty$. We denote this measure by $\Pr_\nu^{\boldsymbol{\pi}}$, and let $E_\nu^{\boldsymbol{\pi}}$ denote its expectation. We use this to define player i's (state) value function:

$$J_{\boldsymbol{\pi}}^i(\nu) := E_\nu^{\boldsymbol{\pi}}\left[\sum_{t=0}^\infty \beta^t c_t^i\right] = E_\nu^{\boldsymbol{\pi}}\left[\sum_{t=0}^\infty \beta^t c(x_t^i, \mu_t, u_t^i)\right].$$

Definition 4: Let $\epsilon \geq 0$, $i \in \mathcal{N}$. $\pi^{*i} \in \Gamma^i$ is called a (uniform) ϵ-*best-response to* $\boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}^{-i}$ if, $\forall \nu \in \mathcal{P}(\mathbf{X})$,

$$J_{(\pi^{*i}, \boldsymbol{\pi}^{-i})}^i(\nu) \leq \inf_{\tilde{\pi}^i \in \Gamma^i} J_{(\tilde{\pi}^i, \boldsymbol{\pi}^{-i})}^i(\nu) + \epsilon.$$

The set of ϵ-best-responses to $\boldsymbol{\pi}^{-i}$ is denoted $\mathrm{BR}_\epsilon^i(\boldsymbol{\pi}^{-i})$.

Definition 5: Let $\epsilon \geq 0$. A joint policy $\boldsymbol{\pi}^* \in \boldsymbol{\Gamma}$ is called a (perfect) *ϵ-equilibrium* if $\pi^{*i} \in \mathrm{BR}^i_\epsilon(\boldsymbol{\pi}^{*-i})$ for all $i \in \mathcal{N}$.

For $\epsilon \geq 0$, we let $\boldsymbol{\Gamma}^{\epsilon\text{-eq}} \subseteq \boldsymbol{\Gamma}$ denote the set of ϵ-equilibrium policies, and let $\boldsymbol{\Gamma}_S^{\epsilon\text{-eq}} := \boldsymbol{\Gamma}^{\epsilon\text{-eq}} \cap \boldsymbol{\Gamma}_S$. In the next section, we describe conditions under which $\boldsymbol{\Gamma}_S^{\epsilon\text{-eq}} \neq \varnothing$.

Definition 6: Let $\mathcal{G}$ be the game in (1). Let

$$\mathcal{V} = \left\{ V^i_{\boldsymbol{\pi}} : \mathbb{Y} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \boldsymbol{\Gamma}_S \right\}, \text{ and}$$
$$\mathcal{W} = \left\{ W^i_{\boldsymbol{\pi}} : \mathbb{Y} \times \mathbb{U} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \boldsymbol{\Gamma}_S \right\}$$

be two families of functions. Then, the pair $(\mathcal{V}, \mathcal{W})$ is called a *subjective function family for $\mathcal{G}$*.

Definition 7: Let $\epsilon \geq 0$ and let $(\mathcal{V}, \mathcal{W})$ be a subjective function family for $\mathcal{G}$. A policy $\pi^{*i} \in \Gamma^i_S$ is called a $(\mathcal{V}, \mathcal{W})$-subjective ϵ-best-response to $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$ if we have

$$V^i_{(\pi^{*i}, \boldsymbol{\pi}^{-i})}(y) \leq \min_{a^i \in \mathbb{U}} W^i_{(\pi^{*i}, \boldsymbol{\pi}^{-i})}(y, a^i) + \epsilon, \quad \forall y \in \mathbb{Y}.$$

This definition of $(\mathcal{V}, \mathcal{W})$-subjective ϵ-best-responding is given in analogy to an ϵ-optimality criterion for an MDP. Here, the functions $V^{*i}_{\boldsymbol{\pi}}$ (resp. $W^{*i}_{\boldsymbol{\pi}}$) are analogs to the state value function (Q-function) for the MDP.

For a fixed player $i \in \mathcal{N}$, $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$, and subjective function family $(\mathcal{V}, \mathcal{W})$, we let $\text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}, \mathcal{W}) \subseteq \Gamma^i_S$ denote i's set of $(\mathcal{V}, \mathcal{W})$-subjective ϵ-best-responses to $\boldsymbol{\pi}^{-i}$.

Definition 8: Let $\epsilon \geq 0$ and let $(\mathcal{V}, \mathcal{W})$ be a subjective function family for $\mathcal{G}$. A joint policy $\boldsymbol{\pi}^* \in \boldsymbol{\Gamma}_S$ is called a $(\mathcal{V}, \mathcal{W})$-subjective ϵ-equilibrium for $\mathcal{G}$ if, for every $i \in \mathcal{N}$, $\pi^{*i} \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{*-i}, \mathcal{V}, \mathcal{W})$.

For any subjective function family $(\mathcal{V}, \mathcal{W})$, we denote the set of $(\mathcal{V}, \mathcal{W})$-subjective ϵ-equilibria by $\mathrm{Subj}_\epsilon(\mathcal{V}, \mathcal{W}) \subseteq \boldsymbol{\Gamma}_S$.

B. On the Observation Channel and Information Structure

So far, we have left the particular observation channel $(\mathbb{Y}, \{\varphi^i\}_{i \in \mathcal{N}})$ unspecified. We now offer two alternatives for the observation channel. The particular choice used in practice will depend on the application area: in some instances, there is a natural restriction of information leading to a particular observation channel; in others, agents may voluntarily compress their observations for function approximation. For an expanded discussion, see [14].[1]

Assumption 1 (Global State Observability): $\mathbb{Y} = \mathbf{X}$ and $\varphi^i(\mathbf{s}) = \mathbf{s}$ for each global state $\mathbf{s} \in \mathbf{X}$ and player $i \in \mathcal{N}$.

Assumption 2 (Mean-Field State Observability): $\mathbb{Y} = X_{\mathrm{loc}} \times \mathrm{Emp}_N$ and $\varphi^i(\mathbf{s}) = (s^i, \mu(\cdot|\mathbf{s}))$ for each global state $\mathbf{s} \in \mathbf{X}$ and player $i \in \mathcal{N}$.

Assumption 2 is the standard observation channel considered in works on mean-field games, see e.g. [4] and the references therein.

C. Relationship with Mean-Field Games

The model above differs from the classical model of mean-field games, which assumes a continuum of agents (as in [1] or [2]). Here, we consider models with a large, finite number of symmetric, weakly coupled agents. Our model resembles the one used in [4], which studies existence of equilibrium in a model with general state and actions spaces.

[1] In [14], we also study a third observation channel, wherein agents observe only their local state and a compressed version of the mean-field state.

III. Stationary Equilibrium Policies: Existence under Two Observation Channels

Lemma 1: Let $\mathcal{G}$ be a partially observed N-player MFG. Fix player $i \in \mathcal{N}$ and let $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$. Then, player i faces a POMDP $M_{\boldsymbol{\pi}^{-i}}$ with partially observed state process $\{\mathbf{x}_t\}_{t \geq 0}$.

Under certain additional conditions, described below, one can show that player $i \in \mathcal{N}$ faces a fully observed MDP. In such cases, the classical theory of MDPs and reinforcement learning can be brought to bear on i's optimization problem, leading to results on the existence of certain equilibrium policies and characterization of one's best-response set.

A. Equilibrium under Global State Observability

Corollary 1: Let $\mathcal{G}$ be a partially observed N-player MFG in which Assumption 1 holds. Fix player $i \in \mathcal{N}$ and let $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$. Then, player i faces a (fully observed) MDP $M_{\boldsymbol{\pi}^{-i}}$ with controlled state process $\{y^i_t\}_{t \geq 0}$.

Under the conditions of Corollary 1, we can consider player i's Q-function for this environment, which we denote by $Q^{*i}_{\boldsymbol{\pi}^{-i}} : \mathbf{X} \times \mathbb{U} \to \mathbb{R}$.

$$Q^{*i}_{\boldsymbol{\pi}^{-i}}\left(\mathbf{s}, a^i\right) := E^{\boldsymbol{\pi}^*}_{\nu_0}\left[\sum_{t=0}^{\infty} \beta^t c^i_t \middle| \mathbf{x}_0 = \mathbf{s}, u^i_0 = a^i \right],$$

for each $(\mathbf{s}, a^i) \in \mathbf{X} \times \mathbb{U}$, where $\boldsymbol{\pi}^* = (\pi^{*i}, \boldsymbol{\pi}^{-i})$ and $\pi^{*i} \in \Gamma^i_S \cap \mathrm{BR}^i_0(\boldsymbol{\pi}^{-i})$.

The value $Q^{*i}_{\boldsymbol{\pi}^{-i}}(\mathbf{s}, a^i)$ represents the *optimal cost-to-go* to player i when play begins at global state $\mathbf{s} \in \mathbf{X}$, player i takes action $a^i \in \mathbb{U}$ at time 0 and follows the policy π^{*i} thereafter, and the remaining players play according to the stationary policy $\boldsymbol{\pi}^{-i}$.

Lemma 2: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying Assumption 1. Then, $\boldsymbol{\Gamma}_S^{0\text{-eq}} \neq \varnothing$.

A partially observed N-player mean-field game with global state observability (Assumption 1) is a special case of the finite N-player stochastic games studied in [24], and so Lemma 2 follows from [24, Theorem 2].

B. Equilibrium Under Mean-Field State Observability

Definition 9: Suppose Assumption 2 holds. Let $i, j \in \mathcal{N}$ and let $\pi^i \in \Gamma^i_S$, $\pi^j \in \Gamma^j_S$. We say that the policies π^i and π^j are *mean-field symmetric* if both are identified with the same transition kernel in $\mathcal{P}(\mathbb{U}|\mathbb{Y})$. For any $I \subset \mathcal{N}$, a collection of policies $\{\pi^i\}_{i \in I}$ is called mean-field symmetric if, for every $i, j \in I$, we have that π^i and π^j are mean-field symmetric.

Definition 10: A set of policies $\boldsymbol{\Pi} \subseteq \boldsymbol{\Gamma}_S$ is called *symmetric* if $\Pi^i = \Pi^j$ for all $i, j \in \mathcal{N}$.

Lemma 3: Let $\mathcal{G}$ be an N-player MFG, let $i \in \mathcal{N}$, and let Assumption 2 hold. If $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$ is mean-field symmetric, then i faces a fully observed MDP $M_{\boldsymbol{\pi}^{-i}}$ with controlled state process $\{y^i_t\}_{t \geq 0}$.

We define the Q-function for player i when playing $\mathcal{G}$ against a mean-field symmetric policy $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$ as

$$Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i) := E^{(\pi^{*i}, \boldsymbol{\pi}^{-i})}_{\nu}\left[\sum_{t=0}^{\infty} \beta^t c^i_t \middle| y^i_0 = y, u^i_0 = a^i \right],$$

for every $y \in \varphi^i(\mathbf{X}) = \{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$ and $a^i \in \mathbb{U}$, where $\pi^{*i} \in \mathrm{BR}^i_0(\boldsymbol{\pi}^{-i}) \cap \Gamma^i_S$ is a best-response to $\boldsymbol{\pi}^{-i}$ and $\nu \in$

$\mathcal{P}(\mathbf{X})$ is arbitrary (see [14] for justification). For elements $y \in \mathbb{Y} \setminus \varphi^i(\mathbf{X})$, we may define $Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i)$ arbitrarily, say $Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, \cdot) \equiv 0$.

To our knowledge, the following result appears to be new.

Theorem 1: Let $\mathcal{G}$ be a partially observed N-player MFG satisfying Assumption 2. For any $\epsilon \geq 0$, $\mathbf{\Gamma}_S^{\epsilon\text{-eq}} \neq \varnothing$.

IV. Convergence of Naive Single-Agent Learning under Stationary Policies

In this section, we study the convergence of learning iterates obtained when player $i \in \mathcal{N}$ naively runs single-agent reinforcement learning algorithms that treat $\{y_t^i\}_{t\geq 0}$ as if it were the state variable of a MDP. This learning process is formalized in Algorithm 1, below, where we have fixed the policies of all players to be stationary. By fixing policies to be stationary, this section focuses on the effect that decentralized information has on independent learners, leaving aside the well-known challenge of non-stationary [25].

Assumption 3: Under any soft stationary policy $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, the global state process $\{\mathbf{x}_t\}_{t\geq 0}$ is an irreducible, aperiodic Markov chain on $\mathbf{X}$.

Theorem 2: Let $\mathcal{G}$ be a partially observed N-player MFG satisfying Assumption 3, let $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$ be soft and $\nu \in \mathcal{P}(\mathbf{X})$. Suppose $i \in \mathcal{N}$ uses Algorithm 1 to obtain $\{\bar{J}_t^i, \bar{Q}_t^i\}_{t\geq 0}$. Then, there exist deterministic functions $\tilde{V}^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \to \mathbb{R}$ and $\tilde{W}^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \times \mathbb{U} \to \mathbb{R}$ such that, $\Pr_\nu^{\boldsymbol{\pi}}$-almost surely, we have $\lim_{t\to\infty} \bar{J}_t^i = \tilde{V}^{*i}_{\boldsymbol{\pi}}$ and $\lim_{t\to\infty} \bar{Q}_t^i = \tilde{W}^{*i}_{\boldsymbol{\pi}}$. If Assumption 1 holds, then $\tilde{V}^{*i}_{\boldsymbol{\pi}}(\mathbf{s}) = J^i_{\boldsymbol{\pi}}(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{X}$ and $\tilde{W}^{*i}_{\boldsymbol{\pi}} = Q^{*i}_{\boldsymbol{\pi}^{-i}}$. If Assumption 2 holds and $\boldsymbol{\pi}^{-i}$ is mean-field symmetric, then $\tilde{V}^{*i}_{\boldsymbol{\pi}}(\varphi^i(\mathbf{s})) = J^i_{\boldsymbol{\pi}}(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{X}$ and $\tilde{W}^{*i}_{\boldsymbol{\pi}} = Q^{*i}_{\boldsymbol{\pi}^{-i}}$.

The proof of Theorem 2 builds on [26, Theorem 4.1] and can be found in [14].

Algorithm 1: Naive Learning in an N-player MFG

Initialize Soft $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, $\bar{Q}_0^i = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{U}}$ and $\bar{J}_0^i = 0 \in \mathbb{R}^{\mathbb{Y}}$

for $t \geq 0$ (t^{th} stage game)

Player i observes y_t^i, selects $u_t^i \sim \pi^i(\cdot|y_t^i)$

Players $-i$ select $\mathbf{u}_t^{-i}$ according to $\boldsymbol{\pi}^{-i}$

Player i observes y_{t+1}^i and cost $c_t^i := c^i(\mathbf{x}_t, u_t^i, \mathbf{u}_t^{-i})$

$n_t^i := \sum_{k=0}^t \mathbf{1}\{(y_k^i, u_k^i) = (y_t^i, u_t^i)\}$

$m_t^i := \sum_{k=0}^t \mathbf{1}\{y_k^i = y_t^i\}$

Q-factor update:

$$\bar{Q}^i_{t+1}(y_t^i, u_t^i) = \left(1 - \frac{1}{n_t^i}\right) \bar{Q}_t^i(y_t^i, u_t^i) + \frac{1}{n_t^i}\left(c_t^i + \beta \min_{a^i \in \mathbb{U}} \bar{Q}_t^i(y_{t+1}^i, a^i)\right),$$

and $\bar{Q}^i_{t+1}(y, a) = \bar{Q}_t^i(y, a)$ for all $(y, a) \neq (y_t^i, u_t^i)$.

Value function update:

$$\bar{J}^i_{t+1}(y_t^i) = \left(1 - \frac{1}{m_t^i}\right) \bar{J}_t^i(y_t^i) + \frac{1}{m_t^i}\left(c_t^i + \beta \bar{J}_t^i(y_{t+1}^i)\right),$$

and $\bar{J}^i_{t+1}(y) = \bar{J}_t^i(y)$ for all $y \neq y_t^i$.

Remark: The limiting quantities $\tilde{V}^{*i}_{\boldsymbol{\pi}}$ and $\tilde{W}^{*i}_{\boldsymbol{\pi}}$ do not, in general, have inherent relevance to player i's objective function in $\mathcal{G}$. These quantities should instead be interpreted as subjective beliefs of player i, obtained through a naive learning process.

We conclude this section by introducing notation for the subjective function family corresponding to each agent's subjective beliefs obtained through Algorithm 1.

Definition 11: Let $\mathcal{V}^* = \left\{V^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \mathbf{\Gamma}_S\right\}$ be the collection of functions defined as follows: for each $i \in \mathcal{N}$ and $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, $V^{*i}_{\boldsymbol{\pi}} := \tilde{V}^{*i}_{\boldsymbol{\pi}}$ if $\boldsymbol{\pi}$ is soft and $V^{*i}_{\boldsymbol{\pi}} \equiv \frac{\|c\|_\infty}{1-\beta} + 1$ otherwise.

Let $\mathcal{W}^* = \left\{W^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \times \mathbb{U} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \mathbf{\Gamma}_S\right\}$ be collection of functions defined as follows: for each $i \in \mathcal{N}$ and $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, $W^{*i}_{\boldsymbol{\pi}} := \tilde{W}^{*i}_{\boldsymbol{\pi}}$ if $\boldsymbol{\pi}$ is soft and $W^{*i}_{\boldsymbol{\pi}} \equiv 0$ otherwise.

The pair $(\mathcal{V}^*, \mathcal{W}^*)$ is called the *naively learned subjective function family* for $\mathcal{G}$.

For non-soft $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, we define $V^{*i}_{\boldsymbol{\pi}}$ and $W^{*i}_{\boldsymbol{\pi}}$ as done above in order to avoid introducing $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium that may not be the outcome of the naive learning process.

A. Existence of Subjective Equilibrium

Lemma 4: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying either Assumptions 1 and 3 or Assumptions 2 and 3. Let $\epsilon > 0$. Then, $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$.

V. Subjective Satisficing Paths

In this section, we ask whether a discrete-time dynamical system on the set of joint policies $\mathbf{\Gamma}_S$ can drive the joint policy to subjective ϵ-equilibrium by changing only the policies of those agents who are not subjectively ϵ-best-responding. For the following definitions, let $\mathcal{G}$ be a partially observed N-player MFG, let $i \in \mathcal{N}$, let $\epsilon \geq 0$, and let $(\mathcal{V}, \mathcal{W})$ be a subjective function family for $\mathcal{G}$.

Definition 12: A sequence $(\boldsymbol{\pi}_k)_{k\geq 0}$ in $\mathbf{\Gamma}_S$ is called a $(\mathcal{V}, \mathcal{W})$-*subjective ϵ-satisficing path* if, $\forall i \in \mathcal{N}, k \geq 0$,

$$\pi_k^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}_k^{-i}, \mathcal{V}, \mathcal{W}) \Rightarrow \pi^i_{k+1} = \pi^i_k.$$

Definition 13: Let $\mathbf{\Pi} \subseteq \mathbf{\Gamma}_S$. The game $\mathcal{G}$ is said to have the $(\mathcal{V}, \mathcal{W})$-*subjective ϵ-satisficing paths property within* $\mathbf{\Pi}$ if, for every $\boldsymbol{\pi} \in \mathbf{\Pi}$, there exists a $(\mathcal{V}, \mathcal{W})$-subjective ϵ-satisficing path $(\boldsymbol{\pi}_k)_{k\geq 0}$ satisfying (i) $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$; (ii) $\boldsymbol{\pi}_k \in \mathbf{\Pi}$ $\forall k \geq 0$; (iii) $\exists K < \infty$ such that $\boldsymbol{\pi}_K \in \mathrm{Subj}_\epsilon(\mathcal{V}, \mathcal{W})$.

A. Naively Learned Subjective Functions and ϵ-Satisficing

We now shift our attention to $(\mathcal{V}^*, \mathcal{W}^*)$, the naively learned subjective function family for $\mathcal{G}$.

1) Paths Under Global State Observability:

Definition 14: Let $\mathcal{G}$ be a partially observed N-player mean-field game for which Assumption 1 is satisfied, and let $i \in \mathcal{N}$. A stationary policy $\pi^i \in \Gamma_S^i$ is said to be *of the mean-field type* if there exists $f^i \in \mathcal{P}(\mathbb{U}|X_{\mathrm{loc}} \times \mathrm{Emp}_N)$ such that $\pi^i(\cdot|\mathbf{s}) = f^i\left(\cdot\middle|s^i, \mu(\cdot|\mathbf{s})\right)$ for every global state $\mathbf{s} \in \mathbf{X}$.

We identify each stationary policy of the mean-field type with its associated transition kernel in $\mathcal{P}(\mathbb{U}|X_{\mathrm{loc}} \times \mathrm{Emp}_N)$. We extend the definitions of mean-field symmetry and symmetric sets of joint policies to this context.

Theorem 3: Let $\epsilon > 0$ and let $\mathcal{G}$ be a partially observed N-player mean-field game for which Assumptions 1 and 3 hold. Suppose $\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is a soft, symmetric subset of policies

satisfying (i) every $\boldsymbol{\pi} \in \mathbf{\Pi}$ is of the mean-field type, and (ii) $\mathbf{\Pi} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$. Then, $\mathcal{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\mathbf{\Pi}$.

2) Paths under Mean-Field State Observability:

Theorem 4: Let $\epsilon > 0$ and let $\mathcal{G}$ be a partially observed N-player mean-field game for which Assumptions 2 and 3 hold. Suppose $\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is a soft, symmetric subset satisfying $\mathbf{\Pi} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$. Then, $\mathcal{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\mathbf{\Pi}$.

The proofs of Theorem 3 and 4 can be found in [14].

B. Quantization of the Policy Space

For algorithm design purposes, it is advantageous to restrict policy selection to a finite subset of $\mathbf{\Gamma}_S$. If the restricted set of policies is obtained via a sufficiently fine, symmetric quantization of the original set, then the performance loss for a given agent is negligible and the structural properties of the previous section hold. For the following definitions, let $\mathcal{G}$ be a partially observed N-player mean-field game, and let $i \in \mathcal{N}$. Recall that d^i is a metric on the set Γ_S^i.

Definition 15: Let $\xi > 0$ and $\tilde{\Pi}^i \subseteq \Gamma_S^i$. A mapping $q^i : \tilde{\Pi}^i \to \tilde{\Pi}^i$ is called a *ξ-quantizer (on $\tilde{\Pi}^i$)* if (i) $q^i(\tilde{\Pi}^i) := \{q^i(\pi^i) : \pi^i \in \tilde{\Pi}^i\}$ is a finite set and (ii) $d^i(\pi^i, q^i(\pi^i)) < \xi$ for all $\pi^i \in \tilde{\Pi}^i$.

Definition 16: Let $\xi > 0$ and let $\tilde{\Pi}^i \subseteq \Gamma_S^i$. A set of policies $\Pi^i \subseteq \tilde{\Pi}^i$ is called a ξ-quantization of $\tilde{\Pi}^i$ if $\Pi^i = q^i(\tilde{\Pi}^i)$, where q^i is some ξ-quantizer on $\tilde{\Pi}^i$.

A set $\Pi^i \subseteq \Gamma_S^i$ is called a quantization of Γ_S^i if it is a ξ-quantization of Γ_S^i for some $\xi > 0$. A quantization Π^i is called *soft* if each policy $\pi^i \in \Pi^i$ is soft. The expression "fine quantization" will be used to reflect that a policy subset is a ξ-quantization for suitably small ξ. We extend the definitions and terminological conventions above to also refer to quantizers and quantizations of sets of joint policies. For instance, $\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is a ξ-quantization of $\mathbf{\Gamma}_S$ if each Π^i is a ξ-quantization of Γ_S^i, and so on.

Lemma 5: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying Assumptions 1 and 3. Let $\epsilon > 0$. There exists $\xi = \xi(\epsilon) > 0$ such that if $\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is any soft ξ-quantization of $\mathbf{\Gamma}_S$, then we have $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \mathbf{\Pi} \neq \varnothing$. Moreover, if $\mathbf{\Pi}_{\mathrm{MF}} \subset \mathbf{\Gamma}_S$ is the set of joint stationary policies of the mean-field type, there exists $\xi = \xi(\epsilon) > 0$ such that if $\mathbf{\Pi}$ is a soft, symmetric ξ-quantization of $\mathbf{\Pi}_{\mathrm{MF}}$, then we have (1) $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \mathbf{\Pi} \neq \varnothing$; (2) $\mathcal{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\mathbf{\Pi}$.

Lemma 6: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying Assumptions 2 and 3. Let $\epsilon > 0$. There exists $\xi = \xi(\epsilon) > 0$ such that if $\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is any soft, symmetric ξ-quantization of $\mathbf{\Gamma}_S$, then (1) $\mathbf{\Gamma}^{\epsilon\text{-eq}} \cap \mathbf{\Pi} \neq \varnothing$ and $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \mathbf{\Pi} \neq \varnothing$; (2) $\mathcal{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\mathbf{\Pi}$.

Lemmas 5 and 6 guarantee that the game $\mathcal{G}$ has the subjective satisficing paths property within finely quantized subsets of policies. This has two desirable consequences for algorithm design. First, players can restrict their search from an uncountable set to a finite subset of policies with only a small loss in performance. Second, since the game $\mathcal{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\mathbf{\Pi}$, play can be driven to subjective ϵ-equilibrium by changing only the policies of those players that are "ϵ-unsatisfied." We thus obtain a stopping condition, whereby player i can settle on a policy whenever it is subjectively ϵ-best-responding.

Taken together, these points remove the need for *coordinated* search of $\mathbf{\Pi}$: play can be driven to ϵ-equilibrium even by random policy updating by those players that are not subjectively ϵ-best-responding. This structure also removes the need for specialized policy updating rules that taken into account special structure in the game.

VI. Learning Algorithm and Convergence Results

Algorithm 2: Independent Learning

Set Parameters
- $\Pi^i \subset \Gamma_S^i$: a fine quantization of Γ_S^i
- $\{T_k\}_{k \geq 0}$: a sequence in $\mathbb{N}$ of learning phase lengths
  - set $t_0 = 0$ and $t_{k+1} = t_k + T_k$ for all $k \geq 0$.
- $e^i \in (0, 1)$: random policy updating probability
- $d^i \in (0, \infty)$: tolerance level for sub-optimality

Initialize $\pi_0^i \in \Pi^i$, $\widehat{Q}_0^i = 0 \in \mathbb{R}^{\mathbb{Y} \times \mathbb{U}}$, $\widehat{J}_0^i = 0 \in \mathbb{R}^{\mathbb{Y}}$

for $k \geq 0$ (k^{th} exploration phase)

  for $t = t_k, t_k + 1, \ldots, t_{k+1} - 1$

    Observe $y_t^i = \varphi^i(\mathbf{x}_t)$

    Select $u_t^i \sim \pi_k^i(\cdot | y_t^i)$

    Observe $c_t^i := c(x_t^i, \mu(\cdot | \mathbf{x}_t), u_t^i)$ and y_{t+1}^i

    Set $n_t^i = \sum_{\tau=t_k}^t \mathbf{1}\{(y_\tau^i, u_\tau^i) = (y_t^i, u_t^i)\}$

    Set $m_t^i = \sum_{\tau=t_k}^t \mathbf{1}\{y_\tau^i = y_t^i\}$

    $\widehat{Q}_{t+1}^i(y_t^i, u_t^i) = \left(1 - \frac{1}{n_t^i}\right) \widehat{Q}_t^i(y_t^i, u_t^i) + \frac{1}{n_t^i}\left[c_t^i + \beta \min_{a^i} \widehat{Q}_t^i(y_{t+1}^i, a^i)\right]$

    $\widehat{J}_{t+1}^i(y_t^i) = \left(1 - \frac{1}{m_t^i}\right) \widehat{J}_t^i(y_t^i) + \frac{1}{m_t^i}\left[c_t^i + \beta \widehat{J}_t^i(y_{t+1}^i)\right]$

  if $\widehat{J}_{t_{k+1}}^i(y) \leq \min_{a^i} \widehat{Q}_{t_{k+1}}^i(y, a^i) + \epsilon + d^i \ \forall y \in \mathbb{Y}$, **then**

    $\pi_{k+1}^i = \pi_k^i$

  else

    $\pi_{k+1}^i \sim (1 - e^i)\delta_{\pi_k^i} + e^i \mathrm{Unif}(\Pi^i)$)

  Reset $\widehat{J}_{t_{k+1}}^i = 0 \in \mathbb{R}^{\mathbb{Y}}$ and $\widehat{Q}_{t_{k+1}}^i = 0 \in \mathbb{R}^{\mathbb{Y} \times \mathbb{U}}$

A. Learning with Global State

We begin by presenting convergence results for Algorithm 2 under global state observability, the richest of the information structures that we consider. In order to state our first result, we now fix $\epsilon > 0$ and make the following assumptions on the various parameters of Algorithm 2.

Assumption 4: Fix $\epsilon > 0$ and for each $i \in \mathcal{N}$ let $\Pi_{\mathrm{MF}}^i := \{\pi^i \in \Gamma_S^i : \pi^i \text{ is of the mean-field type.}\}$. Assume that $\mathbf{\Pi} \subset \mathbf{\Pi}_{\mathrm{MF}}$ is a soft, symmetric quantization of $\mathbf{\Pi}_{\mathrm{MF}}$ satisfying $\mathbf{\Pi} \cap \mathbf{\Gamma}^{\epsilon\text{-eq}} \neq \varnothing$.

For each player $i \in \mathcal{N}$, the tolerance parameter d^i is taken to be positive to account for noise in the learned estimates, but cannot be too large, otherwise poorly performing policies may be mistaken for ϵ-best-responses. The bound $\bar{d}_G$ below is analogous to $\bar{\delta}$ in [27] and depends on both ϵ and $\mathbf{\Pi}$.

Assumption 5: For each player $i \in \mathcal{N}$, $d^i \in (0, \bar{d}_G)$, where $\bar{d}_G = \bar{d}_G(\epsilon, \mathbf{\Pi})$ is specified [14].

Theorem 5: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying Assumptions 1 and 3, and let $\epsilon > 0$. Suppose Assumptions 4 and 5 hold and all players follow Algorithm 2. For any $\xi > 0$, there exists $\tilde{T} = \tilde{T}(\xi, \epsilon, \mathbf{\Pi}, \{d^i\}_{i\in\mathcal{N}})$ such that if $T_k \geq \tilde{T}$ for all k, then

$$\Pr\left(\boldsymbol{\pi}_k \in \mathbf{\Pi} \cap \mathbf{\Gamma}^{\epsilon\text{-eq}}\right) \geq 1 - \xi,$$

for all sufficiently large k.

B. Learning with Mean-Field State Information

Assumption 6: Fix $\epsilon > 0$. Assume $\mathbf{\Pi}$ is a soft, symmetric quantization of $\mathbf{\Gamma}_S$ such that $\mathbf{\Pi} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$.

Assumption 7: For all $i \in \mathcal{N}$, $d^i \in (0, \bar{d}_{\mathrm{MF}})$, where $\bar{d}_{\mathrm{MF}} = \bar{d}_{\mathrm{MF}}(\epsilon, \mathbf{\Pi})$ is specified in [14].

Theorem 6: Let $\mathcal{G}$ be a partially observed N-player mean-field game satisfying Assumptions 2 and 3, and let $\epsilon > 0$. Suppose Assumptions 6 and 7 hold and all players follow Algorithm 2. For any $\xi > 0$, there exists $\tilde{T} = \tilde{T}(\xi, \epsilon, \mathbf{\Pi}, \{d^i\}_{i\in\mathcal{N}})$ such that if $T_k \geq \tilde{T}$ for all k, then

$$\Pr\left(\boldsymbol{\pi}_k \in \mathbf{\Pi} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)\right) \geq 1 - \xi,$$

for all sufficiently large k.

The proofs of Theorems 5 and 6 are given in [14]. In essence, one shows that if the exploration phases are long enough, the learning iterates approximate the subjective functions. Then, the process $\{\boldsymbol{\pi}_k\}_{k\geq 0}$ obtained from Algorithm 2 can be shown to approximate the policy process of a Markov chain on $\mathbf{\Pi}$ whose absorbing states are $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibria. Convergence to $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$ is then shown using the theory of satisficing paths.

VII. Conclusions

In this paper, we have studied independent learning in partially observed N-player mean-field games under two observation channels. We have studied the convergence of stochastic learning iterates used by independent learners and we have presented a notion of subjective equilibrium suitable for analyzing independent learners. Using this notion of subjective equilibrium, we presented results on the existence of subjective ϵ-equilibrium, and we have observed useful structure pertaining to dynamical systems on the set of policies. Exploiting this structure, we have presented a decentralized, independent learning algorithm for playing partially observed N-player mean-field games. Under self-play, this algorithm drives policies to subjective equilibrium.

References

[1] M. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle," *Communications in Information & Systems*, vol. 6, no. 3, pp. 221–252, 2006.

[2] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[3] M. Fischer, "On the connection between symmetric n-player games and mean field games," *The Annals of Applied Probability*, vol. 27, no. 2, pp. 757–810, 2017.

[4] N. Saldi, T. Başar, and M. Raginsky, "Markov–Nash equilibria in mean-field games with discounted cost," *SIAM Journal on Control and Optimization*, vol. 56, no. 6, pp. 4256–4287, 2018.

[5] S. Sanjari, N. Saldi, and S. Yüksel, "Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers," *Mathematics of Operations Research*, 2022.

[6] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994*, pp. 157–163, Elsevier, 1994.

[7] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5527–5540, 2020.

[8] M. Sayin, K. Zhang, D. Leslie, T. Başar, and A. Ozdaglar, "Decentralized Q-learning in zero-sum Markov games," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18320–18334, 2021.

[9] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.

[10] B. Yongacoglu, G. Arslan, and S. Yüksel, "Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information," *IEEE Transactions on Automatic Control*, to appear.

[11] M. L. Littman and C. Szepesvári, "A generalized reinforcement-learning model: Convergence and applications," in *ICML*, vol. 96, pp. 310–318, Citeseer, 1996.

[12] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems.," *Knowledge Engineering Review*, vol. 27, no. 1, pp. 1–31, 2012.

[13] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, pp. 746–752, 1998.

[14] B. Yongacoglu, G. Arslan, and S. Yüksel, "Independent learning in mean-field games: Satisficing paths and convergence to subjective equilibria," *arXiv preprint arXiv:2209.05703*, 2022.

[15] A. C. Kizilkale and P. E. Caines, "Mean field stochastic adaptive control," *IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 905–920, 2012.

[16] H. Yin, P. G. Mehta, S. P. Meyn, and U. V. Shanbhag, "Learning in mean-field games," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 629–644, 2013.

[17] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International Conference on Machine Learning*, pp. 5571–5580, PMLR, 2018.

[18] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games," in *Proceedings of the 18th International Conference on AAMAS*, pp. 251–259, 2019.

[20] R. Elie, J. Perolat, M. Laurière, M. Geist, and O. Pietquin, "On the convergence of model free learning in mean field games," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7143–7150, 2020.

[21] Q. Xie, Z. Yang, Z. Wang, and A. Minca, "Learning while playing in mean-field games: Convergence and optimality," in *International Conference on Machine Learning*, pp. 11436–11447, PMLR, 2021.

[22] M. A. uz Zaman, K. Zhang, E. Miehling, and T. Başar, "Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2278–2284, IEEE, 2020.

[23] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Learning in discrete-time average-cost mean-field games," in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3048–3053, IEEE, 2021.

[24] A. M. Fink, "Equilibrium in a stochastic n-person game," *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.

[25] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," *arXiv preprint arXiv:1707.09183*, 2017.

[26] A. D. Kara and S. Yüksel, "Convergence of finite memory q-learning for pomdps and near optimality of learned policies under filter stability," *arXiv preprint arXiv:2103.12158*, 2021.

[27] B. Yongacoglu, G. Arslan, and S. Yüksel, "Satisficing paths and independent multi-agent reinforcement learning in stochastic games," *arXiv preprint arXiv:2110.04638*, 2022.