Journal of Machine Learning Research 25 (2024) 1-69 Submitted 10/22; Revised 11/23; Published 10/24

Mean-Field Games With Finitely Many Players: Independent Learning and Subjectivity

Bora Yongacoglu BORA.YONGACOGLU@UTORONTO.CA
Department of Electrical and Computer Engineering
University of Toronto

Gürdal Arslan GURDAL@HAWAII.EDU
Department of Electrical Engineering
University of Hawaii at Manoa

Serdar Yüksel YUKSEL@QUEENSU.CA
Department of Mathematics and Statistics
Queen's University

Editor: Csaba Szepesvari

Abstract

Independent learners are agents that employ single-agent algorithms in multi-agent systems, intentionally ignoring the effect of other strategic agents. This paper studies mean-field games from a decentralized learning perspective with two aims: (i) to identify structure that can guide algorithm design, and (ii) to understand emergent behaviour in systems of independent learners. We study a new model of partially observed mean-field games with finitely many players, local action observability, and partial observations of the global state. Specific observation channels considered include (a) global observability, (b) mean-field observability, (c) compressed mean-field observability, and (d) only local observability. We establish conditions under which the control problem of a given agent is equivalent to a fully observed MDP, as well as conditions under which the control problem is equivalent only to a POMDP. Using the connection to MDPs, we prove the existence of perfect equilibrium among memoryless stationary policies under mean-field observability. Using the connection to POMDPs, we prove convergence of learning iterates obtained by independent learners under any of our observation channels. We interpret the limiting values as subjective value functions, which an agent believes to be relevant to its control problem. These subjective value functions are used to propose subjective Q-equilibrium, a new solution concept whose existence is proved under mean-field or global observability. Furthermore, we provide a decentralized independent learning algorithm, and by adapting the recently developed theory of satisficing paths to allow for subjectivity, we prove that it drives play to subjective Q-equilibrium. Our algorithm is decentralized, in that it uses only local information for learning and allows players to use different, heterogeneous policies during play. As such, it departs from the conventional representative agent approach common to other algorithms for mean-field games.

Keywords: Multi-agent reinforcement learning, independent learners, learning in games, mean-field games, decentralized systems

1. Introduction

Mean-field games are a theoretical framework for studying strategic environments with a large number of weakly coupled agents (Huang et al., 2006, 2007; Lasry and Lions, 2007). In a mean-field game, the cost and state dynamics of any particular agent are influenced by the collective behaviour of others only through a distributional *mean-field* term. Due to the ubiquity of large-scale decentralized systems in modern engineering, mean-field games have been used to model a diverse range of applications, such as resource management (Bauso et al., 2012), social conventions (Gomes et al., 2014), traffic control (Chevalier et al., 2015), and opinion dynamics (Stella et al., 2013), among many others.

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 (Zhang et al., 2021). 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. While these efforts have led to a number of empirically successful algorithms, there are fewer works that offer formal convergence analyses of their algorithms, and the bulk of existing work is suitable only for systems with a small number of agents.

There are several competing paradigms for studying multi-agent learning. The rational learning paradigm, popularized by (Kalai and Lehrer, 1993), studies Bayesian agents who maintain beliefs on the strategies (also called policies) used by other agents. Learners select policies to best-respond to their beliefs, while beliefs are updated as new information arrives during the course of play. In this framework, under conditions aligning beliefs with the actual policies used, several results on the convergence of play to equilibrium were proved. However, this framework has two major shortcomings. First, the paradigm does not lend itself to tractable algorithms, as players must maintain beliefs over the strategies used by every other agent, and these strategies can be arbitrarily complex and history-dependent. The second shortcoming of rational learning has to do with the restrictive conditions under which its results hold (Nachbar, 2005).

One desirable feature of the Bayesian framework is its explicit modelling of beliefs. Subjectivity and beliefs are relevant in MARL as they allow for players to be uncertain of the model of the game being played. In (Kalai and Lehrer, 1995), the authors proposed a definition for subjective equilibrium in which players behave optimally with respect to their subjective beliefs, while each player's subjective beliefs satisfy a consistency condition with respect to the observed trajectory of play. Competing notions of subjective equilibrium have been proposed, some of which are structural and non-Bayesian: agents believe their system has a particular structure, and agents act in accordance with their assumptions. We refer the reader to (Arslan and Yüksel, 2023) for a review. In this paper, we propose *subjective Q-equilibrium,* a new non-Bayesian notion of subjective equilibrium that is well-suited to decentralized learning in large-scale systems, where agents suffer from a high degree of uncertainty and may need to make structural assumptions for tractability. This equilibrium definition contributes to an emerging line of work that studies simple agents in complex, possibly non-Markovian environments with a perceived local state (Dong et al., 2022; Chandak et al., 2024; Kara and Yüksel, 2022; Kara and Yüksel, 2024).

Outside of the Bayesian learning paradigm, the majority of theoretical contributions in MARL have focused on highly structured classes of stochastic games, such as two-player zero-sum games (Sayin et al., 2021, 2022) and n-player stochastic teams and their generalizations (Ding et al., 2022; Fox et al., 2022; Leonardos et al., 2022; Mguni et al., 2021; Unlu and Sayin, 2023; Zhang et al., 2022). In much of the existing literature on MARL, a great deal of information is assumed to be available to agents while they learn. These assumptions, such as full state observability (as assumed by (Daskalakis et al., 2020) among others) or action-sharing among all agents (as assumed by (Littman and Szepesvári, 1996; Hu and Wellman, 2003) and (Littman, 2001)), are appropriate in some settings but are unrealistic in many large-scale, decentralized systems. One issue with designing MARL algorithms that use global information about the local states and actions of all players is that such algorithms do not scale with the number of players. The so-called *curse of many agents* is a widely cited challenge to MARL (Wang et al., 2020).

Independent learners (Matignon et al., 2009, 2012) are a class of MARL algorithms that are characterized by intentional obliviousness to the strategic environment: independent learners ignore the presence of other players, effectively treating other players as part of the environment rather than as non-stationary learning agents. By naively running a single-agent reinforcement learning algorithm using only local information, independent learners are relieved of the burden of excessive information, which may lead to scalable algorithms for large-scale decentralized systems. However, additional care must be taken when designing independent learners, as direct application of single-agent reinforcement learning has had mixed success in small empirical studies (Condon, 1990; Tan, 1993; Sen et al., 1994; Claus and Boutilier, 1998). Here, we offer a new perspective on independent learners by framing their obliviousness as a subjective structural belief on the system within which they exist. Accordingly, agents may act rationally with respect to their (possibly incorrect) subjective beliefs. Building on this interpretation, we propose a notion of subjective equilibrium that is inspired by, and well-suited to, the analysis of independent learners in decentralized MARL. Structural properties of such subjective equilibria, such as their (non-) existence and their comparison to traditional solution concepts, may explain why independent learners lead to stable outcomes in certain MARL applications and not others.

The Learning Setting

We study independent learners in partially observed n-player mean-field games. Our model is closely related to that of (Saldi et al., 2018), but allows for a variety of observation channels through which agents can observe non-local information. Decentralization and learning are our primary focuses, and we are interested in online algorithms that involve minimal coordination between agents. Some salient characteristics of our model are the following:

Limited View of the Global State: Using a single, unified model, we consider four alternative observation channels by which players observe non-local state information.

Local Action Information: A given player does not know the policies used by the remaining players, and the player does not observe the actions generated by these policies.

Decentralized, Online Feedback: In this study, players to not have access to the model governing the system, nor do they have access to a simulator/oracle/generative model for sampling data in an arbitrary order, nor do they have access to offline data sets that can be used for training. Instead, we assume that players view only the stream of data encountered during play, which is not shared across agents.

Decentralized Training: This work does not belong to the paradigm of *centralized training with decentralized execution* (Lowe et al., 2017; Foerster et al., 2018), in which a global value function is learned during training and players select policies that can be implemented in a decentralized manner informed by this global quantity.

Heterogeneous Policies: Since agents learn using different local information encountered during play, we avoid the common requirement that all agents employ the same policy at any given time.

Contributions:

1) In Section 2, we propose a new model of partially observed n-player mean-field games. This model is suitable for the study of decentralized learning and is flexible enough to model various alternative observation channels. In particular, each agent observes its own local actions and partially observes the global state. Observation channels considered include (a) global state observability, (b) (local state and) mean-field observability (c) compressed mean-field observability, and (d) only local state observability.

2) In Section 3, we show that if the remaining players follow (memoryless) stationary policies, then an individual faces a control problem equivalent to a POMDP (Lemma 6). Moreover, under mean-field observability and a symmetry condition on the policies of other players, Theorem 8 states that this control problem is equivalent to a fully observed MDP. We also show that the MDP equivalence does not hold in general when one relaxes either mean-field observability or symmetry of policies. This non-equivalence to an MDP is an important distinction between the model studied here and the limit model with infinitely many agents. In Theorem 12, for the case of mean-field observability, we prove the existence of perfect equilibrium in the set of (memoryless) stationary policies. This result builds on Theorem 8, and is of independent interest.

3) We study learning iterates obtained by independent learners in a partially observed n-player mean-field game. In Theorem 13, we show that when each agent uses a (memoryless) stationary policy and naively runs single-agent value estimation algorithms (Algorithm 1), its learning iterates converge almost surely under mild assumptions. In Appendix B, we characterize these almost sure limits in terms of underlying, implied MDPs. We offer a new perspective on independent learning, interpreting it as a subjective model adopted by the learning agent. The limiting values of the agent's learning iterates are then interpreted as subjective value functions, in analogy to objective value functions in single-agent MDPs. Then, we define subjective ϵ-best-responding in Definition 15 and we use this to define subjective ϵ-equilibrium (also called subjective Q-equilibrium) in Definition 16. In Lemma 17, we show that subjective ϵ-equilibrium exists under mean-field observability.

4) We present Algorithm 3, a decentralized independent learning algorithm for partially observed n-player mean-field games. We use a modified theory of satisficing paths to show

that Algorithm 3 drives play to subjective ϵ-equilibrium under mean-field observability (Theorem 25). Analogous results for other observation channels can be found in Appendix J.

Organization

In §1.1 and §1.2, we survey related literature. The formal model is presented in Section 2. In Section 3, we examine the connection between partially observed n-player mean-field games and partially/fully observed Markov decision problems (POMDP/MDPs), and we prove several structural results. Independent learners and subjectivity are discussed in Section 4, where we study the convergence and interpretation of learning iterates under partial observability. We subsequently define our notion of subjective equilibrium and state an existence result under mean-field observability. Algorithmic results, including a decentralized learning algorithm (Algorithm 3) and its analysis, are given in Section 5. Results of a simulation study are presented in Section 6. Proofs, background material, and further results can be found in the expansive appendix sections, including an additional result on objective ϵ-equilibrium under local observability that is presented in Appendix H.

Notation: We use Pr to denote a probability measure on some underlying probability space, with additional superscripts and subscripts as needed. For a finite set S, we let $\mathbb{R}^S$ denote the real vector space of dimension $|S|$ where vector components are indexed by elements of S. We let $0 \in \mathbb{R}^S$ denote the zero vector of $\mathbb{R}^S$ and $1 \in \mathbb{R}^S$ denote the vector in $\mathbb{R}^S$ for which each component is 1. For standard Borel sets S, S', we let $\Delta(S)$ denote the set of probability measures on S with the Borel σ-algebra on S, and we let $\mathscr{P}(S'|S)$ denote the set of transition kernels on S' given S. We use $Y \sim \mu$ to denote that the random variable Y has distribution μ. For an event $\{\cdot\}$, we let $\mathbf{1}\{\cdot\}$ denote the indicator function of the event's occurrence. If a probability distribution μ is a mixture of distributions $\mu_1, \ldots, \mu_n$ with mixture weights $p_1, \ldots, p_n$, we write $Y \sim \sum_{i=1}^n p_i \mu_i$. For $s \in S$, we use $\delta_s \in \Delta(S)$ to denote the Dirac measure centered at s.

1.1 Related Work

A number of research items attempt to approximate solution concepts for mean-field games using learning algorithms. By and large, these works approach learning in mean-field games by analyzing the single-agent problem for a representative agent in the limiting case as $n \to \infty$, and equilibrium is defined using a best-responding condition as well as a consistency condition; see, for instance, (Subramanian and Mahajan, 2019, Definition 2.1). The solution concepts sought in these works are inherently symmetric: at equilibrium, all agents use the same policy as the representative agent. As a result, such equilibria may be difficult to learn in decentralized learning settings, where agents will use heterogeneous policies during the learning process. We now describe a representative sample of works in this centralized learning tradition, and we refer the reader to (Laurière et al., 2022) for a survey.

In (Guo et al., 2019), existence and uniqueness of mean-field equilibrium is shown under stringent assumptions, and an algorithm is proposed and analyzed under an assumption that a generative model for sampling transition and cost data is available.

Stationary mean-field games are studied in (Subramanian and Mahajan, 2019). Optimality and equilibrium are defined in terms of the limiting invariant distribution of the

mean-field state. The authors present multiple notions of equilibrium and study two-timescale policy gradient methods, where a representative agent updates its policy to best respond to its (estimated) environment on the fast timescale and updates its estimate for the mean-field flow on the slow timescale. With access to a simulator for obtaining data, convergence to a local notion of equilibrium is proved.

A variant of fictitious play for mean-field games is proposed in (Elie et al., 2020), though the question of learning to best-respond is black-boxed. Other algorithms based on fictitious play have also been proposed in (Mguni et al., 2018) and (Xie et al., 2021).

A value iteration algorithm for computing stationary mean-field equilibrium in both discounted and average-cost mean-field games is presented in (Anahtarci et al., 2020) and extended in (Anahtarci et al., 2023a). Regularized mean-field games are studied and a learning algorithm based on fitted Q-learning is proposed in (Anahtarci et al., 2023b). Algorithms and operators based on regularization are also studied in (Cui and Koeppl, 2021). The preceding analytical works are closely related to the two-timescale algorithm of (Zaman et al., 2023). A different two-timescale algorithm is considered in (Angiuli et al., 2022).

In the preceding works, the authors largely restrict their attention to Markov policies that depend only on local state observations, without dependence on the mean-field distribution. At least two research items have considered learning with a focus on *population-dependent policies*, which condition one's action choice on both the local state and the mean-field distribution. In this vein, the existence of so-called master policies is studied and a learning algorithm for their computation is proposed in (Perrin et al., 2022). An algorithm for learning mean-field equilibrium in finite horizon games is considered in (Mishra et al., 2023), using results from (Vasal, 2023).

Most of the works cited above are *centralized* methods for a decentralized system, as they consider a population of agents using symmetric policies at all times. As a result, the problems studied are single-agent rather than multi-agent in flavour, due to the lack of strategic interaction in the mean-field limit. The principal aim of these papers is to use learning techniques to compute a (near) equilibrium for the mean-field games they study. By contrast, this paper aims to understand patterns of behaviour that emerge when agents use reasonable (if limited) learning algorithms in a shared environment. Our focus, then, is less computational and more descriptive in nature.

In many realistic settings, agents may employ different learning algorithms for a variety of reasons, such as differing prior beliefs. Moreover, since distinct agents observe distinct local observation histories and feed these local observation histories to their learning algorithms, distinct agents may use radically different policies over the course of learning. Work in the computational tradition largely avoids such learning dynamics, and therefore does not encounter plausible equilibrium policies consisting of various heterogeneous policies used by a population of homogeneous players. In this paper, we depart from the traditional approach of mandating all agents to follow the same policy during learning.

The learning environment studied in (Yardim et al., 2023) is perhaps the closest to the one considered in this paper. (Yardim et al., 2023) study a decentralized learning problem in an n-player symmetric anonymous game with local observability and propose a mirror descent algorithm. Unlike the papers described earlier, the n agents are allowed to use het-

erogeneous policies during the course of learning. Under Lipschitz continuity assumptions on various operators, the authors show convergence to equilibrium in the regularized game.

The primary objective of the literature surveyed above is to provide algorithms for the approximation of conventional variants of mean-field equilibrium. These works consider a variety of models for system interaction and a variety of objective criteria and time horizons. To facilitate the design of such algorithms, many of these works conduct structural analysis on mean-field games, typically in the limit with infinitely many players, and almost always under symmetry conditions mandating that the entire population uses the policy of the representative agent.

This paper, too, makes contributions on the structure and analysis of mean-field games, focusing on the case with finitely many players and obtaining clear insights both with and without symmetry in policies. This is illustrated by the results and examples in Section 3. On the other hand, unlike the papers surveyed above, the goal of this paper is not the approximation of mean-field equilibria *per se*. Instead, the goal here is to better understand the outcomes of decentralized learning processes in large-scale systems. Since players use heterogeneous, asymmetric policies during the learning process and iteratively adjust their behaviour as new information arrives, we do not restrict ourselves to the symmetric solution concepts of mean-field equilibrium. In Section 4, we propose a novel solution concept that is well-suited for the study of decentralized learners. In considering alternative solution concepts for mean-field games, this paper is also related to the works (Lee et al., 2021; Subramanian et al., 2022b; Muller et al., 2022; Campi and Fischer, 2022).

In (Lee et al., 2021), mean-field games with strategic complementarities are studied and a trembling hand equilibrium concept is proposed. The authors of (Subramanian et al., 2022b) propose a definition of decentralized mean-field equilibrium that attempts to capture heterogeneous policies while preserving the overall consistency-optimality form of conventional mean-field equilibrium. Mean-field correlated (and coarse correlated) equilibrium are proposed and studied in (Muller et al., 2022). A related solution concept called the *correlated solution* of a mean-field game is proposed by (Campi and Fischer, 2022).

1.2 Applications of Mean-Field Games

We now describe some relevant recent applications. For a survey of some classical applications of mean-field games, see (Gomes and Saúde, 2014).

A peer-to-peer energy market is modelled as a mean-field game in (Xia et al., 2019). The mean-field quality of the market owes to the large population size and the random matching of consumers and producers, according to which a player's best response depends on the distribution of other agents' budgets (states).

In (Li et al., 2019), vehicle dispatch in a ride-sharing application is modelled as independent learning in a mean-field game. The approach used there is well-aligned with the present paper: players observe an aggregate environmental state (describing weather conditions, traffic congestion, and so on) as well as local information about rides in their vicinity, and use this information to guide their action choice. The authors propose an independent MARL algorithm based on Q-learning, which manipulates an object resembling a Q-function: for agent i, they consider a quantity denoted $Q_i(o_i, a_i)$, which attempts to assign a value to playing action a_i after observing the symbol o_i. Strictly speaking, this

object is not a Q-function, since the agent does not face an MDP. Nevertheless, the authors observe that this approach leads to tractable, decentralized algorithms that outperform their centralized predecessors.

Traffic routing is often modelled as a mean-field game, where overall travel time depends on the distribution of agents through the traffic network. In (Salhab et al., 2018), traffic routing is modelled as a mean-field game with mean-field observability, where availability of the mean-field is justified through the use of a mobile application. In a related study, (Cabannes et al., 2022), traffic routing is modelled as a mean-field game with only local observability.

Another application of mean-field games involves resource allocation when a large number of agents compete for limited resources, often through auctions. Examples include service in a cellular network (Manjrekar et al., 2019), auctions for advertisements (Iyer et al., 2014), resource distribution in smart grids (Li et al., 2018), and device-to-device wireless transmission (Li et al., 2016).

2. Model

In this section, we present our model of partially observed n-player mean-field games, where several strategic agents interact in a shared environment. In subsequent sections, we will investigate the deep connection between our model of mean-field games and the single-agent model of partially observed Markov decision problems (POMDPs). Background material on POMDPs and fully observed MDPs can be found in Appendix A.

2.1 Partially Observed Mean-Field Games

For $n \in \mathbb{N}$, a partially observed n-player mean-field game is described by a list $\mathbf{G}$:

$$\mathbf{G} = (\mathcal{n}, \mathbb{X}, \mathbb{Y}, \mathbb{A}, \{\varphi^i\}_{i\in\mathcal{n}}, c, \gamma, P_{\text{loc}}, \nu_0). \tag{1}$$

The game $\mathbf{G}$ consists of n players/agents, indexed by $\mathcal{n} = \{1, \dots, n\}$. At time $t \in \mathbb{Z}_{\geq 0}$, the local state for player $i \in \mathcal{n}$ is denoted by $x_t^i \in \mathbb{X}$, where $\mathbb{X}$ is a finite set of *local states*. We let $\mathbf{X} := \times_{i\in\mathcal{n}}\mathbb{X}$ denote the n-fold Cartesian product of $\mathbb{X}$, indexed by $\mathcal{n}$. An element of $\mathbf{X}$ is called a *global state*, and the global state at time t is denoted by $\mathbf{x}_t = (x_t^i)_{i\in\mathcal{n}}$. The *mean-field* associated with global state $\mathbf{x}_t$ is denoted by $\mu_t \in \Delta(\mathbb{X})$ and defined as the empirical measure of $\mathbf{x}_t$:

$$\mu_t(B) := \frac{1}{n}\sum_{i\in\mathcal{n}} \delta_{x_t^i}(B), \quad \forall B \subseteq \mathbb{X}.$$

Each player $i \in \mathcal{n}$ observes the global state $\mathbf{x}_t$ indirectly, making the local observation $y_t^i = \varphi^i(\mathbf{x}_t)$, where $\mathbb{Y}$ is a finite set and $\varphi^i : \mathbf{X} \to \mathbb{Y}$ is player i's observation function. Each player i uses its locally observable history variable $h_t^i \in H_t$, to be defined shortly, to select an action $a_t^i \in \mathbb{A}$, where $\mathbb{A}$ is a finite set of actions. We denote the *joint action* of all players by $\mathbf{a}_t = (a_t^i)_{i\in\mathcal{n}}$, and note that player i only observes its own action a_t^i. After choosing action a_t^i, player i incurs a stage cost of $c_t^i = c(x_t^i, \mu_t, a_t^i)$, where $c : \mathbb{X}\times\Delta(\mathbb{X})\times\mathbb{A} \to \mathbb{R}$. Player i's local state then evolves according to $x_{t+1}^i \sim P_{\text{loc}}(\cdot|x_t^i, \mu_t, a_t^i)$, where $P_{\text{loc}} \in \mathscr{P}(\mathbb{X}|\mathbb{X}\times\Delta(\mathbb{X})\times\mathbb{A})$. Each player discounts its stream of costs using discount factor $\gamma \in (0, 1)$, and the initial global state $\mathbf{x}_0$ has distribution ν_0. The process repeats at time $t + 1$ and so on.

Observation Channels and Local Information

In what follows, we refer to the pair $(\mathbb{Y}, \{\varphi^i\}_{i\in n})$ as the *observation channel* for the game $\mathbf{G}$. To simplify the statement of assumptions on the observation channel, we define a function $\mu : \mathbf{X} \to \Delta(\mathbb{X})$ that maps a global state to its empirical distribution:

$$(\mu(\mathbf{s}))(B) := \frac{1}{n}\sum_{i\in n} \delta_{s^i}(B), \quad B \subseteq \mathbb{X}.$$

With this notation, we have the mean-field state $\mu_t = \mu(\mathbf{x}_t)$ in the description of play above. We also define a finite subset $\mathrm{Emp}_n \subset \Delta(\mathbb{X})$ as $\mathrm{Emp}_n := \{\mu(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$.

In order to cover a diverse range of large, decentralized learning environments with a unified model, we now state some alternative observation channels that may be assumed.

Assumption 1 (Global 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 Observability) $\mathbb{Y} = \mathbb{X} \times \mathrm{Emp}_n$ *and* $\varphi^i(\mathbf{s}) = (s^i, \mu(\mathbf{s}))$ *for each global state* $\mathbf{s} \in \mathbf{X}$ *and player* $i \in \mathcal{n}$.

Assumption 3 (Compressed Observability) *For* $k \in \mathbb{N}$, *let* $[k] := \{1, 2, \ldots, k\}$ *and let* $f : \Delta(\mathbb{X}) \to [k]$. *Then,* $\mathbb{Y} = \mathbb{X} \times [k]$ *and for each* $i \in \mathcal{n}$, $\mathbf{s} \in \mathbf{X}$, *we have* $\varphi^i(\mathbf{s}) = (s^i, f(\mu(\mathbf{s}))$.

Assumption 4 (Local Observability) $\mathbb{Y} = \mathbb{X}$ *and* $\varphi^i(\mathbf{s}) = s^i$ *for each* $i \in \mathcal{n}, \mathbf{s} \in \mathbf{X}$.

Assumptions 1–4 are presented in order of increasing decentralization. In practice, the particular choice of informational assumption will depend on one's application area: in some instances, there will be a natural restriction of information leading to a particular observation channel. In other instances, agents may voluntarily compress a more informative observation variable for the purposes of function approximation. We now briefly describe these alternative observation channel assumptions. Additional discussion on this topic is provided in Section 7.

In decentralized systems, global observability is often unrealistic, but this case is included in our discussion as it is helpful for developing insights. This observation channel has previously been used in some works on linear-quadratic mean-field games (Zaman et al., 2020a,b). Mean-field observability is perhaps the most commonly considered observation channel in works on mean-field games: see (Saldi et al., 2018; Zaman et al., 2023) and the references therein. Local observability is also a commonly adopted assumption (see, for instance, (Muller et al., 2022; Yardim et al., 2023) and the references therein).

Compressed observability is an intermediate setting between mean-field and local observability. To our knowledge, this has not been explicitly considered in prior theoretical work. Compressed observability can be motivated using the discussion above: it serves to lessen the computational burden at a given learning agent in a partially observed n-player mean-field game and, as we discuss in Section 7, may be a more appropriate modelling assumption in some applications. By taking $k = 1$, we see that local state observability

is in fact a special case of compressed state observability, where the compressed information about the mean-field state is uninformative. We include Assumption 4 separately to highlight the importance of this set-up.

Where possible, we conduct our analysis in a unified manner, without specifying the observation channel. In some instances, stronger results can be proved under richer observation channels. In such cases, our exposition and analysis centres on mean-field observability, with discussion on other information structures either omitted or postponed.

Local Histories and Policies

We now formalize the action selection process by defining policies. We make distinctions between the overall system history and each player's locally observable histories. For any $t \in \mathbb{Z}_{\geq 0}$, we define the sets

$$\begin{aligned} \mathbf{H}_t &:= (\mathbf{X} \times \mathbf{A})^t \times \mathbf{X}, \\ H_t &:= \Delta(\mathbf{X}) \times (\mathbb{Y} \times \mathbb{A} \times \mathbb{R})^t \times \mathbb{Y}. \end{aligned}$$

For any $t \geq 0$, the set $\mathbf{H}_t$ represents 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 in the game $\mathbf{G}$ may observe. Elements of $\mathbf{H}_t$ are called *system histories of length* t, and elements of H_t are called *observable histories of length* t.

For each $t \geq 0$ and player $i \in \mathcal{n}$, we let

$$\begin{aligned} \mathbf{h}_t &:= (\mathbf{x}_0, \mathbf{a}_0, \cdots, \mathbf{x}_{t-1}, \mathbf{a}_{t-1}, \mathbf{x}_t), \\ h_t^i &:= (\nu_0, y_0^i, a_0^i, c_0^i, \cdots, y_{t-1}^i, a_{t-1}^i, c_{t-1}^i, y_t^i) \end{aligned}$$

denote the t^{th} *system history variable* and player i's t^{th} *locally observable history variable*, respectively. Note that $\mathbf{h}_t$ is a random quantity taking values in $\mathbf{H}_t$, while h_t^i is a random quantity taking values in H_t.

Definition 1 *A sequence $\pi^i = (\pi_t^i)_{t \geq 0}$ with $\pi_t^i \in \mathcal{P}(\mathbb{A}|H_t)$ for every $t \geq 0$ is called a* policy *for player i. We let Π^i denote the set of all policies for player i.*

Definition 2 *A policy $\pi^i \in \Pi^i$ is called* (memoryless) stationary (or simply stationary) *if, for some $f^i \in \mathcal{P}(\mathbb{A}|\mathbb{Y})$, the following holds: 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 Π_S^i denote the set of stationary policies for player i.*

Remark: The set of policies – and thus learning algorithms – available to an agent depends on the set of locally observable histories, which itself depends on the observation channel $(\mathbb{Y}, \{\varphi^i\}_{i \in \mathcal{n}})$. In this paper, our focus is on *independent learners*, which are learners that do not use the joint action information in their learning algorithms. Here, we have chosen to incorporate this constraint into the information structure. Moreover, to underscore the importance of learning in our study, we also do not assume that the players know the cost function c. Instead, we assume only that they receive feedback costs in response to particular system interactions. These assumptions on the information structure resemble those of other work on independent learners (Daskalakis et al., 2020; Sayin et al., 2021;

Yongacoglu et al., 2022; Matignon et al., 2009, 2012; Wei and Luke, 2016; Yongacoglu et al., 2023). A rather different information structure is that of *joint action learners*, studied by (Claus and Boutilier, 1998) among others, where the locally observable history variables also include the joint action history.

Notation: We let $\mathbf{\Pi} := \times_{i\in\mathcal{I}}\Pi^i$ denote the set of *joint policies*. To isolate player i's component in a particular joint policy $\boldsymbol{\pi} \in \mathbf{\Pi}$, 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 $\mathbf{\Pi} = \Pi^i \times \mathbf{\Pi}^{-i}$, a joint action may be written as $\mathbf{a} = (a^i, \mathbf{a}^{-i}) \in \mathbf{A} := \mathbb{A}^{\mathcal{I}}$, and so on.

Value Functions, Best-Responding, and Objective Equilibrium

For each player $i \in \mathcal{I}$, we identify the set of stationary policies Π^i_S with the set $\mathscr{P}(\mathbb{A}|\mathbb{Y})$ of transition kernels on $\mathbb{A}$ given $\mathbb{Y}$. When convenient, a stationary policy $\pi^i \in \Pi^i_S$ is treated as if it were an element of $\mathscr{P}(\mathbb{A}|\mathbb{Y})$, and reference to the locally observable history variable is omitted. For each $i \in \mathcal{I}$, we introduce the metric d^i on Π^i_S, defined by

$$d^i(\pi^i, \tilde{\pi}^i) := \max\{|\pi^i(a^i|y) - \tilde{\pi}^i(a^i|y)| : y \in \mathbb{Y}, a^i \in \mathbb{A}\}, \quad \forall \pi^i, \tilde{\pi}^i \in \Pi^i_S.$$

We metrize the set of stationary joint policies $\mathbf{\Pi}_S$ with a metric $\mathbf{d}$, defined as

$$\mathbf{d}(\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}}) := \max_{i\in\mathcal{I}} d^i(\pi^i, \tilde{\pi}^i), \quad \forall \boldsymbol{\pi}, \tilde{\boldsymbol{\pi}} \in \mathbf{\Pi}_S.$$

A metric $\mathbf{d}^{-i}$ for the set $\mathbf{\Pi}^{-i}_S$ is defined analogously to $\mathbf{d}$. We have that the sets $\{\Pi^i_S\}_{i\in\mathcal{I}}$, $\{\mathbf{\Pi}^{-i}_S\}_{i\in\mathcal{I}}$ and $\mathbf{\Pi}_S$ are all compact in the topologies induced by the corresponding metrics.

For any joint policy $\boldsymbol{\pi} = (\pi^i)_{i\in\mathcal{I}}$ and initial distribution $\nu \in \Delta(\mathbf{X})$, there exists a unique probability measure $\Pr^{\boldsymbol{\pi}}_\nu$ on trajectories in $(\mathbf{X}\times\mathbf{A})^\infty$ satisfying the following, for all $t \geq 0$: (i) $\Pr^{\boldsymbol{\pi}}_\nu(\mathbf{x}_0 \in \cdot) = \nu(\cdot)$. (ii) For any $i \in \mathcal{I}$, $\Pr^{\boldsymbol{\pi}}_\nu\left(a^i_t \in \cdot \middle| h^i_t\right) = \pi^i_t\left(\cdot|h^i_t\right)$. (iii) The collection $\{a^j_t\}_{j\in\mathcal{I}}$ is jointly independent given $\mathbf{h}_t$. (iv) For any $i \in \mathcal{I}$ and $t \geq 0$, local states evolve according to $\Pr^{\boldsymbol{\pi}}_\nu\left(x^i_{t+1} \in \cdot \middle| \mathbf{h}_t, \mathbf{a}_t\right) = P_{\text{loc}}\left(\cdot \middle| x^i_t, \mu_t, a^i_t\right)$. (v) The collection $\{x^j_{t+1}\}_{j\in\mathcal{I}}$ is jointly independent given $(\mathbf{h}_t, \mathbf{a}_t)$.

We let $E^{\boldsymbol{\pi}}_\nu$ denote the expectation associated with $\Pr^{\boldsymbol{\pi}}_\nu$ and we use it to define player i's objective function, also called the (state) value function:

$$J^i(\boldsymbol{\pi}, \nu) := E^{\boldsymbol{\pi}}_\nu\left[\sum_{t=0}^\infty \gamma^t c^i_t\right] = E^{\boldsymbol{\pi}}_\nu\left[\sum_{t=0}^\infty \gamma^t c(x^i_t, \mu_t, a^i_t)\right].$$

Lemma 3 *For any initial measure $\nu \in \Delta(\mathbf{X})$ and any player $i \in \mathcal{I}$, the mapping $\boldsymbol{\pi} \mapsto J^i(\boldsymbol{\pi}, \nu)$ is continuous on $\mathbf{\Pi}_S$.*

The proof is omitted, as it resembles that of (Yongacoglu et al., 2023, Lemma 2.10).

From the final expression in the definition of $J^i(\boldsymbol{\pi}, \nu)$, one can see that player i's objective is only weakly coupled with the rest of the system: player i's costs depend on the global state and joint action sequences $\{\mathbf{x}_t, \mathbf{a}_t\}_{t\geq 0}$ only through player i's components $\{x^i_t, a^i_t\}_{t\geq 0}$, the mean-field state sequence $\{\mu_t\}_{t\geq 0}$, and the subsequent influence that $\{\mu_t\}_{t\geq 0}$ has on the evolution of $\{x^i_t\}_{t\geq 0}$. Nevertheless, player i's objective function does depend on the policies of the remaining players. This motivates the following definitions.

Definition 4 *Let $\epsilon \geq 0$, $\nu \in \Delta(\mathbf{X})$, $i \in \mathcal{N}$, and $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}^{-i}$. A policy $\pi^{*i} \in \Pi^i$ is called an* ϵ-best-response to $\boldsymbol{\pi}^{-i}$ with respect to ν *if*

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

For $i \in \mathcal{N}$, $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}^{-i}$, $\epsilon \geq 0$, and $\nu \in \Delta(\mathbf{X})$, we let $\mathrm{BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \nu) \subseteq \Pi^i$ denote player i's set of ϵ-best-responses to $\boldsymbol{\pi}^{-i}$ with respect to ν. If, additionally, $\pi^i \in \mathrm{BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \nu)$ for all $\nu \in \Delta(\mathbf{X})$, then π^i is called a *uniform ϵ-best-response to $\boldsymbol{\pi}^{-i}$*. The set of uniform ϵ-best-responses to a policy $\boldsymbol{\pi}^{-i}$ is denoted $\mathrm{BR}^i_\epsilon(\boldsymbol{\pi}^{-i})$.

Definition 5 *Let $\epsilon \geq 0$, $\nu \in \Delta(\mathbf{X})$, and $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}$. The joint policy $\boldsymbol{\pi}^*$ is called an* ϵ-equilibrium with respect to ν *if π^{*i} is an ϵ-best-response to $\boldsymbol{\pi}^{*-i}$ with respect to ν for every player $i \in \mathcal{N}$. Additionally, if $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}$ is an ϵ-equilibrium with respect to every $\nu \in \Delta(\mathbf{X})$, then $\boldsymbol{\pi}^*$ is called a* perfect ϵ-equilibrium.

For $\epsilon \geq 0$ and $\nu \in \Delta(\mathbf{X})$, we let $\boldsymbol{\Pi}^{\epsilon\text{-eq}}(\nu) \subset \boldsymbol{\Pi}$ denote the set of ϵ-equilibrium policies with respect to ν, and we let $\boldsymbol{\Pi}^{\epsilon\text{-eq}}$ denote the set of perfect ϵ-equilibrium policies. Furthermore, we let $\boldsymbol{\Pi}_S^{\epsilon\text{-eq}}(\nu) := \boldsymbol{\Pi}^{\epsilon\text{-eq}}(\nu) \cap \boldsymbol{\Pi}_S$ for each $\nu \in \Delta(\mathbf{X})$ and we let $\boldsymbol{\Pi}_S^{\epsilon\text{-eq}} := \boldsymbol{\Pi}^{\epsilon\text{-eq}} \cap \boldsymbol{\Pi}_S$.

Further Comments on Models of Mean-Field Games

The model above differs from the classical model of mean-field games, which assumes a continuum of agents. Here, we consider models with a possibly large but finite number of symmetric, weakly coupled agents. Our model closely resembles that of (Saldi et al., 2018), which studies existence of equilibrium and allows for general state and actions spaces. Unlike (Saldi et al., 2018), we consider games with finite state and action spaces, and we consider a variety of observation channels. Mean-field games can be viewed as limit models of n-player symmetric stochastic games, where players are exchangeable and symmetric. A number of papers have formally examined the connection between games with finitely many players and the corresponding limit model, including the works of (Fischer, 2017; Saldi et al., 2018; Sanjari et al., 2022).

3. From Games to MDPs and Back: Existence of Equilibrium

In this section, we explore the deep connection between POMDPs and n-player mean-field games. We begin by observing that, in any partially observed n-player mean-field game, player i faces a stochastic control problem equivalent to a POMDP whenever its counterparts use a stationary policy. Next, under mean-field observability and a symmetry condition on the joint policy $\boldsymbol{\pi}^{-i}$, we prove that player i faces a problem equivalent to a *fully observed* MDP. This result is of independent interest, but it additionally enables a proof that, under mean-field observability, the set of stationary joint policies admits a perfect equilibrium.

Lemma 6 *Let $\mathbf{G}$ be a partially observed n-player mean-field game. Fix $i \in \mathcal{N}$ and let $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$ be a stationary policy for the remaining players. Then, player i faces a partially observed Markov decision problem $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$, with partially observed state process $\{\mathbf{x}_t\}_{t\geq 0}$* and observation process $\{y^i_t\}_{t\geq 0}$.

Lemma 6, whose proof is straightforward and omitted, gives conditions under which a player faces a POMDP. Under certain additional conditions, such as those presented in Theorem 8, one can show that player $i \in \mathcal{N}$ faces a fully observed MDP. When player i faces an MDP in its observation variable, the classical theory of MDPs and reinforcement learning can be brought to bear on player i's optimization problem, leading to results on the existence of certain equilibrium policies and characterization of one's best-response set.

3.1 An MDP Reduction under Mean-Field Observability

We now state and prove Theorem 8, which provides sufficient conditions for a given player to face an MDP in the mean-field game. Theorem 8 will later play a critical role in the proof of Theorem 12, the second main result of this section, which asserts that the game admits a perfect equilibrium in the set of stationary policies.

Definition 7 *For players $i, j \in \mathcal{N}$, let $\pi^i \in \Pi^i_S$, $\pi^j \in \Pi^j_S$ be stationary policies. We say that the policies π^i and π^j are* symmetric *if both are identified with the same transition kernel in $\mathcal{P}(\mathbb{A}|\mathbb{Y})$. For any subset of players $I \subset \mathcal{N}$, a collection of policies $\{\pi^i\}_{i\in I}$ is called symmetric if, for every $i, j \in I$, we have that π^i and π^j are symmetric.*

Theorem 8 *Let $\mathbf{G}$ be a partially observed n-player mean-field game with mean-field observability (that is, Assumption 2 holds). Let $i \in \mathcal{N}$. If $\boldsymbol{\pi}^{-i} \in \mathbf{\Pi}^{-i}_S$ is symmetric, then i faces a fully observed MDP $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$ with controlled state process $\{y^i_t\}_{t\geq 0}$, where $y^i_t = \varphi^i(\mathbf{x}_t) = (x^i_t, \mu(\mathbf{x}_t))$ for all $t \geq 0$.*

For the proof of Theorem 8, we introduce the following notation: Σ_η is the set of permutations on $\mathcal{N}$. For a given permutation $\sigma \in \Sigma_\eta$ and a global state $\mathbf{s} = (s^i)_{i\in\mathcal{N}}$, let $\sigma(\mathbf{s})$ be the global state in which $\sigma(\mathbf{s})^i := s^{\sigma(i)}$ for each player $i \in \mathcal{N}$. In other words, the local state of player i in global state $\sigma(\mathbf{s})$ is given by the local state of player $\sigma(i)$ in global state $\mathbf{s}$. Similarly, for a joint action $\mathbf{a} = (a^i)_{i\in\mathcal{N}}$, we let $\sigma(\mathbf{a})^i := a^{\sigma(i)}$ for each player $i \in \mathcal{N}$.

Proof Fix $i \in \mathcal{N}$. Recall that $\mathbb{Y} = \mathbb{X} \times \mathrm{Emp}_n$ and player i's observations are given by $\varphi^i(\mathbf{s}) = (s^i, \mu(\mathbf{s}))$ for any $\mathbf{s} \in \mathbf{X}$. For any $k \geq 0$, we let $y^i_k := \varphi^i(\mathbf{x}_k) = (x^i_k, \mu(\mathbf{x}_k))$.

Let $\nu \in \Delta(\mathbf{X})$ be any initial distribution of the global state variable and let $\boldsymbol{\pi}^{-i} \in \mathbf{\Pi}^{-i}_S$ be a symmetric policy for the remaining players. Let $\pi^i \in \Pi^i$ be any policy for player i, and let $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i})$. Note that player i's cost at time t, c^i_t, is a measurable function of (y^i_t, a^i_t):

$$c^i_t = c(x^i_t, \mu(\mathbf{x}_t), a^i_t) = c(y^i_t, a^i_t),\ \forall t \geq 0.$$

That is, player i's observation variable y^i_t summarizes the cost-relevant part of the system history for player i. Therefore, we must show that the following holds for all $t \geq 0$, any $\Upsilon \subseteq \mathbb{Y}$, and in a time invariant manner:

$$\Pr^{\boldsymbol{\pi}}_\nu \left[y^i_{t+1} \in \Upsilon \middle| \{y^i_k, a^i_k : 0 \leq k \leq t\}\right] = \Pr^{\boldsymbol{\pi}}_\nu \left[y^i_{t+1} \in \Upsilon \middle| y^i_t, a^i_t\right],$$

$\Pr^{\boldsymbol{\pi}}_\nu$-almost surely.[1]

1. For brevity, we omit the qualifier "$\Pr^{\boldsymbol{\pi}}_\nu$-almost surely" for all subsequent equalities involving conditional expectations. The time invariance described here is for time homogeneity of the MDP, and requires that the right side of this expression does not depend on t.

Fix $t \geq 0$. For any $\Upsilon \subseteq \mathbb{Y}$, let $\mathbf{1}_\Upsilon$ denote the indicator function of the event $\{y^i_{t+1} \in \Upsilon\}$. Using the law of iterated expectations and conditioning on the (finer) σ-algebra generated by the random variables $\{\mathbf{x}_k, a^i_k : 0 \leq k \leq t\}$, we have

$$\begin{aligned} E^{\boldsymbol{\pi}}_\nu \left(\mathbf{1}_\Upsilon \middle| \{y^i_k, a^i_k : 0 \leq k \leq t\}\right) =& E^{\boldsymbol{\pi}}_\nu \left[E^{\boldsymbol{\pi}}_\nu \left(\mathbf{1}_\Upsilon \middle| \{\mathbf{x}_k, a^i_k : 0 \leq k \leq t\}\right) \middle| \{y^i_k, a^i_k : 0 \leq k \leq t\}\right] \\ =& E^{\boldsymbol{\pi}}_\nu \left[E^{\boldsymbol{\pi}}_\nu \left(\mathbf{1}_\Upsilon \middle| \mathbf{x}_t, a^i_t\right) \middle| \{y^i_k, a^i_k : 0 \leq k \leq t\}\right]. \end{aligned}$$

Thus, it suffices to show that

$$E^{\boldsymbol{\pi}}_\nu \left(\mathbf{1}_\Upsilon \middle| \mathbf{x}_t, a^i_t\right) = E^{\boldsymbol{\pi}}_\nu \left(\mathbf{1}_\Upsilon \middle| y^i_t, a^i_t\right) \tag{2}$$

holds for all $\Upsilon \subseteq \mathbb{Y}$: since the left side of (2) does not vary with t, this will establish that $\{y^i_t\}_{t\geq 0}$ is a (time homogeneous) controlled Markov process controlled by $\{a^i_t\}_{t\geq 0}$. Moreover, since $\mathbb{Y}$ is a finite set, it suffices to show that this holds for all singletons $\Upsilon = \{w\}$, $w \in \mathbb{Y}$.

Fix $w \in \mathbb{Y}$ and note that the events $\{y^i_{t+1} = w\}$ and $\{\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w)\}$ are equivalent. We claim that

$$\begin{aligned} &\Pr^{\boldsymbol{\pi}}_\nu \left[\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \mathbf{s}, a^i_t = a^i\right] \\ =&\Pr^{\boldsymbol{\pi}}_\nu \left[\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t \in \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s})), a^i_t = a^i\right] \end{aligned} \tag{3}$$

holds for any $\mathbf{s} \in \mathbf{X}$ and any $a^i \in \mathbb{A}$, where $\left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))$ denotes the pre-image of $\varphi^i(\mathbf{s})$. By mean-field observability, we can explicitly characterize the set $\left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))$ as

$$\begin{aligned} \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s})) &= \left\{\tilde{\mathbf{s}} \in \mathbf{X} : \varphi^i(\tilde{\mathbf{s}}) = \varphi^i(\mathbf{s})\right\} \\ &= \left\{\tilde{\mathbf{s}} \in \mathbf{X} : \tilde{s}^i = s^i \text{ and } \boldsymbol{\mu}(\tilde{\mathbf{s}}) = \boldsymbol{\mu}(\mathbf{s})\right\} \\ &= \{\sigma(\mathbf{s}) | \sigma \in \Sigma_\eta : \sigma(i) = i\}. \end{aligned} \tag{4}$$

The final expression comes from the fact that $\varphi^i(\mathbf{s}) = \varphi^i(\tilde{\mathbf{s}}) \iff \tilde{\mathbf{s}}$ can be obtained from $\mathbf{s}$ by permuting the local states of other players while leaving i's local state fixed.

To verify the claim in (3), let $\tilde{\mathbf{s}} \in \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))$ be an arbitrary global state with $\varphi^i(\tilde{\mathbf{s}}) = \varphi^i(\mathbf{s})$. Let $\sigma \in \Sigma_\eta$ be a permutation of the players such that $\sigma(i) = i$ and $\tilde{s}^j = s^{\sigma(j)}$ for all $j \in \mathcal{N}\setminus\{i\}$. Using the notational conventions described earlier, we have that $\tilde{\mathbf{s}} = \sigma(\mathbf{s})$.

Since the policies π^j and $\pi^{\sigma(j)}$ are symmetric, for any $\mathbf{s} \in \mathbf{X}$ and $\mathbf{a} \in \mathbf{A}$, one has

$$\Pr^{\boldsymbol{\pi}}_\nu \left(\mathbf{a}_t = \mathbf{a} \middle| \mathbf{x}_t = \mathbf{s}, a^i_t = a^i\right) = \Pr^{\boldsymbol{\pi}}_\nu \left(\mathbf{a}_t = \sigma(\mathbf{a}) \middle| \mathbf{x}_t = \sigma(\mathbf{s}), a^i_t = a^i\right) \tag{5}$$

Since the local states $\{x^j_{t+1}\}_{j\in\eta}$ are conditionally independent given $(\mathbf{x}_t, \mathbf{a}_t)$, the global state transition function has a product structure, which is invariant under permutations:

$$\begin{aligned} \Pr^{\boldsymbol{\pi}}_\nu \left(\mathbf{x}_{t+1} = \mathbf{s}' \middle| \mathbf{x}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a}\right) &= \prod_{j\in\mathcal{N}} P_{\text{loc}}\left(s'^j \middle| s^j, \boldsymbol{\mu}(\mathbf{s}), a^j\right) \\ &= \prod_{j\in\mathcal{N}} P_{\text{loc}}\left(s'^{\sigma(j)} \middle| s^{\sigma(j)}, \boldsymbol{\mu}(\sigma(\mathbf{s})), a^{\sigma(j)}\right) \\ &= \Pr^{\boldsymbol{\pi}}_\nu \left(\mathbf{x}_{t+1} = \sigma(\mathbf{s}') \middle| \mathbf{x}_t = \sigma(\mathbf{s}), \mathbf{a}_t = \sigma(\mathbf{a})\right). \end{aligned} \tag{6}$$

Using the law of total probability, conditioning on $\mathbf{a}_t$, and recalling that $\tilde{\mathbf{s}} = \sigma(\mathbf{s})$, it follows that for any $\mathbf{s}' \in \mathbf{X}$,

$$\Pr_\nu^{\boldsymbol{\pi}}\left[\mathbf{x}_{t+1} = \mathbf{s}' \middle| \mathbf{x}_t = \mathbf{s}, a_t^i = a^i\right] = \Pr_\nu^{\boldsymbol{\pi}}\left[\mathbf{x}_{t+1} = \sigma(\mathbf{s}') \middle| \mathbf{x}_t = \tilde{\mathbf{s}}, a_t^i = a^i\right]. \tag{7}$$

Noting that $\mu(\mathbf{s}') = \mu(\sigma(\mathbf{s}'))$, we have that $\mathbf{s}' \in \left(\varphi^i\right)^{-1}(w) \iff \sigma(\mathbf{s}') \in \left(\varphi^i\right)^{-1}(w)$. Thus,

$$\left(\varphi^i\right)^{-1}(w) = \bigcup_{\mathbf{s}' \in (\varphi^i)^{-1}(w)} \{\mathbf{s}'\} = \bigcup_{\mathbf{s}' \in (\varphi^i)^{-1}(w)} \{\sigma(\mathbf{s}')\}. \tag{8}$$

It follows that

$$\begin{aligned}
\Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \mathbf{s}, a_t^i = a^i\right) &= \sum_{\mathbf{s}' \in (\varphi^i)^{-1}(w)} \Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} = \mathbf{s}' \middle| \mathbf{x}_t = \mathbf{s}, a_t^i = a^i\right) \\
&= \sum_{\mathbf{s}' \in (\varphi^i)^{-1}(w)} \Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} = \sigma(\mathbf{s}') \middle| \mathbf{x}_t = \sigma(\mathbf{s}), a_t^i = a^i\right) \\
&= \Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \sigma(\mathbf{s}), a_t^i = a^i\right).
\end{aligned}$$

Since $\tilde{\mathbf{s}} \in \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))$ was arbitrary, using (4), we see that

$$\Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \mathbf{s}, a_t^i = a^i\right) = \Pr_\nu^{\boldsymbol{\pi}}\left(\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \bar{\mathbf{s}}, a_t^i = a^i\right), \tag{9}$$

for any $\bar{\mathbf{s}} \in \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))$. We conclude that (3) holds by applying iterated expectations to its right-hand side, conditioning on $\{\mathbf{x}_t, a_t^i\}$, and using (9) to simplify the resulting summation.

Finally, the event $\{\mathbf{x}_t \in \left(\varphi^i\right)^{-1}(\varphi^i(\mathbf{s}))\}$ is equivalent to the event $\{y_t^i = \varphi^i(\mathbf{s})\}$ and $\{\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w)\}$ is equivalent to $\{y_{t+1}^i = w\}$. We have therefore shown that

$$\Pr^{\boldsymbol{\pi}}\left[\mathbf{x}_{t+1} \in \left(\varphi^i\right)^{-1}(w) \middle| \mathbf{x}_t = \mathbf{s}, a_t^i = a^i\right] = \Pr^{\boldsymbol{\pi}}\left[y_{t+1}^i = w \middle| y_t^i = \varphi^i(\mathbf{s}), a_t^i = a^i\right],$$

for any $w \in \mathbb{Y}$, $\mathbf{s} \in \mathbf{X}$, and $a^i \in \mathbb{A}$. The result follows. ■

Comments on the Proof of Theorem 8

The proof of Theorem 8 depends critically on the connection between local observations and hidden global states described in (4) : under mean-field observability, two global states $\mathbf{s}, \tilde{\mathbf{s}} \in \mathbf{X}$ give rise to the same local observation for player i, i.e. $\varphi^i(\mathbf{s}) = \varphi^i(\tilde{\mathbf{s}})$, if and only if they agree on the local state of player i (i.e. $s^i = \tilde{s}^i$) *and* one can be obtained from the other by permuting the local states of the remaining agents. This specific characterization of pre-images of φ^i holds only for mean-field observability and fails under compressed or local observability. As a result, the proof above cannot be used in those settings. Indeed, we will show by counterexample in Section 3.3 that both mean-field observability and symmetry of the policy $\boldsymbol{\pi}^{-i}$ were necessary in Theorem 8 and cannot be relaxed without loss of generality.

A byproduct of the proof of Theorem 8 concerns the conditional distribution of player i's observation-action trajectories, $\{y_t^i, a_t^i\}_{t\geq 0}$, conditional on the initial observation y_0^i. In particular, with π^i arbitrary and $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$ symmetric, putting $t = 0$ in equation (2) and then invoking Theorem 8, one obtains the following equality for any initial measures $\nu, \nu' \in \Delta(\mathbf{X})$:

$$\begin{aligned}&\Pr_{\nu}^{\boldsymbol{\pi}}\left(\{y_t^i, a_t^i\}_{t\geq 0} \in \cdot \,\middle|\, y_0^i = y, a_0^i = a^i\right)\\ =&\Pr_{\nu'}^{\boldsymbol{\pi}}\left(\{y_t^i, a_t^i\}_{t\geq 0} \in \cdot \,\middle|\, y_0^i = y, a_0^i = a^i\right), \quad \forall (y, a^i) \in \mathbb{Y} \times \mathbb{A}. \end{aligned} \tag{10}$$

Q-Functions Under Mean-Field Observability and Symmetry

In light of Theorem 8, we define the Q-function for player i when playing $\mathbf{G}$ against a symmetric policy $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$ as

$$Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i) := E_{\nu}^{(\pi^{*i}, \boldsymbol{\pi}^{-i})}\left[\sum_{t=0}^{\infty} \gamma^t c(x_t^i, \mu(\mathbf{x}_t), a_t^i) \middle| y_0^i = y, a_0^i = a^i\right], \quad \forall a^i \in \mathbb{A},$$

for every $y \in \varphi^i(\mathbf{X}) = \{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$, where $\pi^{*i} \in \mathrm{BR}_0^i(\boldsymbol{\pi}^{-i}) \cap \Pi_S^i$ is a best response to $\boldsymbol{\pi}^{-i}$ and $\nu \in \Delta(\mathbf{X})$ is arbitrary.[2] For elements $y \in \mathbb{Y} \setminus \varphi^i(\mathbf{X})$, which do not arise as the observation output of any global state, we may define $Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i)$ arbitrarily, say $Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, \cdot) \equiv 0$.

For any player $i \in \mathcal{N}$, we let $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i} \subset \boldsymbol{\Pi}_S^{-i}$ denote the set of symmetric joint policies for the remaining players, and we let $\boldsymbol{\Pi}_{S,\mathrm{sym}} \subset \boldsymbol{\Pi}_S$ denote the set of stationary joint policies that are symmetric. We note that the sets Π_S^i and $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$ are in bijection, and we define $\mathrm{sym}^i : \Pi_S^i \to \boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$ by $\mathrm{sym}^i(\pi^i) = \left(\pi^i\right)_{j \in \mathcal{N}\setminus\{i\}}$, for all $\pi^i \in \Pi_S^i$. We metrize $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$ using the metric d^i on Π_S^i:

$$\mathbf{d}_{\mathrm{sym}}^{-i}\left(\mathrm{sym}^i(\pi^i), \mathrm{sym}^i(\tilde{\pi}^i)\right) := d^i(\pi^i, \tilde{\pi}^i), \quad \forall \pi^i, \tilde{\pi}^i \in \Pi_S^i. \tag{11}$$

We note that the metric $\mathbf{d}_{\mathrm{sym}}^{-i}$ is equivalent to the metric $\mathbf{d}^{-i}$ restricted to $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$. We also define an analogous metric $\mathbf{d}_{\mathrm{sym}}$ on $\boldsymbol{\Pi}_{S,\mathrm{sym}}$.

We now state some lemmas on the continuity of various functions, where continuity is with respect to the metrics defined above. The proofs of these lemmas resemble those of (Yongacoglu et al., 2023, Lemmas 2.10-2.13) and are omitted. In each of the lemmas below, we let $\mathbf{G}$ be an n-player mean-field game with mean-field observability (Assumption 2), and let $i \in \mathcal{N}$ be any player.

Lemma 9 *Fix $(y, a^i) \in \mathbb{Y} \times \mathbb{A}$. The mapping $\boldsymbol{\pi}^{-i} \mapsto Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i)$ is continuous on $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$.*

Lemma 10 *Fix $y \in \mathbb{Y}$. The mapping $\boldsymbol{\pi}^{-i} \mapsto \min_{a^i \in \mathbb{A}} Q^{*i}_{\boldsymbol{\pi}^{-i}}(y, a^i)$ is continuous on $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$.*

Lemma 11 *Fix $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$. The following mapping is continuous on Π_S^i:*

$$\pi^i \mapsto \max_{s \in \mathbf{X}} \left(J^i(\pi^i, \boldsymbol{\pi}^{-i}, \mathbf{s}) - \min_{a^i \in \mathbb{A}} Q^{*i}_{\boldsymbol{\pi}^{-i}}(\varphi^i(\mathbf{s}), a^i) \right).$$

2. That ν can be arbitrarily chosen follows from the preceding discussion culminating in (10).

3.2 Existence of Stationary Equilibrium under Mean-Field Observability

Theorem 12, stated below, asserts that a stationary perfect equilibrium exists when the game has mean-field observability. The proof technique parallels that of (Fink, 1964, Theorem 2), making the required modifications to account for partial observability of the global state.

Theorem 12 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumption 2. For any* $\epsilon \geq 0$*, there exists a perfect* ϵ*-equilibrium policy in* $\mathbf{\Pi}_S$*. That is,* $\mathbf{\Pi}_S^{\epsilon\text{-eq}} \neq \varnothing$*.*

Proof Fix player $i \in \mathcal{N}$. We define a point-to-set mapping $\mathscr{B}^i : \Pi_S^i \to 2^{\Pi_S^i}$ with

$$\mathscr{B}^i(\pi^i) = \mathrm{BR}_0^i(\mathrm{sym}^i(\pi^i)) \cap \Pi_S^i, \quad \forall \pi^i \in \Pi_S^i.$$

By Theorem 8, player i is facing an MDP with finite state and action spaces when playing against $\mathrm{sym}^i(\pi^i)$, and the set of stationary optimal policies for this MDP is given by $\mathscr{B}^i(\pi^i) \subseteq \Pi_S^i$. Thus, $\mathscr{B}^i(\pi^i)$ is non-empty, convex, and compact for each $\pi^i \in \Pi_S^i$.

If $(\mathrm{sym}^i(\tilde{\pi}_k^i))_{k\geq 0}$ is a sequence of symmetric joint policies in $\mathbf{\Pi}_{S,\mathrm{sym}}^{-i}$ converging to some $\mathrm{sym}^i(\tilde{\pi}_\infty^i) \in \mathbf{\Pi}_{S,\mathrm{sym}}^{-i}$ and $(\pi_k^i)_{k\geq 0}$ is a sequence in Π_S^i such that (1) $\lim_{k\to\infty} \pi_k^i = \pi_\infty^i \in \Pi_S^i$ and (2) for every $k \geq 0$, we have that $\pi_k^i \in \mathrm{BR}_0^i(\mathrm{sym}^i(\tilde{\pi}_k^i))$, then one can use Lemma 3, Lemmas 9–11, and Lemma 31 to conclude that $\pi_\infty^i \in \mathrm{BR}_0^i(\mathrm{sym}^i(\tilde{\pi}_\infty^i)) \cap \Pi_S^i = \mathscr{B}^i(\tilde{\pi}_\infty^i)$. This implies that the point-to-set mapping $\mathscr{B}^i$ is upper hemicontinuous.

We invoke Kakutani's fixed point theorem on $\mathscr{B}^i$ to obtain a fixed point

$$\pi^{*i} \in \mathscr{B}^i(\pi^{*i}) = \mathrm{BR}_0^i(\mathrm{sym}^i(\pi^{*i})) \cap \Pi_S^i.$$

By symmetry, $\big(\pi^{*i}, \mathrm{sym}^i(\pi^{*i})\big) \in \Pi_S^i \times \mathbf{\Pi}_{S,\mathrm{sym}}^{-i} \subset \mathbf{\Pi}_S$ is a perfect 0-equilibrium for $\mathbf{G}$, and *a fortiori* a perfect ϵ-equilibrium for $\mathbf{G}$. ■

To the best of our knowledge, this result is new. It differs from earlier results in that it considers a setting with finitely many players, discrete time, discounted costs, and partial observability. The (person-by-person) optimality of each player's policy holds without approximation ($\epsilon = 0$) and holds over all admissible policies and not merely over the player's stationary policies. Moreover, this optimality holds for any initial distribution, and for any (finite) number of players. To better situate this result in the literature, we now describe some related results.

In the continuous time literature, (Arapostathis et al., 2017) prove the existence of equilibrium in games with finitely many players and an ergodic (average) cost criterion. (Lacker, 2015) studies existence of Markov equilibrium in the limiting regime with infinitely many players. For games with finitely many players, several results about equilibria among feedback strategies with global observability are given in (Lacker, 2020). Further discussion on the topic can be found in (Fischer, 2017).

In discrete time, (Perrin et al., 2022) consider the limiting regime and study what they call population-dependent policies, with the goal of obtaining solutions that perform well across initial distributions. This consideration of the initial distribution is the defining feature of uniform best-responding and perfect equilibrium as studied here. In our framework, their notion of population-dependent policies corresponds to our notion of stationary policies under mean-field observability.

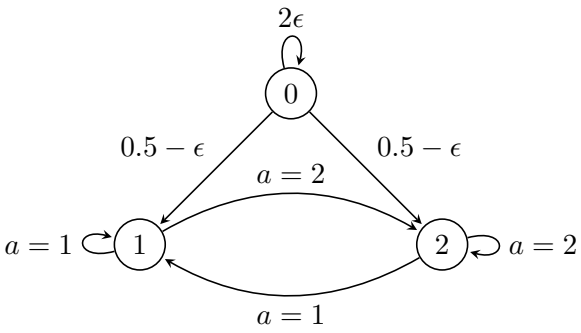

Figure 1: Local State Transition Probabilities for Example 1

Also in the discrete-time setting, (Biswas, 2015) studies mean-field games with finitely many players and local observability. Using the ergodic (average) cost criterion, the existence of a certain kind of equilibrium is proved. The equilibrium considered by (Biswas, 2015) is somewhat less demanding than the one considered here: in our definition of a perfect equilibrium, each player i's policy must be optimal over its complete set of admissible policies Π^i, rather than optimal over the restricted set $\Pi^i_S \subsetneq \Pi^i$.

3.3 Counterexamples to MDP Equivalence

We now provide explicit counterexamples showing that, in general, player i will not face an MDP in $\{y^i_t\}_{t\geq 0}$ when either mean-field observability or symmetry of $\boldsymbol{\pi}^{-i}$ fail to hold.

Example 1: On Relaxing the Symmetry of $\boldsymbol{\pi}^{-i}$

If one relaxes the symmetry assumption in Theorem 8, then the controlled Markov property of $\{y^i_t\}_{t\geq 0}$ is lost in general, even under mean-field observability. We illustrate this using a simple example, with $\mathbb{X} = \mathbb{A} = \{0,1,2\}$ and $n = 3$, though the example extends to any number of players. Local state transitions depend only on local state and action without dependence on the mean-field term. The relevant probabilities are summarized in Figure 1, with $\epsilon \in (0, 0.5)$ arbitrary and transitions from states 1 or 2 to state 0 omitted for clarity. In words, the transition probabilities out of state 0 are independent of both action and the mean-field state, with transitions to states 1 and 2 being equally likely, with probability $0.5-\epsilon$. Transitions out of states 1 and 2 are deterministic, with $x^j_{t+1} = a^j_t$ for any agent j. That is, $P_{\mathrm{loc}}(a|s,\mu,a) = 1$, for $s \in \{1,2\}$, $\mu \in \mathrm{Emp}_n$, and $a \in \{0,1,2\}$.

We define a cost function $c(x,\mu,a) := \kappa \cdot \mathbf{1}\{\mu(x) \geq 0.5\}$ for any $(x,\mu,a) \in \mathbb{X}\times\Delta(\mathbb{X})\times\mathbb{A}$, where $\kappa > 0$ is a penalty for being in the same state as the majority of all agents. Although the mean-field term μ_t does not appear in its local state transition probabilities, the mean-field term is nevertheless relevant to a given player's cost.

We define two stationary policies, denoted $\pi_{\mathrm{stay}} \in \mathscr{P}(\mathbb{A}|\mathbb{Y})$ and $\pi_{\mathrm{go2}} \in \mathscr{P}(\mathbb{A}|\mathbb{Y})$. These policies are given by $\pi_{\mathrm{stay}}(\cdot|s,\mu) := \delta_s$ and $\pi_{\mathrm{go2}}(\cdot|s,\mu) := \delta_2$, for all $(s,\mu) \in \mathbb{Y}$. In words, the policy π_{stay} always chooses $a^j_t = x^j_t$, while π_{go2} always chooses $a^j_t = 2$.

We fix $i = 3$ to be the index of the last player in the game, and we consider the *non-symmetric* stationary joint policy $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$ given by $\pi^1 = \pi_{\text{stay}}$ and $\pi^2 = \pi_{\text{go2}}$, and we let $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i})$, where $\pi^i \in \Pi^i$ is arbitrary.

Suppose that the initial distribution $\nu \in \Delta(\mathbf{X})$ is such that each player j's initial state is $x_0^j = 0$ with probability 1. Writing an element $(x^i, \mu) \in \mathbb{Y}$ as $\big(x^i, [\mu(0), \mu(1), \mu(2)]\big)$, one can establish the following inequality

$$\Pr_\nu^{\boldsymbol{\pi}}\left(y_{t+1}^i = (2, [0,0,1]) \middle| y_t^i = \left(0, \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]\right), a_t^i = a^i\right) > 0.$$

On the other hand, conditioning also on y_{t-1}^i, one has the following:

$$\Pr_\nu^{\boldsymbol{\pi}}\left(y_{t+1}^i = (2, [0,0,1]) \middle| y_t^i = y_{t-1}^i = \left(0, \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]\right)\right) = 0.$$

The condition $y_t^i = y_{t-1}^i$ here implies that (a) player 2 (who changes state from 1 to 2 given the chance) was already in state 2 at time $t-1$, and (b) player 1 (who remains in whatever state it is already in) was in state 1 at time $t-1$. Hence, neither will choose to change its state at time t.

In summary, as a_t^i did not feature in the analysis above, we have observed that

$$\Pr_\nu^{\boldsymbol{\pi}}\left(y_{t+1}^i \in \cdot \middle| y_t^i, a_t^i\right) \neq \Pr_\nu^{\boldsymbol{\pi}}\left(y_{t+1}^i \in \cdot \middle| \left\{y_k^i, a_k^i : 0 \leq k \leq t\right\}\right),$$

which establishes that $\{y_t^i, a_t^i\}_{t\geq 0}$ is not a controlled Markov process. ⋄

Example 2: On Relaxing Mean-Field Observability

We now show that if one relaxes the mean-field observability assumption in Theorem 8, then player i does not face an MDP with state $\{y_t^i\}_{t\geq 0}$. In this example, one sees that the local observation can fail to capture the cost-relevant history and furthermore can fail to satisfy the controlled Markov property, even under symmetry of the joint policy $\boldsymbol{\pi}^{-i}$. Thus, both defining features of MDPs are violated in the counterexample below.

Consider an n-player mean-field game $\mathbf{G}$ with local observability, $\mathbb{X} = \{1, \ldots, 10\}$, and $\mathbb{A} = \text{Emp}_n$. Define the cost function as $c(s, \mu, a) = \|\mu - a\|_2$ for each $(s, \mu, a) \in \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A}$. Fixing $\delta > 0$, the transition kernel P_{loc} is given by

$$P_{\text{loc}}(s'|s, \mu, a) = \begin{cases} \delta, & \text{if } s' = 10 \\ 1 - \delta, & \text{if } s' = \left\lceil \sum_{\bar{s}\in\mathbb{X}} \bar{s}\mu(\bar{s}) \right\rceil \\ 0, & \text{else.} \end{cases}$$

In words, player i's local state transitions either to the mean of the mean-field (rounding up to the nearest integer) or to the maximum value, 10. Decreases to one's local state inform the player about the mean of the preceding mean-field term: if $x_{t+1}^i = \omega + 1 < x_t^i$, then the mean of μ_t lies in $(\omega, \omega + 1]$. Note also that the means of $\{\mu_t\}_{t\geq 0}$ are non-decreasing.

The discount factor $\gamma \in (0, 1)$ is set arbitrarily, and the initial distribution $\nu_0 \in \Delta(\mathbf{X})$ is chosen such that $\{x_0^j\}_{j\in\mathcal{I}}$ is independently and identically distributed according to the uniform distribution on $\mathbb{X}$.

Each player i's cost depends only on its chosen action and the empirical distribution associated with the global state. Thus, each player i minimizes its costs by tracking the (unobserved) mean-field process $\{\mu_t\}_{t\geq 0}$. It is clear that player i's cost at time t is not measurable with respect to $(y_t^i, a_t^i) = (x_t^i, a_t^i)$. By itself, this shows that $\{y_t^i\}_{t\geq 0}$ is not the state variable for an MDP, since it does not summarize the cost-relevant history.

Let $\boldsymbol{\pi}$ be *any* joint policy. We have $\Pr_{\nu_0}^{\boldsymbol{\pi}}\left(x_{t+1}^i \leq 4 \middle| x_t^i = 10\right) > 0$. However, decreases in one's local state are informative about the mean-field, so conditioning on y_{t-1}^i and y_{t-2}^i can change this strict inequality to an equality:

$$\Pr_{\nu_0}^{\boldsymbol{\pi}}\left(x_{t+1}^i = 4 \middle| x_t^i = 10, x_{t-1}^i = 5, x_{t-2}^i = 10\right) = 0.$$

This shows that $\{y_t^i = x_t^i\}_{t\geq 0}$ also fails to satisfy the controlled Markov property.

In summary, the discussion above shows that the control problem faced by player i in a partially observed n-player mean-field game with compressed or local observability is not generally equivalent to a fully observed MDP. This may occur either because the observation variable fails to summarize the cost-relevant history of the system, because the observation variable fails to satisfy the controlled Markov property, or both. ◇

3.4 Further Comments on Equilibrium in n-Player Mean-Field Games

On MDP Equivalence and Equilibrium Under Global Observability: When an n-player mean-field game has global observability, the result is a fully observed stochastic game. In this case, results analogous to Theorems 8 and 12 are well-known (Fink, 1964). The results above do not follow from this well-known theory due to the partial observability of the global state variable.

Toward Equilibrium under Compressed Observability: Under compressed observability, one cannot guarantee the existence of a perfect equilibrium among the stationary policies without loss of generality, as evidenced by the game in Example 2. Nevertheless, it may be desirable to identify (restrictive) sufficient conditions for existence. One approach involves taking the number of players to be large enough that the strategic coupling is negligible and the mean-field sequence $\{\mu_t\}_{t\geq 0}$ is essentially constant over the effective planning horizon. For a result on existence of ϵ-equilibrium under Assumption 4 as well as further discussion, see Appendix H.

4. Independent Learning and Subjectivity

In this section, we study independent learning in n-player mean-field games by harnessing the connections between (PO)MDPs and the control problem of a player in the game.

4.1 Independent Learning

In large, decentralized systems, agents have limited information about the overall system. For instance, the policies, actions, and local states of other agents may be unobservable or difficult to estimate using local data. Moreover, maintaining estimates of various global quantities introduces a massive computational and communication burden at each agent.

Independent learning is one approach to learning in decentralized settings. In this paradigm, agents use only local information and run single-agent learning algorithms, treat-

ing their environment as though it were a fully observed (single-agent) MDP. Such intentional obliviousness is characteristic of independent learners and is used to alleviate the computational burden at each agent. In effect, each independent learning agent makes an (erroneous) simplifying assumption about its environment to obtain tractable, scalable algorithms in complex multi-agent settings. In this paper, we interpret this simplifying assumption as implying *subjective beliefs* held by the player about its system.

The independent learning paradigm is similar to various algorithms designed for single-agent, non-Markovian environments. In such settings, optimal policies are generally non-stationary and history-dependent. This leads to learning and planning problems that are generally intractable. Recently, some authors have proposed use of simpler surrogates for the full system history, in an effort to inform tractable algorithm design in complex environments. Examples of work in this line include the agent state concept introduced in (Dong et al., 2022) and leveraged in (Sinha et al., 2024), the approximate information state concept of (Subramanian et al., 2022a), the approximate belief state of (Kara and Yüksel, 2022), and the quantized belief state of (Kara and Yüksel, 2024). Viewed as an instance of this approach, independent learners in games are simple agents in complex, non-stationary environments and local observations can be seen as agent states or approximate belief states.

In the literature on multi-agent reinforcement learning with independent learning agents, learning and policy adjustment are typically interleaved: agents follow a particular stationary policy to collect feedback data for a number of interactions, and then update their policy parameters using estimates of value functions or some other object. (For example, see the work on regret testers by (Foster and Young, 2006).) It is therefore important to understand the properties of estimates arising from the learning process. In this section, we study the convergence properties of value function estimates obtained by independent learners in n-player mean-field games. A full learning algorithm, with policy dynamics interleaved with value estimation, is presented in Section 5.

4.2 Single-Agent Learning Estimates under Stationary Policies

We now study the learning iterates of agents who attempt to naively estimate a Q-function and state value function, treating their local observations as though they were state variables for an MDP. We study the evolution of these learning iterates under stationary policies, leaving aside the challenge of non-stationarity that arises when policies are adjusted during play, which will be considered later.

In Algorithm 1, below, each player runs two stochastic approximation algorithms, one resembling Q-learning, the other resembling value function estimation. Since player i does not generally face an MDP, the Q-function and state value function of single-agent theory need not be expressible as a function of player i's local observations. For this reason, we interpret Algorithm 1 as learning *subjective* value functions that are compatible with player i's beliefs but need not have a meaningful connection to the objective function.

We now study the convergence properties of the iterate sequences $\{\bar{Q}^i_t, \bar{J}^i_t\}_{t\geq 0}$, which are produced by Algorithm 1, and we provide sufficient conditions for their almost sure convergence. As we will describe, these conditions are relatively mild and can be relaxed further if desired. Importantly, the convergence of iterates does not depend on whether or not player i faces an MDP in its observation variable $\{y^i_t\}_{t\geq 0}$.

Algorithm 1: Independent Learning of (Subjective) Value Functions

Initialize Soft $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$, $\bar{Q}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{A}}$ and $\bar{J}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}}$

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

Simultaneously, each player $j \in \mathcal{n}$ selects $a^j_t \sim \pi^j(\cdot|y^j_t)$

Player i observes y^i_{t+1} and cost $c^i_t := c(x^i_t, \mu(\mathbf{x}_t), a^i_t)$

$n^i_t := \sum_{k=0}^{t} \mathbf{1}\{(y^i_k, a^i_k) = (y^i_t, a^i_t)\}$

$m^i_t := \sum_{k=0}^{t} \mathbf{1}\{y^i_k = y^i_t\}$

Q-factor update:

$$\bar{Q}^i_{t+1}(y^i_t, a^i_t) = \left(1 - \frac{1}{n^i_t}\right) \bar{Q}^i_t(y^i_t, a^i_t) + \frac{1}{n^i_t}\left(c^i_t + \gamma \min_{a^i \in \mathbb{A}} \bar{Q}^i_t(y^i_{t+1}, a^i)\right),$$

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

Value function update:

$$\bar{J}^i_{t+1}(y^i_t) = \left(1 - \frac{1}{m^i_t}\right) \bar{J}^i_t(y^i_t) + \frac{1}{m^i_t}\left(c^i_t + \gamma \bar{J}^i_t(y^i_{t+1})\right),$$

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

We begin with an assumption on the transition kernel P_{loc}. For intuitive simplicity, we state this assumption in terms of the underlying state process.

Assumption 5 *Under any stationary policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$ and initial distribution $\nu \in \Delta(\mathbf{X})$, the global state process $\{\mathbf{x}_t\}_{t\geq 0}$ is an irreducible, aperiodic Markov chain on $\mathbf{X}$.*

Assumption 5 is satisfied, for instance, when $P_{\text{loc}}(s'|s, \mu, a) > 0$ for any arguments $(s', s, \mu, a) \in \mathbb{X} \times \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A}$, but applies much more broadly. Assumption 5 is sufficient but not necessary for the following results: it can be relaxed by assuming only that, for each soft policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$, there exists a unique probability measure $\nu^\infty_{\boldsymbol{\pi}} \in \Delta(\mathbf{X})$ such that the time averaged occupancy measure converges to $\nu^\infty_{\boldsymbol{\pi}}$.[3] That is, one can relax Assumption 5 by assuming only that, for any initial measure ν_0, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \Pr^{\boldsymbol{\pi}}_{\nu_0}[\mathbf{x}_t \in \cdot] \to \nu^\infty_{\boldsymbol{\pi}}, \text{ as } T \to \infty.$$

Theorem 13 *Let $\mathbf{G}$ be a partially observed n-player mean-field game for which Assumption 5 holds, and let $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$ be soft. Suppose player $i \in \mathcal{n}$ uses Algorithm 1. For deterministic functions $V^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \to \mathbb{R}$ and $W^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \times \mathbb{A} \to \mathbb{R}$, defined in Appendix B, we have the following:*

3. A policy is soft if it places positive probability on each action in any context. See Definition 29.

1. *For any $\nu \in \Delta(\mathbf{X})$, $\Pr_\nu^{\boldsymbol{\pi}}$ almost surely, we have*

$$\lim_{t\to\infty} \bar{J}_t^i = V_{\boldsymbol{\pi}}^{*i} \quad \textit{and} \quad \lim_{t\to\infty} \bar{Q}_t^i = W_{\boldsymbol{\pi}}^{*i}.$$

2. *If Assumption 1 holds, then $V_{\boldsymbol{\pi}}^{*i}(\mathbf{s}) = J^i(\boldsymbol{\pi}, \mathbf{s})$ for all $\mathbf{s} \in \mathbf{X}$ and $W_{\boldsymbol{\pi}}^{*i} = Q_{\boldsymbol{\pi}^{-i}}^{*i}$.*
3. *Under mean-field observability (Assumption 2), if $\boldsymbol{\pi}^{-i}$ is symmetric, then $V_{\boldsymbol{\pi}}^{*i}(\varphi^i(\mathbf{s})) = J^i(\boldsymbol{\pi}, \mathbf{s})$ for all $\mathbf{s} \in \mathbf{X}$ and $W_{\boldsymbol{\pi}}^{*i} = Q_{\boldsymbol{\pi}^{-i}}^{*i}$.*

The proof of Theorem 13 can be found in Appendix C. The first part of Theorem 13 holds for any observation channel and any soft stationary joint policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$. In particular, it holds under mean-field, compressed, or local observability and even when $\boldsymbol{\pi}$ is not symmetric, and thus holds when player i faces a POMDP and not an MDP. As a result, the convergence of the iterates $\{\bar{Q}_t^i, \bar{J}_t^i\}_{t\geq 0}$ is not a simple consequence of celebrated results in stochastic approximation, such as those of (Tsitsiklis, 1994).

Remarks: In Appendix B, we explicitly define the functions $V_{\boldsymbol{\pi}}^{*i}$ and $W_{\boldsymbol{\pi}}^{*i}$. In (Kara and Yüksel, 2022), it is shown that the function $W_{\boldsymbol{\pi}}^{*i}$, restricted to $\{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\} \times \mathbb{A}$, is in fact the Q-function for *some* fully observed Markov decision problem with state space $\varphi^i(\mathbf{X}) := \{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$. The same argument used there can be used to establish that $V_{\boldsymbol{\pi}}^{*i}$, restricted to $\varphi^i(\mathbf{X})$, is the state value function for the *same* MDP. A complete specification of this MDP is deferred to Appendix B. We note:

(i) The MDP with state space $\varphi^i(\mathbf{X})$ described above is an instance of what (Kara and Yüksel, 2022) calls *an approximate belief MDP with memory length 0.*

(ii) Under Assumption 5, each soft policy $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_S$ gives rise to a unique invariant measure $\nu_{\tilde{\boldsymbol{\pi}}} \in \Delta(\mathbf{X})$, and we have that, for any $\nu \in \Delta(\mathbf{X})$, $\Pr_\nu^{\tilde{\boldsymbol{\pi}}}(\mathbf{x}_t \in \cdot) \to \nu_{\tilde{\boldsymbol{\pi}}}(\cdot)$ in total variation as $t \to \infty$. The remark after Assumption 5 says that it suffices to replace this convergence of laws by the convergence of their ergodic averages to $\nu_{\tilde{\boldsymbol{\pi}}}$.

(iii) Under compressed observability, the limiting quantities of Algorithm 1 depend, in general, on the policy π^i used by player i. This is in contrast to MDP settings (such as the specific cases in Parts 2 and 3 of Theorem 13), where the limiting values of Q-learning are the same for different (soft/sufficiently exploratory) policies.

(iv) In general, player i does not actually face the particular MDP on $\varphi^i(\mathbf{X}) \subseteq \mathbb{Y}$ described in item (iii). The limiting quantities $V_{\boldsymbol{\pi}}^{*i}$ and $W_{\boldsymbol{\pi}}^{*i}$ do not, in general, have any inherent relevance to player i's objective function in the game $\mathbf{G}$. These quantities should instead be interpreted as the subjective beliefs of player i, which were arrived at through a naive independent learning process.

4.3 Independent Learners and Subjectivity

In an n-player mean-field game, each cost minimizing player's true objective is to select a best-response to the policies of its counterparts. The solution concepts of (ϵ-) equilibrium, both perfect and with respect to an initial distribution, are then defined in terms of these objective functions. These definitions are inherently *objective* notions: optimality and equilibrium are defined without any approximation, estimation, or modelling of beliefs.

When studying the game $\mathbf{G}$ as an environment for decentralized multi-agent learning, one remarks that a given agent i may not be able to objectively verify whether a given policy π^i is an ϵ-best-response to a policy $\boldsymbol{\pi}^{-i}$ of its counterparts. Indeed, player i does not know the system data defining $\mathbf{G}$, does not know the joint policy $\boldsymbol{\pi}^{-i}$, and does not even observe the actions of other agents.[4] In this section, we formalize our notion of *subjective Q-equilibrium* and motivate it as a learning-relevant solution concept that accounts for decentralized information and independent learning in mean-field games.

Independent learners model their environment as an MDP and run single-agent algorithms, treating their local observations as though they are controlled Markovian state variables. This simplifying assumption is made to reduce the computational burden, so that tractable and scalable – if suboptimal – algorithms can be employed. By making a simplifying assumption on its environment, player i effectively adopts a *subjective model* of the system. Due to model uncertainty and decentralized information inherent to its environment, player i cannot verify whether a given policy is an objective ϵ-best-response to the behaviour of others. On the other hand, player i can check whether a given policy is *subjectively* ϵ-optimal for its assumed model, an MDP, using reinforcement learning techniques to estimate its (subjective) Q-function and value function.

Definition 14 (Subjective Function Family) *For each player $i \in \mathcal{N}$ and stationary joint policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$, let $V^{*i}_{\boldsymbol{\pi}}$ and $W^{*i}_{\boldsymbol{\pi}}$ be the functions appearing in Theorem 13, which are defined in Appendix B. We define two families of functions*

$$\mathcal{V}^* := \left\{V^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \boldsymbol{\Pi}_S\right\} \quad \text{and} \quad \mathcal{W}^* := \left\{W^{*i}_{\boldsymbol{\pi}} : \mathbb{Y} \times \mathbb{A} \to \mathbb{R} \middle| i \in \mathcal{N}, \boldsymbol{\pi} \in \boldsymbol{\Pi}_S\right\}$$

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

Definition 15 (Subjective Best-Responding) *Let $i \in \mathcal{N}$, $\epsilon \geq 0$, $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$. A policy $\pi^{*i} \in \Pi_S^i$ is called a $(\mathcal{V}^*, \mathcal{W}^*)$-*subjective ϵ-best-response to $\boldsymbol{\pi}^{-i}$ *if we have*

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

Definition 16 (Subjective (Q-) Equilibrium) *Let $\epsilon \geq 0$. A joint policy $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}_S$ is called a $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium for $\mathbf{G}$ if π^{*i} is a $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-best-response to $\boldsymbol{\pi}^{*-i}$ for every $i \in \mathcal{N}$.*

We will alternate between several terms for this solution concept: it will be called $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium when ambiguity is possible; it may be called subjective Q-equilibrium to distinguish it from earlier notions of subjective equilibrium; or it may simply be called subjective (ϵ-) equilibrium. Similarly, we will often refer to $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ($\epsilon$-) best-responding simply as subjective (ϵ-) best-responding.

The (possibly empty) set of $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-best-responses for player $i \in \mathcal{N}$ against stationary $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$ will be denoted by $\text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*) \subseteq \Pi_S^i$. Similarly, we let $\text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \subseteq \boldsymbol{\Pi}_S$ denote the (possibly empty) set of $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium policies for $\mathbf{G}$.

4. We reiterate that we wish to avoid the *representative agent* approach. Players are free to employ different policies. As a result, a given agent i will not know another agent j's policy.

4.4 Interpretations of Subjective Best-Responding and Equilibrium

Below, we describe some desirable qualities of our definition of subjective equilibrium.

Centrality of Local Information and Subjectivity: Independent learning agents make a simplifying assumption about their environment so as to obtain tractable, scalable algorithms. Since the independent learner in a mean-field game subjectively models its environment as an MDP, we have defined subjective ϵ-best-responding in analogy to a characterization of ϵ-optimality of MDPs (Lemma 31). That is, the definition of subjective best-responding is given with reference to the world model adopted by the learning agent, which has implications for the explanatory power of the definition.

Room for Asymmetry: In online, decentralized learning, different players observe different feedback trajectories, which will then be used to update their policies in an uncoordinated manner. As a result, in the absence of centralization or significant communication, it is unreasonable to expect the whole population to follow the same policy. The subjective equilibrium of Definition 16 allows for asymmetry of policies, and applies equally well to a heterogeneous population of agents, each using a distinct policy. Thus, it is suitable for online, decentralized learning, unlike various prior mean-field solution concepts.

Explanatory and Predictive Power: The merits and advantages of considering subjective notions of best-responding and equilibrium do not lie in their ability to approximate their objective analogs. Rather, these notions have merit because of their explanatory and predictive power when studying independent learners in decentralized settings. Independent learners engage in an iterative learning process to (subjectively) evaluate their policies. If the joint policy constitutes a subjective equilibrium, then the learning process and modelling assumption of each agent jointly ensure that the agent will assess itself to be behaving optimally. As a result, each agent will continue to use the policy it is already using. In this sense, this definition of subjective equilibrium is self-reinforcing and individually rational given one's subjective beliefs. Indeed, in the coming sections, we will show that simple independent learners converge to subjective equilibrium. That is, one sees that subjective equilibrium policies are stable points for independent learning, and stability may be desirable from a system design point of view. With this perspective, the subjective framework can serve to explain the advantages and shortcomings of existing works that apply single-agent algorithms to multi-agent settings.

4.5 Existence of Subjective Equilibrium

Lemma 17 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game with mean-field observability (Assumption 2), and suppose Assumption 5 holds. Let* $\epsilon > 0$*. Then, there exists a* $(\mathcal{V}^*, \mathcal{W}^*)$*-subjective* ϵ*-equilibrium. That is,* $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*.*

A proof of Lemma 17 can be found in Appendix D. Lemma 17 does not rule out the existence of *subjective-but-not-objective* ϵ-equilibrium policies, i.e. policies in $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \setminus \mathbf{\Pi}^{\epsilon\text{-eq}}$.

A result analogous to Lemma 17 holds when mean-field observability (Assumption 2) is strengthened to global observability (Assumption 1). We refrain from including this result, since it mirrors the result above. Moreover, under global observability, the notion of subjective equilibrium essentially coincides with objective equilibrium.

4.6 On Subjective Equilibrium under Compressed Observability

The proof techniques used to guarantee the existence of both objective and subjective equilibria under Assumption 1 and 2 do not apply to partially observed n-player mean-field games with the observation channels of compressed observability (Assumption 3) or local observability (Assumption 4). In the convergence analysis of the coming sections, we will assume the existence of subjective equilibria when one of Assumptions 3 or 4 holds.

Establishing existence of subjective equilibrium under compressed or local observability is an open problem, the resolution of which is closely related to other solution concepts in both n-player mean-field games and the $n \to \infty$ limit models. Under various names, several solution concepts have been proposed for mean-field games, and a number of approximation results relate these solutions concepts to one another. Some of these approximation results relate solution concepts for models with finitely many players to solution concepts for models with an infinite number of players, while others relate alternative solution concepts within a single model. For a sampling of works in this line, we cite (Weintraub et al., 2005, 2008; Adlakha et al., 2015; Saldi et al., 2018) and (Arslan and Yüksel, 2023) and the references therein.

Due to similarities between the model considered here and the model considered in (Arslan and Yüksel, 2023), it may be possible to modify the analysis of (Arslan and Yüksel, 2023, Theorem 4.1) and use that line of argument to establish the existence of subjective equilibrium when n is sufficiently large by relying on the existence of (objective) equilibrium in the associated limit model. In Appendix H of this paper, we present a result of the aforementioned type (Theorem 47). We also describe how this approximation result may be used to establish the existence of subjective ϵ-equilibrium in some partially observed n-player mean-field games.

5. Subjective Satisficing: Algorithm and Convergence Result

In this section, we present an independent learning algorithm suitable for decentralized learning in partially observed n-player mean-field games. This algorithm belongs to the paradigm of *win-stay, lose-shift algorithms*, in which agents assess the performance of a given policy by learning, and then update their policies in response to this assessment. This paradigm is popular in works on independent learners, as the structure is relatively simple and (subjectively) individually rational. Due to its win-stay, lose-shift structure, the algorithm we present in Section 5.4 is representative of a variety of independent learning algorithms. Analysis of the convergence behaviour of Algorithm 3 is, therefore, informative for interpreting or predicting the behaviour of independent learners and win-stay, lose-shift algorithms in large-scale, decentralized systems. We will show that Algorithm 3 drives policies to subjective ϵ-equilibrium, which illustrates the predictive and explanatory power of this new solution concept.

For clarity of presentation, this section studies games with mean-field observability. Analogous structural results (without learning) under global, compressed, and local observability are given in Appendix I. Analogous learning results under these observation channels are given in Appendix J.

5.1 Win-Stay, Lose-Shift Algorithms

Win-Stay, Lose-Shift Algorithms are a class of algorithms with the following form: players do not switch policies when they are satisfied with their performance (winning), but are free to experiment with other policies when they are unsatisfied (losing). This approach is based on the *satisficing* principle (Simon, 1956), which posits that a boundedly rational agent may tolerate some suboptimality in its performance. Thus, if an agent *thinks* that its performance is ϵ-optimal, it will not switch policies. Otherwise, if the agent thinks its performance is not ϵ-optimal, it may switch policies in an exploratory manner. In this work, we use the term *satisficing* synonymously with win-stay, lose-shift.

The overarching structure described above allows for different satisfaction criteria, and unifies a number of different MARL algorithms (Chien and Sinclair, 2011; Chasparis et al., 2013; Candogan et al., 2013; Posch, 1999; Yongacoglu et al., 2022). In what follows, we adopt the satisfaction criterion of subjective ϵ-best-responding, using the subjective functions of $(\mathcal{V}^*, \mathcal{W}^*)$ described in Section 4 and defined in Appendix B. We will focus on win-stay, lose-shift algorithms in which players select their successor policies from a subset of their stationary policies. In symbols, player i will select its policies from some set $\widehat{\Pi}^i \subset \Pi^i_S$. This restriction can be motivated either by bounded rationality or by player i's ability to represent policies with only finite precision. With these specifications in mind, the win-stay, lose-shift dynamics we consider take the following form: at time k, player $i \in \mathcal{n}$ selects its policy π^i_{k+1} with reference to $\boldsymbol{\pi}_k = (\pi^i_k, \boldsymbol{\pi}^{-i}_k)$ as follows:

$$\pi^i_{k+1} = \begin{cases} \pi^i_k, & \text{if } \pi^i_k \in \text{Subj-BR}^i_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \\ \tilde{\pi}^i \sim \lambda^i(\cdot|\boldsymbol{\pi}_k) & \text{else,} \end{cases}$$

where $\lambda^i(\cdot|\boldsymbol{\pi}_k)$ is a distribution over $\widehat{\Pi}^i$ that may depend on the joint policy $\boldsymbol{\pi}_k$.

5.2 A Subjective Satisficing Algorithm with Oracle Access

This section studies win-stay, lose-shift algorithms in a simplified setting, where an oracle provides each player with the relevant subjective functions for its policy update. A full learning algorithm is presented in Section 5.4. The complete learning algorithm can be interpreted as a noisy analog of the oracle algorithm below.

Algorithm 2 induces a time homogeneous Markov chain on $\widehat{\boldsymbol{\Pi}} = \times_{i\in\mathcal{n}}\widehat{\Pi}^i$. Policy $\boldsymbol{\pi}^* \in \widehat{\boldsymbol{\Pi}}$ is absorbing for this Markov chain $\iff$ it is a $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium. At this point, the set $\widehat{\boldsymbol{\Pi}}$ has not been explicitly defined, so the question of whether $\widehat{\boldsymbol{\Pi}}$ admits a subjective equilibrium requires examination. Appropriate selection of $\widehat{\boldsymbol{\Pi}}$ is discussed in the next subsection. Moreover, even if $\widehat{\boldsymbol{\Pi}}$ contains subjective equilibria, it is not clear *a priori* that such subjective equilibria are accessible from particular non-equilibrium joint policies by a process of policy adjustment in which satisfied agents do not change their policies. This consideration leads to the following definitions. For the following definitions, let $\mathbf{G}$ be a partially observed n-player mean-field game, let $i \in \mathcal{n}$, let $\epsilon \geq 0$, and recall that $(\mathcal{V}^*, \mathcal{W}^*)$ denotes the subjective function family for $\mathbf{G}$ and was defined in Definition 14.

Definition 18 *A sequence of policies $(\boldsymbol{\pi}_k)_{k\geq 0}$ in $\boldsymbol{\Pi}_S$ is called a* subjective ϵ-satisficing path *if, for every $i \in \mathcal{n}$ and $k \geq 0$, we have*

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

Algorithm 2: Subjective ϵ-satisficing Policy Revision (for player $i \in \mathcal{n}$)

Set Parameters
  $e^i \in (0,1)$: experimentation probability when not subjectively ϵ-best-responding
  $\widehat{\Pi}^i \subset \widehat{\Pi}^i_S$: a subset of stationary policies.
Initialize $\pi^i_0 \in \widehat{\Pi}^i$: initial policy
for $k \geq 0$ (k^{th} policy update)
  Receive $V^{*i}_{\boldsymbol{\pi}_k}$ and $W^{*i}_{\boldsymbol{\pi}_k}$ from oracle
  if $V^{*i}_{\boldsymbol{\pi}_k}(y) \leq \min_{a^i} W^{*i}_{\boldsymbol{\pi}_k}(y, a^i) + \epsilon$ *for all* $y \in \mathbb{Y}$ **then**
    $\pi^i_{k+1} = \pi^i_k$
  else
    $\pi^i_{k+1} \sim (1 - e^i)\delta_{\pi^i_k} + e^i \text{Unif}(\widehat{\Pi}^i)$

Intuitively, when agents jointly update their policies along a subjective ϵ-satisficing path, an agent only switches its policy when it (subjectively) deems the policy to be performing poorly. No further restrictions are placed on how an agent is allowed to switch (or not switch) its policy when it is subjectively unsatisfied.

Definition 19 *Let* $\widehat{\boldsymbol{\Pi}} \subseteq \boldsymbol{\Pi}_S$. *The game* $\mathbf{G}$ *is said to have the* subjective ϵ-satisficing paths property in $\widehat{\boldsymbol{\Pi}}$ *if the following holds: for every* $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$, *there exists a subjective* ϵ*-satisficing path* $(\boldsymbol{\pi}_k)_{k \geq 0}$ *such that* (i) $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, (ii) $\boldsymbol{\pi}_k \in \widehat{\boldsymbol{\Pi}}$ *for all* $k \geq 0$, *and* (iii) *for some* $K < \infty$, $\boldsymbol{\pi}_K \in \text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$.

This defines a subjective analog of the *objective* satisficing paths theory first proposed in (Yongacoglu et al., 2023) for Markov games, also studied in (Yongacoglu et al., 2024a) and (Yongacoglu et al., 2024b). We now describe how one may use *quantization* to select the policy subset $\widehat{\boldsymbol{\Pi}}$ such that the game $\mathbf{G}$ has the subjective ϵ-satisficing paths property within $\widehat{\boldsymbol{\Pi}}$.

5.3 Quantization of the Policy Space

In this section, we describe the quantization/discretization of policy sets. We argue that if the restricted set of policies $\widehat{\boldsymbol{\Pi}}$ is obtained via a sufficiently fine quantization of the original set $\boldsymbol{\Pi}_S$, then the performance loss for agent i will be negligible if it optimizes over $\widehat{\Pi}^i$ instead of Π^i_S. Moreover, we argue that if the restricted subset of policies is suitably fine, symmetric, and soft (Definition 29), then it will contain a subjective ϵ-equilibrium and, further, satisficing dynamics will drive policies to such subjective equilibria from any initial joint policy.

For the following definitions, let $\mathbf{G}$ be a partially observed n-player mean-field game with mean-field observability, and let $i \in \mathcal{n}$. Recall that d^i is a metric on the set Π^i_S.

Definition 20 *Let $\xi > 0$ and $\tilde{\Pi}^i \subseteq \Pi^i_S$. A mapping $q^i : \tilde{\Pi}^i \to \tilde{\Pi}^i$ is called a* ξ-quantizer (on $\tilde{\Pi}^i$) *if* (1) $q^i(\tilde{\Pi}^i) := \{q^i(\pi^i) : \pi^i \in \tilde{\Pi}^i\}$ *is a finite set and* (2) $d^i(\pi^i, q^i(\pi^i)) < \xi$ *for all* $\pi^i \in \tilde{\Pi}^i$.

Definition 21 *Let $\xi > 0$ and let $\tilde{\Pi}^i \subseteq \Pi^i_S$. A set of policies $\widehat{\Pi}^i \subseteq \tilde{\Pi}^i$ is called a* ξ-*quantization of $\tilde{\Pi}^i$ if $\widehat{\Pi}^i = q^i(\tilde{\Pi}^i)$, where q^i is some ξ-quantizer on $\tilde{\Pi}^i$.*

A set $\widehat{\Pi}^i \subseteq \Pi^i_S$ is called a quantization of Π^i_S if it is a ξ-quantization of Π^i_S for some $\xi > 0$. A quantization $\widehat{\Pi}^i$ is called *soft* if each policy $\pi^i \in \widehat{\Pi}^i$ is soft in the sense of Definition 29. The expression "fine quantization" will be used to reflect that a policy subset is a ξ-quantization for suitably small ξ. We extend these definitions and terminological conventions to also cover joint policies. For instance, $\widehat{\mathbf{\Pi}} = \times_{i \in \mathcal{n}} \widehat{\Pi}^i \subset \mathbf{\Pi}_S$ is a ξ-quantization of $\mathbf{\Pi}_S$ if each $\widehat{\Pi}^i$ is a ξ-quantization of Π^i_S, and so on.

Definition 22 *Let $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ be a quantization of $\mathbf{\Pi}_S$. We say that $\widehat{\mathbf{\Pi}}$ is* symmetric *if $\widehat{\Pi}^i = \widehat{\Pi}^j$ for each $i, j \in \mathcal{n}$.*

Lemma 23 *Let $\mathbf{G}$ be an n-player mean-field game with mean-field observability (Assumption 2) and suppose Assumption 5 holds. Let $\epsilon > 0$. There exists $\xi = \xi(\epsilon) > 0$ such that if $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ is any soft, symmetric ξ-quantization of $\mathbf{\Pi}_S$, then we have*

1) $\mathbf{\Pi}^{\epsilon\text{-eq}} \cap \widehat{\mathbf{\Pi}} \neq \varnothing$ and $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \widehat{\mathbf{\Pi}} \neq \varnothing$, and

2) $\mathbf{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property in $\widehat{\mathbf{\Pi}}$.

The first part can be seen using the proof of Lemma 17. The proof of the second part resembles that of (Yongacoglu et al., 2023, Theorem 3.6), which considered n-player symmetric stochastic games with full (global) state observability. One modification is needed: Corollary 2.9 of (Yongacoglu et al., 2023) must be replaced by a partial observations analog, Lemma 39, whose statement and proof is given in Appendix E

Lemma 23 guarantee that the game $\mathbf{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within finely quantized subsets of policies. This has two desirable consequences for algorithm design purposes. First, players can restrict their policy search from an uncountable set (all stationary policies) to a finite subset of policies with only a small loss in performance. Second, since the game $\mathbf{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property within $\widehat{\mathbf{\Pi}}$, play can be driven to (subjective) ϵ-equilibrium by changing only the policies of those players that are "ϵ-unsatisfied," so to speak. 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 the joint policy space $\mathbf{\Pi}_S$: since $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths to equilibrium exist within $\widehat{\mathbf{\Pi}}$ and $\widehat{\mathbf{\Pi}}$ is finite, play can be driven to subjective ϵ-equilibrium even by random policy updating by those players that are not subjectively ϵ-best-responding. Moreover, this also removes the need for specialized policy update rules, such as inertial best-responding (Arslan and Yüksel, 2017), that take into account special structure in the game.

The preceding remarks are formalized in Lemma 24, below.

Lemma 24 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumptions 2 and 5, and let* $\epsilon > 0$*. Let* $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ *be a quantization of* $\mathbf{\Pi}_S$ *and suppose* $\widehat{\mathbf{\Pi}}$ *satisfies (1)* $\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*; (2)* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for all* $i, j \in \mathcal{N}$*; and (3) every policy* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$ *is soft.*

Suppose that each agent $i \in \mathcal{N}$ *updates is policy sequence* $\{\pi_k^i\}_{k \geq 0}$ *according to Algorithm 2 and that, for each* $k \geq 0$*, the policy updates for* $\boldsymbol{\pi}_{k+1}$ *are conditionally independent across agents given* $\boldsymbol{\pi}_k$*. Then,* $\lim_{k \to \infty} \Pr(\boldsymbol{\pi}_k \in \widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)) = 1$.

The proof of Lemma 24 can be found in Appendix E. The proof crucially leverages the subjective satisficing structure described in the second part of Lemma 23.

Remark: The choice to update the policy according to $\pi_{k+1}^i \sim (1 - e^i)\delta_{\pi_k^i} + e^i \mathrm{Unif}(\widehat{\Pi}^i)$, in Line 10 of Algorithm 2, was somewhat arbitrary. In particular, the choice to uniformly mix over $\widehat{\Pi}^i$ with probability $e^i > 0$ was made to ensure the paths to equilibrium exist in the Markov chain of Lemma 24. The choice to remain with one's old policy with probability $1 - e^i$ was arbitrarily picked for ease of exposition, and can be replaced by any suitable distribution over $\widehat{\Pi}^i$ according to the taste of the system designer. Some choices may include gradient descent projected back onto $\widehat{\Pi}^i$ or selecting a best-response to $\boldsymbol{\pi}_k^{-i}$ within $\widehat{\Pi}^i$. Such changes may result in a significant speed-up of convergence to equilibrium when they are well-suited to the underlying game, but the guarantee holds in any case due to the uniform randomization.

5.4 An Independent Learning Algorithm for n-Player Mean-Field Games

We now present Algorithm 3, a win-stay, lose-shift algorithm that does not rely on non-local information and does not require the many agents of the system to use identical policies during learning. As such, Algorithm 3 is suitable for online, decentralized learning applications in large-scale systems, where agents have a limited understanding of the system.

At a high level, Algorithm 3 is a noise-perturbed, learning-based analog of Algorithm 2. Rather than receiving subjective function information from an oracle, here the subjective functions are learned using system feedback. The learned subjective functions are then used to estimate whether the agent is (subjectively) ϵ-best-responding.

Learning with Mean-Field State Information

We now present a result on the convergence behaviour of Algorithm 3 under mean-field observability. Analogous results for other observation channels can be found in Appendix J.

Under mean-field observability, policy iterates obtained from Algorithm 3 drive play to $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-equilibrium with high probability. In order to state this result formally in Theorem 25, we now fix $\epsilon > 0$ and make the following assumptions on the various parameters of Algorithm 3.

Assumption 6 *Fix* $\epsilon > 0$*. Let* $\widehat{\mathbf{\Pi}}$ *be a fine quantization of* $\mathbf{\Pi}_S$ *satisfying:* (1) $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for all* $i, j \in \mathcal{N}$*;* (2) $\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*;* (3) *For any* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$*,* $\boldsymbol{\pi}$ *is soft.*

Next we present a restriction on the parameters $\{d^i\}_{i \in \mathcal{N}}$. 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

Algorithm 3: Independent Learning

Set Parameters
  $\widehat{\Pi}^i \subset \Pi^i_S$: a fine quantization of Π^i_S
  $\{T_k\}_{k\geq 0}$: a sequence in $\mathbb{N}$ of "exploration 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 $\pi^i_0 \in \widehat{\Pi}^i$ (arbitrary), $\widehat{Q}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{A}}$, $\widehat{J}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}}$
for $k \geq 0$ (k^{th} exploration phase)
  for $t = t_k, t_k+1, \ldots, t_{k+1}-1$
    Observe $y^i_t = \varphi^i(\mathbf{x}_t)$
    Select $a^i_t \sim \pi^i_k(\cdot|y^i_t)$
    Observe $c^i_t := c(x^i_t, \mu(\mathbf{x}_t), a^i_t)$ and y^i_{t+1}
    Set $n^i_t = \sum_{\tau=t_k}^{t} \mathbf{1}\{(y^i_\tau, a^i_\tau) = (y^i_t, a^i_t)\}$
    Set $m^i_t = \sum_{\tau=t_k}^{t} \mathbf{1}\{y^i_\tau = y^i_t\}$
    $\widehat{Q}^i_{t+1}(y^i_t, a^i_t) = \left(1 - \frac{1}{n^i_t}\right) \widehat{Q}^i_t(y^i_t, a^i_t) + \frac{1}{n^i_t}\left[c^i_t + \gamma \min_{\tilde{a}^i} \widehat{Q}^i_t(y^i_{t+1}, \tilde{a}^i)\right]$
    $\widehat{J}^i_{t+1}(y^i_t) = \left(1 - \frac{1}{m^i_t}\right) \widehat{J}^i_t(y^i_t) + \frac{1}{m^i_t}\left[c^i_t + \gamma \widehat{J}^i_t(y^i_{t+1})\right]$
  if $\widehat{J}^i_{t_{k+1}}(y) \leq \min_{a^i} \widehat{Q}^i_{t_{k+1}}(y, a^i) + \epsilon + d^i \ \forall y \in \mathbb{Y}$, **then**
    $\pi^i_{k+1} = \pi^i_k$
  else
    $\pi^i_{k+1} \sim (1-e^i)\delta_{\pi^i_k} + e^i \mathrm{Unif}(\widehat{\Pi}^i)$
  Reset $\widehat{J}^i_{t_{k+1}} = 0 \in \mathbb{R}^{\mathbb{Y}}$ and $\widehat{Q}^i_{t_{k+1}} = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{A}}$

small, so that poorly performing policies are not mistaken for subjective ϵ-best-responses. The bound $\bar{d}_{\mathrm{MF}}$ below is defined analogous to the term $\bar{d}$ in Yongacoglu et al. (2023) and depends on both ϵ and $\widehat{\mathbf{\Pi}}$.

Assumption 7 *For all $i \in \mathcal{n}$, $d^i \in (0, \bar{d}_{\mathrm{MF}})$, where $\bar{d}_{\mathrm{MF}} = \bar{d}_{\mathrm{MF}}(\epsilon, \widehat{\mathbf{\Pi}})$ is specified in Appendix G.*

Theorem 25 *Let $\mathbf{G}$ be a partially observed n-player mean-field game satisfying Assumptions 2 and 5, and let $\epsilon > 0$. Suppose the policy set $\widehat{\mathbf{\Pi}}$ and the tolerance parameters $\{d^i\}_{i\in\mathcal{n}}$ satisfy Assumptions 6 and 7, and suppose all players follow Algorithm 3. For any $\xi > 0$, there exists $\tilde{T} = \tilde{T}(\xi, \epsilon, \widehat{\mathbf{\Pi}}, \{d^i\}_{i\in\mathcal{n}})$ such that if $T_k \geq \tilde{T}$ for all k, then*

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

for all sufficiently large k.

A proof of Theorem 25 is available in Appendix G. This result does not rule out the possibility that play settles at a subjective equilibrium that is not an objective equilibrium.

That is, it is possible that $\left(\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)\right) \setminus \mathbf{\Pi}^{\epsilon\text{-eq}} \neq \varnothing$, and that Algorithm 3 drives play to such a policy.

Remark: The guarantee of Theorem 25 continues to hold even if $\delta_{\pi_k^i}$ is replaced by any transition kernel in $\mathscr{P}(\widehat{\Pi}^i | \widehat{\Pi}^i \times \mathbb{R}^{\mathbb{Y}} \times \mathbb{R}^{\mathbb{Y}\times\mathbb{A}})$, where the distribution over $\widehat{\Pi}^i$ depends on π_k^i, $\widehat{J}_{t_{k+1}}^i$, and $\widehat{Q}_{t_{k+1}}^i$. Furthermore, each agent $i \in \mathcal{N}$ may use a different transition kernel for this update, allowing for heterogeneity in the learning dynamics.

Learning Under Other Observation Channels

Analogous convergence results can be stated under global observability (Theorem 57), compressed observability, or local observability. The latter two cases are described in Theorem 58, where the convergence result requires an additional assumption on the existence of subjective ϵ-equilibrium.

6. Simulation Study

We ran Algorithm 3 on a 20-player mean-field game with compressed observability, described below in (12). This game can be interpreted as a model for decision-making during an epidemic or as a model for vehicle use decisions in a traffic network.

The game $\mathbf{G}$ used for our simulation is given by the following list:

$$\mathbf{G} = (\mathcal{N}, \mathbb{X}, \mathbb{Y}, \mathbb{A}, \{\varphi^i\}_{i\in\mathcal{N}}, P_{\mathrm{loc}}, c, \gamma, \nu_0,). \tag{12}$$

The components of $\mathbf{G}$ are as follows. $\mathcal{N} = \{1, 2, \ldots, 20\}$ is a set of 20 players. The local state of an agent is interpreted as the agent's overall health, and takes values in $\mathbb{X} = \{\text{bad, medium, good}\}$. The agent selects between actions in $\mathbb{A} = \{\text{go, wait, heal}\}$. The action "go" corresponds to proceeding with one's usual activities, "wait" corresponds to avoiding activities with high risk of disease transmission, and "heal" corresponds to seeking healthcare. Each agent receives a signal that the proportion of agents in bad condition is either "low" or "high", i.e. $\mathbb{Y} = \mathbb{X}\times\{\text{low, high}\}$. For each $i \in \mathcal{N}$ and $\mathbf{s} \in \mathbf{X}$, player i's observation is given by $\varphi^i(\mathbf{s}) = (s^i, f(\mu(\mathbf{s})))$, where for any $\nu \in \Delta(\mathbb{X})$, $f(\nu)$ is

$$f(\nu) = \begin{cases} \text{low} & \text{if } \nu(\text{bad}) < 0.35, \\ \text{high} & \text{if } \nu(\text{bad}) \geq 0.35. \end{cases}$$

The stage cost function $c : \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A} \to \mathbb{R}$ is given by

$$c(s, \nu, a) := -R_{\text{go}} \cdot \mathbf{1}\{a = \text{go}\} + R_{\text{bad}} \cdot \mathbf{1}\{s = \text{bad}\} + R_{\text{heal}} \cdot \mathbf{1}\{a = \text{heal}\}$$

for all $(s, \nu, a) \in \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A}$, where $R_{\text{go}} = 5$ is a reward for undertaking one's usual business, $R_{\text{bad}} = 10$ is a penalty for being in bad condition, and $R_{\text{heal}} = 3$ is the cost of seeking healthcare. We note that the stage cost $c(s, \nu, a)$ depends only on (s, a), and the only strategic coupling in $\mathbf{G}$ is through coupled state dynamics.

The discount factor $\gamma = 0.8$, and $\nu_0 \in \Delta(\mathbf{X})$ is the product uniform distribution: $x_0^i \sim \mathrm{Unif}(\mathbb{X})$ for each $i \in \mathcal{N}$ and the random variables $\{x_0^i\}_{i\in\mathcal{N}}$ are jointly independent.

We now describe the transition kernel P_{loc}, which captures the state transition probabilities as a function of one's local state, the mean-field state, and one's local action. First,

for any $s \in \mathbb{X}$ and $\nu \in \Delta(\mathbb{X})$, we have $P_{\text{loc}}(\cdot|s, \nu, \text{wait}) = \delta_{\{s\}}$. That is, when a player waits, its local state is left unchanged with probability 1. For each $s \in \mathbb{X}$ and $\nu \in \Delta(\mathbb{X})$, we summarize $P_{\text{loc}}(\cdot|s, \nu, \text{heal})$ and $P_{\text{loc}}(\cdot|s, \nu, \text{go})$ in the tables below:

	$P_{\text{loc}}(\text{bad}\|s, \nu, \text{heal})$	$P_{\text{loc}}(\text{medium}\|s, \nu, \text{heal})$	$P_{\text{loc}}(\text{good}\|s, \nu, \text{heal})$
$s = \text{bad}$	0.25	0.5	0.25
$s = \text{medium}$	0.0	0.25	0.75
$s = \text{good}$	0.0	0.0	1.0

To succinctly describe $P_{\text{loc}}(\cdot|s, \nu, \text{go})$ for each $s \in \mathbb{X}$ and $\nu \in \Delta(\mathbb{X})$, we introduce a function to describe the likelihood of minor condition degradation (from good to medium or medium to bad) as a function of $\nu(\text{bad})$:

$$m(\nu) := \frac{\exp(12[\nu(\text{bad}) - 0.3])}{1 + \exp(12[\nu(\text{bad}) - 0.3])}$$

We also encode a fixed 0.15 chance of severe condition deterioration playing "go", independent of the mean-field term. This leads to the following transition probabilities:

s	$P_{\text{loc}}(\text{bad}\|s, \nu, \text{go})$	$P_{\text{loc}}(\text{medium}\|s, \nu, \text{go})$	$P_{\text{loc}}(\text{good}\|s, \nu, \text{go})$
bad	1.0	0.0	0.0
medium	$0.15 + 0.85 \cdot m(\nu)$	$1 - (0.15 + 0.85 \cdot m(\nu))$	0.0
good	0.15	$0.85 \cdot m(\nu)$	$1 - (0.15 + 0.85 \cdot m(\nu))$

As the number of agents in bad condition increases, the risk of degrading one's own condition when playing "go" also increases, slowly at first, then abruptly as the proportion of agents approaches and surpasses a threshold quantity.

The model data in (12) has been chosen to encourage the following splitting behaviour: an agent in bad condition should always seek healthcare; an agent in good condition should always play "go"; and an agent in medium condition should play "wait" when it observes that the proportion of agents in the bad state is high, while it should play "go" otherwise.

This behaviour is not a dominant strategy in the game theoretic sense: if, for some reason, a large number of agents play "go" when in bad condition, then an agent in medium condition does not gain anything by waiting, since the proportion of agents in bad condition will remain high. In this circumstance, the agent may prefer to either risk playing "go" or to simply play "heal". As such, the dynamics to subjective equilibrium (if it exists) are non-trivial.

Using the game in (12), we ran 250 independent trials of self-play under Algorithm 3, were each trial consisted of 20 exploration phases and each exploration phase consisted for 25,000 stage games. Our chosen parameters were $\epsilon = 5$, $d^i \equiv 1.5$, and $e^i \equiv 0.25$.

We define $\tilde{\Pi} \subset \mathscr{P}(\mathbb{A}|\mathbb{Y})$ to be the following finite subset of stationary policies:

$$\tilde{\Pi} := \{\tilde{\pi} \in \mathscr{P}(\mathbb{A}|\mathbb{Y}) : 10^2 \cdot \tilde{\pi}(a|y) \in \mathbb{Z}\,,\ \forall y \in \mathbb{Y}, a \in \mathbb{A}\}.$$

That is, policies in $\tilde{\Pi}$ can be described using only two places after the decimal point in each component probability distribution. We then define a quantizer quant : $\mathscr{P}(\mathbb{A}|\mathbb{Y}) \to \mathscr{P}(\mathbb{A}|\mathbb{Y})$ as

$$\text{quant}(\pi) = \begin{cases} \pi, & \text{if } \pi \text{ is 0.025-soft} \\ 0.9 \cdot \pi + 0.1 \cdot \pi_{\text{uniform}} & \text{otherwise,} \end{cases}$$

Exploration Phase Index k	$\frac{1}{250}\sum_{\tau=1}^{250} \mathbf{1}\{\boldsymbol{\pi}_{k,\tau} \in \text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)\}$
$k = 0$	0.0
$k = 5$	0.576
$k = 10$	0.74
$k = 15$	0.816
$k = 19$	0.844

Table 1: Frequency of Subjective ϵ-equilibrium. $\boldsymbol{\pi}_{k,\tau}$ denotes the policy for EP k during the τ^{th} trial.

where π_{uniform} is the policy such that $\pi_{\text{uniform}}(\cdot|y) = \text{Unif}(\mathbb{A})$ for each $y \in \mathbb{Y}$. In our simulation, players select policies from the finite set $\widehat{\Pi} := \text{quant}(\tilde{\Pi})$.

Our results are summarized in Figures 2 and 3 and in Table 1. In Figure 2, we plot the frequency of subjective ϵ-equilibrium against the exploration phase index. For our purposes, a policy at a given exploration phase is considered a subjective ϵ-equilibrium if all 20 players perceive themselves to be ϵ-best-responding, given their subjective state and action functions. We observe that the frequency steadily rises to over 84% by the 20th exploration phase.

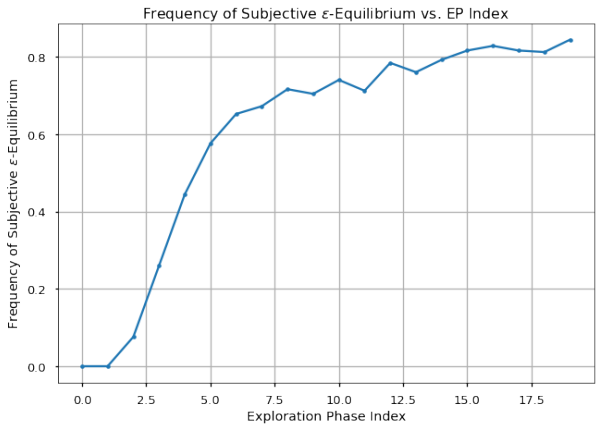

Figure 2: Frequency of subjective ϵ-equilibrium plotted against the exploration phase index, averaged over 250 trials.

In Figure 3, we plot two curves. The blue curve reports the mean number of agents, out of 20, that are subjectively ϵ-best-responding at a given exploration phase index. This quantity quickly rises to, and remains at, 19.75, reflecting that a large majority of agents are ϵ-satisfied even when the system is not at subjective ϵ-equilibrium. Figure 3 also reports the number of agents using the quantized version of $\pi_{\text{sensible}} \in \mathscr{P}(\mathbb{A}|\mathbb{Y})$, defined for each (x, σ) pair as follows: $\pi_{\text{sensible}}(\cdot|\text{good, low}) = \pi_{\text{sensible}}(\cdot|\text{good, high}) = \delta_{\text{go}}$ (i.e. uniformly playing "go" when in good condition); $\pi_{\text{sensible}}(\cdot|\text{bad, low}) = \pi_{\text{sensible}}(\cdot|\text{bad, high}) = \delta_{\text{heal}}$

(i.e. uniformly playing "heal" when in bad condition); $\pi_{\text{sensible}}(\cdot|\text{medium, low}) = \delta_{\text{go}}$ while $\pi_{\text{sensible}}(\cdot|\text{medium, high}) = \delta_{\text{wait}}$.

The orange curve in Figure 3 reports the mean number of agents following the policy $\text{quant}(\pi_{\text{sensible}})$. Interestingly, this curve lags below the blue curve, reflecting that play often settled at an asymmetric joint policy that forms a subjective ϵ-equilibrium.

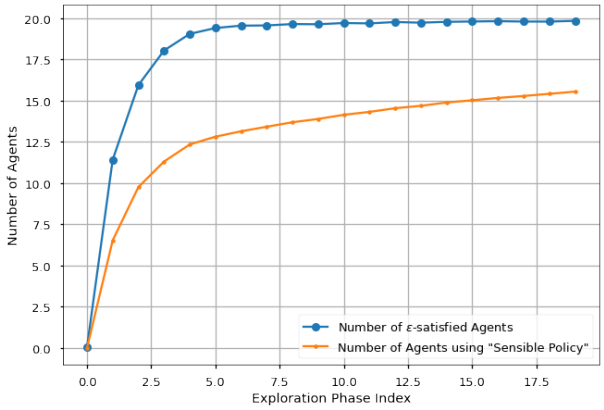

Figure 3: Mean Number of players who are subjectively ϵ-best-responding vs. exploration phase index, averaged over 250 trials.

7. Discussion

We now consider situations in which compressed observability (Assumption 3) may be the most appropriate model for decentralized information. It is natural to begin by asking how a player's observations are actually obtained. (Note that the actual observation channel encountered need not be the same as the observation channel used in the model of the game.) We envision three (actual) observation channels as being the most plausible.

In the first scenario, agents obtain readings on the global state through local sensors. Such an (actual) observation channel is truly decentralized and would result in a limited view of the overall system, giving rise to compressed observability. In this case, the actual observation channel may be used as the model's observation channel. In a related second scenario, agents supplement local sensor readings by communicating with neighbours. Here, too, one may naturally take the model's observation channel to be the same as the actual observation channel, and Assumption 3 may be more appropriate than Assumption 2.

In a third scenario, a centralized entity monitors the global state and broadcasts a compressed signal about the mean-field to the agents. This is the case, for example, in vehicle routing. Here, local states are locations in a traffic network, and action selection corresponds to selection of a path to one's destination. Agents may rely on a satellite navigation system to locate themselves in the network and to identify which paths are congested. If

the navigation system reports congestion using a tiered system of low-, moderate-, and high-congestion roads, then compressed observability is the relevant observation channel.

In addition to applications where the actual observation channel results in a limited view of the overall system, compressed observability may also arise in systems with rich (actual) observation channels if players voluntarily choose to discard information in their learning process. When the number of players, n, and the number of local states, $|\mathbb{X}|$, are both moderately large, the set of empirical measures $\text{Emp}_n = \{\mu(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$ becomes unwieldy for the purposes of (tabular) learning. As such, it is reasonable to expect that naive independent learning agents will employ some form of function approximation, with quantization of Emp_n offering the simplest form of function approximation. Moreover, a common compression scheme may be shared by all agents in such an application, perhaps as the result of some shared "conventional wisdom" about the system. Building on the ideas of (Patil et al., 2024), the question of whether non-tabular algorithms can learn well-performing, history-dependent policies involving compressed mean-field observations is an interesting direction for future research.

8. Conclusion

In this paper, we considered partially observed n-player mean-field games from the point of view of decentralized independent learners. Independent learning is characterized by ignoring the presence of other strategic agents in the system, treating one's environment as if it were a single-agent MDP, and naively running single-agent learning algorithms to select one's policy. We studied the convergence of naive single-agent learning iterates in the game setting, proving almost sure convergence under mild assumptions. By analogy to near-optimality criteria for MDPs, we used the limiting values of these learning iterates to develop a notion of subjective ϵ-equilibrium. After establishing the existence of objective perfect equilibrium as well as subjective ϵ-equilibrium under mean-field observability, we extended the notion of (objective) ϵ-satisficing paths of Yongacoglu et al. (2023) to subjective value functions. In this framework, we studied the structural properties of n-player mean-field games and showed that subjective ϵ-satisficing paths to subjective ϵ-equilibrium exist under various information structures for partially observed n-player mean-field games.

Apart from the structural and conceptual contributions described above, we have also presented Algorithm 3, a decentralized independent learner for playing partially observed n-player mean-field games, and we have argued that Algorithm 3 drives policies to subjective ϵ-equilibrium under self-play. Unlike the bulk of results on learning in mean-field games, the convergence guarantees of Algorithm 3 do not mandate that players use the same policy at a given time or that they use the same policy update rule to switch policies at the end of an exploration phase. As such, our algorithm is capable of describing learning dynamics for a population of homogeneous agents that may arrive at a joint policy consisting of heterogeneous policies. The learning dynamics presented here result in system stability, in that policies settle to a particular joint policy, but the emergent behaviour need not be an objective equilibrium. Convergence to non-equilibrium policies may arise in real-world strategic environments, and our notion of subjective equilibrium may present an interpretation for such real-world stability in some instances.

This paper leaves open the question of whether subjective ϵ-equilibrium always exist in mean-field games with compressed or local observability. Determining sufficient conditions for the existence of subjective equilibrium is an interesting topic for future research.

References

Sachin Adlakha, Ramesh Johari, and Gabriel Y. Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Learning mean-field games with discounted and average costs. *Journal of Machine Learning Research*, 24(17):1–59, 2023a.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023b.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

Ari Arapostathis, Anup Biswas, and Johnson Carroll. On solutions of mean field games with ergodic cost. *Journal de Mathématiques Pures et Appliquées*, 107(2):205–251, 2017.

Gürdal Arslan and Serdar Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2017.

Gürdal Arslan and Serdar Yüksel. Subjective equilibria under beliefs of exogenous uncertainty for dynamic games. *SIAM Journal on Control and Optimization*, 61(3):1038–1062, 2023.

Dario Bauso, Hamidou Tembine, and Tamer Başar. Robust mean field games with application to production of an exhaustible resource. *IFAC Proceedings Volumes*, 45(13): 454–459, 2012.

Anup Biswas. Mean field games with ergodic cost for discrete time Markov processes. *arXiv preprint arXiv:1510.08968*, 2015.

Theophile Cabannes, Mathieu Laurière, Julien Perolat, Raphael Marinier, Sertan Girgin, Sarah Perrin, Olivier Pietquin, Alexandre M. Bayen, Eric Goubault, and Romuald Elie. Solving N-player dynamic routing games with congestion: A mean-field approach. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1557–1559, 2022.

Luciano Campi and Markus Fischer. Correlated equilibria and mean field games: A simple model. *Mathematics of Operations Research*, 47(3):2240–2259, 2022.

Ozan Candogan, Asuman Ozdaglar, and Pablo A. Parrilo. Near-potential games: Geometry and dynamics. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–32, 2013.

Siddharth Chandak, Pratik Shah, Vivek S. Borkar, and Parth Dodhia. Reinforcement learning in non-Markovian environments. *Systems & Control Letters*, 185:105751, 2024.

Georgios C. Chasparis, Ari Arapostathis, and Jeff S. Shamma. Aspiration learning in coordination games. *SIAM J. Control and Optimization*, 51(1):465–490, 2013.

Geoffroy Chevalier, Jerome Le Ny, and Roland Malhamé. A micro-macro traffic model based on mean-field games. In *2015 American Control Conference (ACC)*, pages 1983–1988. IEEE, 2015.

Steve Chien and Alistair Sinclair. Convergence to approximate Nash equilibria in congestion games. *Games and Economic Behavior*, 71(2):315–327, 2011.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, pages 746–752, 1998.

Anne Condon. On algorithms for simple stochastic games. *Advances in Computational Complexity Theory*, 13:51–72, 1990.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5527–5540, 2020.

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.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 23(1):11627–11680, 2022.

Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7143–7150, 2020.

Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 28(1):89–93, 1964.

Markus Fischer. On the connection between symmetric n-player games and mean field games. *The Annals of Applied Probability*, 27(2):757–810, 2017.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Dean Foster and H. Peyton Young. Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1:341–367, 2006.

Roy Fox, Stephen M. McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

Diogo A. Gomes and João Saúde. Mean field games models: A brief survey. *Dynamic Games and Applications*, 4:110–154, 2014.

Diogo A. Gomes, Roberto M. Velho, and Marie-Therese Wolfram. Socio-economic applications of finite state mean field games. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028), 2014.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

Junling Hu and Michael P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(Nov):1039–1069, 2003.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Minyi Huang, Peter E. Caines, and Roland P. Malhamé. Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9):1560–1571, 2007.

Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

Ehud Kalai and Ehud Lehrer. Rational learning leads to Nash equilibrium. *Econometrica*, pages 1019–1045, 1993.

Ehud Kalai and Ehud Lehrer. Subjective games and equilibria. *Games and Economic Behavior*, 8(1):123–163, 1995.

Ali D. 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, 2022.

Ali D. Kara and Serdar Yüksel. Q-learning for stochastic control under general information structures and non-Markovian environments. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Daniel Lacker. Mean field games via controlled martingale problems: Existence of Markovian equilibria. *Stochastic Processes and their Applications*, 125(7):2856–2894, 2015.

Daniel Lacker. On the convergence of closed-loop Nash equilibria to the mean field game limit. *The Annals of Applied Probability*, 30(4):1693 – 1761, 2020.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. Learning mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022.

Kiyeob Lee, Desik Rengarajan, Dileep Kalathil, and Srinivas Shakkottai. Reinforcement learning for mean field games with strategic complementarities. In *International Conference on Artificial Intelligence and Statistics*, pages 2458–2466. PMLR, 2021.

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.

Jian Li, Rajarshi Bhattacharyya, Suman Paul, Srinivas Shakkottai, and Vijay Subramanian. Incentivizing sharing in realtime d2d streaming networks: A mean field game perspective. *IEEE/ACM Transactions on Networking*, 25(1):3–17, 2016.

Jian Li, Bainan Xia, Xinbo Geng, Hao Ming, Srinivas Shakkottai, Vijay Subramanian, and Le Xie. Mean field games in nudge systems for societal networks. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 3(4):1–31, 2018.

Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference*, pages 983–994, 2019.

Michael L. Littman. Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning*, volume 1, pages 322–328, 2001.

Michael L. Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *International Conference on Machine Learning*, volume 96, pages 310–318, 1996.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Mayank Manjrekar, Vinod Ramaswamy, Vamseedhar Reddyvari Raja, and Srinivas Shakkottai. A mean field game approach to scheduling in cellular systems. *IEEE Transactions on Control of Network Systems*, 7(2):568–578, 2019.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Coordination of independent learners in cooperative Markov games. *HAL preprint hal-00370889*, 2009.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *Knowledge Engineering Review*, 27(1):1–31, 2012.

David Mguni, Joel Jennings, and Enrique Munoz de Cote. Decentralised learning in systems with many, many strategic agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

David Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*, pages 7688–7699. PMLR, 2021.

Rajesh Mishra, Sriram Vishwanath, and Deepanshu Vasal. Model-free reinforcement learning for mean field games. *IEEE Transactions on Control of Network Systems*, 2023.

Paul Muller, Romuald Elie, Mark Rowland, Mathieu Laurière, Julien Perolat, Sarah Perrin, Matthieu Geist, Georgios Piliouras, Olivier Pietquin, and Karl Tuyls. Learning correlated equilibria in mean-field games. *arXiv preprint arXiv:2208.10138*, 2022.

John H. Nachbar. Beliefs in repeated games. *Econometrica*, 73(2):459–480, 2005.

Gandharv Patil, Aditya Mahajan, and Doina Precup. On learning history-based policies for controlling Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3511–3519. PMLR, 2024.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9413–9421, 2022.

Martin Posch. Win–stay, lose–shift strategies for repeated games—Memory length, aspiration levels and noise. *Journal of Theoretical Biology*, 198(2):183–195, 1999.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Rabih Salhab, Jerome Le Ny, and Roland P Malhamé. A mean field route choice game model. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1005–1010. IEEE, 2018.

Sina Sanjari, Naci Saldi, and Serdar Yüksel. Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers. *Mathematics of Operations Research*, 2022.

Muhammed O. Sayin, Kaiqing Zhang, David Leslie, Tamer Başar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

Muhammed O. Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization*, 60(4):2095–2114, 2022.

Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 426–431, 1994.

Herbert A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129, 1956.

Amit Sinha, Matthieu Geist, and Aditya Mahajan. Periodic agent-state based Q-learning for POMDPs. *Advances in Neural Information Processing Systems*, 2024.

Leonardo Stella, Fabio Bagagiolo, Dario Bauso, and Giacomo Como. Opinion dynamics and stubbornness through mean-field games. In *52nd IEEE Conference on Decision and Control*, pages 2519–2524. IEEE, 2013.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 251–259, 2019.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(1):483–565, 2022a.

Sriram G. Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. Decentralized mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9439–9447, 2022b.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337, 1993.

J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

Onur Unlu and Muhammed O. Sayin. Episodic Logit-Q dynamics for efficient learning in stochastic teams. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1985–1990. IEEE, 2023.

Deepanshu Vasal. Sequential decomposition of discrete-time mean-field games. *Dynamic Games and Applications*, pages 1–19, 2023.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pages 10092–10103. PMLR, 2020.

Christopher Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

Ermo Wei and Sean Luke. Lenient learning in independent-learner stochastic cooperative games. *Journal of Machine Learning Research*, 17(1):2914–2955, 2016.

Gabriel Y. Weintraub, C. Lanier Benkard, and Benjamin Van Roy. Oblivious equilibrium: A mean field approximation for large-scale dynamic games. *Advances in Neural Information Processing Systems*, 18, 2005.

Gabriel Y. Weintraub, C. Lanier Benkard, and Benjamin Van Roy. Markov perfect industry dynamics with many firms. *Econometrica*, 76(6):1375–1411, 2008.

Bainan Xia, Srinivas Shakkottai, and Vijay Subramanian. Small-scale markets for a bilateral energy sharing economy. *IEEE Transactions on Control of Network Systems*, 6(3):1026–1037, 2019.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR, 2023.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information. *IEEE Transactions on Automatic Control*, 67(10):5230–5245, 2022.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Satisficing paths and independent multiagent reinforcement learning in stochastic games. *SIAM Journal on Mathematics of Data Science*, 5(3):745–773, 2023.

Bora Yongacoglu, Gürdal Arslan, Lacra Pavel, and Serdar Yüksel. Generalizing better response paths and weakly acyclic games. *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024a.

Bora Yongacoglu, Gurdal Arslan, Lacra Pavel, and Serdar Yuksel. Paths to equilibrium in games. *Advances in Neural Information Processing Systems*, 2024b.

Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2278–2284. IEEE, 2020a.

Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Approximate equilibrium computation for discrete-time linear-quadratic mean-field games. In *2020 American Control Conference (ACC)*, pages 333–339. IEEE, 2020b.

Muhammad Aneeq Uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Başar. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*, pages 10178–10206. PMLR, 2023.

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.

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*, 35:1923–1935, 2022.

Appendices

Appendix A. Fully and Partially Observed Markov Decision Problems

A finite, partially observed Markov decision problem (POMDP) with the discounted cost criterion is given by a list $\mathcal{m}$:

$$\mathcal{m} = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, C, P_{\mathcal{m}}, \phi, \gamma, \nu_0)\,. \tag{13}$$

At time $t \in \mathbb{Z}_{\geq 0}$, the system's state is denoted X_t and takes values in the finite set $\mathcal{X}$, with $X_0 \sim \nu_0$. An observation Y_t taking values in the finite set $\mathcal{Y}$ is generated according to a noisy reading of X_t through the observation channel $\phi \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, as $Y_t \sim \phi(\cdot|X_t)$. The agent uses its observable history variable, $\hbar_t^{\mathrm{ob}}$, to be defined shortly, to select its action A_t from the finite set $\mathcal{A}$. The agent then incurs a stage cost C_t, given by the cost function C as $C_t := C(X_t, A_t)$. The system's state subsequently evolves according to $X_{t+1} \sim P_{\mathcal{m}}(\cdot|X_t, A_t)$, where $P_{\mathcal{m}} \in \mathcal{P}(\mathcal{X}|\mathcal{X} \times \mathcal{A})$. A discount factor $\gamma \in (0,1)$ is used to discount the sequence of costs incurred by the agent.

We define the *system history sets* $\{\mathcal{H}_t\}_{t \in \mathbb{Z}_{\geq 0}}$ as follows:

$$\mathcal{H}_t := (\mathcal{X} \times \mathcal{Y} \times \mathcal{A})^t \times \mathcal{X} \times \mathcal{Y}, \quad \forall t \geq 0.$$

Elements of $\mathcal{H}_t$ are called *system histories of length* t. We define a random quantity $\hbar_t$, taking values in $\mathcal{H}_t$ and given by

$$\hbar_t = (X_0, Y_0, A_0, \cdots, X_{t-1}, Y_{t-1}, A_{t-1}, X_t, Y_t),$$

and we use $\hbar_t$ to denote the t^{th} *system history variable*. To capture the information actually observed by the agent controlling the system, we define *observable history sets* $\{\mathcal{H}_t^{\mathrm{ob}}\}_{t \geq 0}$ as

$$\mathcal{H}_t^{\mathrm{ob}} := \Delta(\mathcal{X}) \times (\mathcal{Y} \times \mathcal{A} \times \mathbb{R})^t \times \mathcal{Y}, \quad \forall t \geq 0.$$

Elements of $\mathcal{H}_t^{\mathrm{ob}}$ are called *observable histories of length* t, and we let

$$\hbar_t^{\mathrm{ob}} = (\nu_0, Y_0, A_0, C_0, \cdots, C_{t-1}, Y_t),$$

denote the $\mathcal{H}_t^{\mathrm{ob}}$-valued random quantity representing the t^{th} *observable history variable*.

Definition 26 *A sequence $\pi = (\pi_t)_{t \geq 0}$ such that $\pi_t \in \mathcal{P}(\mathcal{A}|\mathcal{H}_t^{\mathrm{ob}})$ for each t is called a* policy *for the POMDP $\mathcal{m}$.*

We denote the set of all policies for the POMDP $\mathcal{m}$ by $\Pi_{\mathcal{m}}$. Fixing a policy $\pi \in \Pi_{\mathcal{m}}$ and an initial measure $\nu \in \Delta(\mathcal{X})$ induces a unique probability measure $\Pr_\nu^\pi$ on the trajectories of play, i.e. on sequences $(X_t, Y_t, A_t, C_t)_{t=0}^\infty$. We let E_ν^π denote the expectation associated with $\Pr_\nu^\pi$ by define the agent's objective function as

$$\mathcal{J}_{\mathcal{m}}(\pi, \nu) := E_\nu^\pi \left[\sum_{t \geq 0} \gamma^t C(X_t, A_t) \right], \quad \forall \pi \in \Pi_{\mathcal{m}}, \nu \in \Delta(\mathcal{X}).$$

In the special case that $\nu = \delta_s$ for some state $s \in \mathcal{X}$, we simply write $\mathcal{J}_{\mathcal{m}}(\pi, s)$.

Definition 27 (Optimal Policy) *For $\epsilon \geq 0$, a policy $\pi^* \in \Pi_{\mathcal{M}}$ is called* ϵ-optimal with respect to $\nu \in \Delta(\mathcal{X})$ *if*

$$\mathcal{J}_{\mathcal{M}}(\pi^*, \nu) \leq \inf_{\pi \in \Pi_{\mathcal{M}}} \mathcal{J}_{\mathcal{M}}(\pi, \nu) + \epsilon.$$

If π^* is ϵ-optimal with respect to every $\nu \in \Delta(\mathcal{X})$, then π^* is called uniformly ϵ-optimal.

If $\epsilon = 0$ in the preceding definition, we simply refer to π^* as being optimal, either uniformly or with respect to a given initial distribution ν.

Definition 28 (Stationary Policies) *A policy $\pi \in \Pi_{\mathcal{M}}$ is called* (memoryless) stationary (or simply stationary) *if, for some $g \in \mathcal{P}(\mathcal{A}|\mathcal{Y})$, the following holds: for any $t \geq 0$ and $\tilde{h}_t^{\text{ob}} = (\tilde{\nu}, \tilde{Y}_0, \ldots, \tilde{Y}_t) \in \mathcal{H}_t^{\text{ob}}$, we have*

$$\pi_t(\cdot|\tilde{h}_t^{\text{ob}}) = g(\cdot|\tilde{Y}_t).$$

We let $\Pi_{\mathcal{M},S}$ denote the set of stationary policies for the POMDP $\mathcal{M}$.

Definition 29 (Soft Policies) *For $\xi > 0$, a policy $\pi \in \Pi_{\mathcal{M}}$ is called* ξ-soft *if, for any $t \geq 0$ and $\tilde{h}_t^{\text{ob}} \in \mathcal{H}_t^{\text{ob}}$, we have $\pi(a|\tilde{h}_t^{\text{ob}}) \geq \xi$ for all $a \in \mathcal{A}$. A policy $\pi \in \Pi_{\mathcal{M}}$ is called* soft *if it is ξ-soft for some $\xi > 0$.*

The goal for an agent controlling the POMDP $\mathcal{M}$ is to find an optimal policy.

Remark: We have chosen to present a model in which costs are measurable functions of the state and action. One can also allow for costs to be random variables with expectation given by the cost function C. Since our objective involves an expectation, this does not change the ensuing analysis, and we opt for the simpler model.

One can also generalize the POMDP model above (or the MDP model below) to allow the transition/observation probabilities and cost function to additionally depend on time. Such models may be called *time inhomogeneous*. By contrast, the models presented here are *time homogeneous*, as the transition/observation probabilities and cost function do not vary across time. Any reference to a POMDP (or fully observed MDP) in this article shall be understood to refer to a time homogeneous POMDP (or MDP).

A.1 Fully Observed Markov Decision Problems

Definition 30 (MDP) *A fully observed Markov decision problem (or simply an MDP) is a POMDP for which $\mathcal{X} = \mathcal{Y}$ and $\phi(\cdot|s) = \delta_s$ for each state $s \in \mathcal{X}$.*

It is well-known that if $\mathcal{M}$ is an MDP, then there exists a (uniformly) optimal policy $\pi^* \in \Pi_{\mathcal{M},S}$, and furthermore it suffices to check for optimality along Dirac distributions $\{\delta_s : s \in \mathcal{X}\}$. Using the existence of an optimal policy $\pi^* \in \Pi_{\mathcal{M},S}$, we define the (optimal) Q-function for the MDP $\mathcal{M}$ as follows: $Q_{\mathcal{M}}^* : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is given by

$$Q_{\mathcal{M}}^*(s,a) := E_{\nu}^{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t C(X_t, A_t) \middle| X_0 = s, A_0 = a\right], \quad \forall (s,a) \in \mathcal{X} \times \mathcal{A},$$

where $\pi^* \in \Pi_{\mathcal{M},S}$ is an optimal policy for $\mathcal{M}$ and $\nu \in \Delta(\mathcal{X})$ is any initial state distribution.[5] One can show that, for any $s \in \mathcal{X}$, we have $\mathcal{J}_{\mathcal{M}}(\pi^*, s) = \min_{a \in \mathcal{A}} Q^*_{\mathcal{M}}(s, a)$.

Lemma 31 *Let $\pi \in \Pi_{\mathcal{M},S}$, $\epsilon \geq 0$. Then, π is ϵ-optimal for the MDP $\mathcal{M}$ if and only if*

$$\mathcal{J}_{\mathcal{M}}(\pi, s) \leq \min_{a \in \mathcal{A}} Q^*_{\mathcal{M}}(s, a) + \epsilon, \quad \forall s \in \mathcal{X}. \tag{14}$$

Under mild conditions on the MDP $\mathcal{M}$, the action value function $Q^*_{\mathcal{M}}$ can be learned iteratively using the Q-learning algorithm (Watkins, 1989). Similarly, for stationary policies $\pi \in \Pi_{\mathcal{M},S}$, the value function $\mathcal{J}_{\mathcal{M}}(\pi, \cdot)$ can be learned iteratively. That is, the agent can produce sequences of iterates $\{\widehat{J}_t, \widehat{Q}_t\}_{t \geq 0}$, with $\widehat{J}_t \in \mathbb{R}^{\mathcal{X}}$ and $\widehat{Q}_t \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ for each t, such that, almost surely,

$$\widehat{Q}_t \to Q^*_{\mathcal{M}} \quad \text{and} \quad \widehat{J}_t(s) \to \mathcal{J}_{\mathcal{M}}(\pi, s), \quad \forall s \in \mathcal{X}.$$

Thus, an agent controlling the MDP $\mathcal{M}$ using a stationary policy $\pi \in \Pi_{\mathcal{M},S}$ may use an estimated surrogate of the inequality of (14)—involving stochastic estimates of $Q^*_{\mathcal{M}}$ and $\mathcal{J}_{\mathcal{M}}(\pi, \cdot)$—as a stopping condition when searching for an ϵ-optimal policy. This idea will feature heavily in the subsequent sections. In particular, we will use an analogous condition for our definition of subjective best-responding. As a preview of forthcoming definitions, we now state a definition that is motivated by analogy to (14).

Definition 32 *Let $\widehat{J} \in \mathbb{R}^{\mathcal{X}}$ and $\widehat{Q} \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$. The function $\widehat{J}$ is said to be* subjectively ϵ-optimal with respect to $\widehat{Q}$ *if*

$$\widehat{J}(s) \leq \min_{a \in \mathcal{A}} \widehat{Q}(s, a) + \epsilon, \quad \forall s \in \mathcal{X}.$$

The intuition underlying Definition 32 is that $\widehat{J}(s)$ plays the role of $\mathcal{J}_{\mathcal{M}}(\pi, s)$ while $\widehat{Q}(s, a)$ plays the role of $Q^*_{\mathcal{M}}(s, a)$ in (14), with $\widehat{J}$ and $\widehat{Q}$ arising from a learning process. In employing such a comparison to test for ϵ-optimality of π, the agent replaces objective quantities with learned estimates that represent subjective knowledge obtained through system interaction and learning.

Appendix B. Implied MDPs and Subjective Functions

In this section, we explicitly characterize the subjective value functions of Theorem 13 in Section 4. These subjective value functions arise as the limit points of an independent learning process, in which each agent used a stationary policy during the learning process and obtained a sequence of stochastic iterates meant to approximate action values and state values of an assumed MDP. However, as described in Lemma 6, each agent does not face a fully observed MDP but rather a POMDP. To understand the limiting values of the learning process in Algorithm 1, we introduce *implied MDPs* associated with POMDPs. We then use the connection between n-player mean-field games and POMDPs to define the subjective functions of Theorem 13 using the language of implied MDPs.

5. The function $Q^*_{\mathcal{M}}$ is also called the action-value function for $\mathcal{M}$ and is usually defined as the fixed point of a Bellman operator. The brief presentation above invokes some well-known properties of MDPs.

B.1 Implied MDPs and Subjective Functions for Ergodic POMDPs

To begin, let $\mathcal{M} = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, C, P_{\mathcal{M}}, \phi, \gamma, \nu_0)$ be a POMDP as described in (13). Recall that $\Pi_{\mathcal{M},S}$ denotes the set of stationary policies for the POMDP $\mathcal{M}$, and is identified with the set $\mathcal{P}(\mathcal{A}|\mathcal{Y})$ of transition kernels on $\mathcal{A}$ given $\mathcal{Y}$. We now define several objects relevant to the analysis of learning algorithms for POMDPs.

Definition 33 *For $\lambda \in \Delta(\mathcal{X})$, the* backward channel $B^\lambda \in \mathcal{P}(\mathcal{X}|\mathcal{Y})$ *is a transition kernel on $\mathcal{X}$ given $\mathcal{Y}$ defined by*

$$B^\lambda(x|y) := \frac{\phi(y|x)\lambda(x)}{\sum_{s\in\mathcal{X}} \phi(y|s)\lambda(s)}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

One may interpret the quantity $B^\lambda(x|y)$ as the posterior probability that $x_0 = x$ given that $x_0 \sim \lambda$ and the first observation symbol was $y_0 = y$.

Definition 34 *For $\lambda \in \Delta(\mathcal{X})$, the* implied transition kernel $\mathcal{T}_\lambda \in \mathcal{P}(\mathcal{Y}|\mathcal{Y}\times\mathcal{A})$ *is given by*

$$\mathcal{T}_\lambda(y'|y,a) := \sum_{s'\in\mathcal{X}} \phi(y'|s') \sum_{s\in\mathcal{X}} \mathcal{T}(s'|s,a)B^\lambda(s|y), \quad \forall y, y' \in \mathcal{Y}, a \in \mathcal{A}.$$

The quantity $\mathcal{T}_\lambda(y'|y,a)$ represents the conditional probability that the observation $y_1 = y'$ given $y_0 = y, u_0 = u$ and $x_0 \sim \lambda$.

Definition 35 *For $\lambda \in \Delta(\mathcal{X})$, the* implied cost function $C_\lambda : \mathcal{Y}\times\mathcal{A} \to \mathbb{R}$ *is given by*

$$C_\lambda(y,a) := \sum_{x\in\mathcal{X}} C(x,a)B^\lambda(x|y), \quad \forall y \in \mathcal{Y}, a \in \mathcal{A}.$$

We will employ the following standing assumption for the remainder of this section.

Assumption 8 *For any policy $\pi \in \Pi_{\mathcal{M},S}$, there exists unique $\lambda_\pi \in \Delta(\mathcal{X})$ such that*

$$\lim_{t\to\infty} \Pr{}^\pi_{\lambda_0}(X_t \in \cdot) = \lambda_\pi(\cdot), \quad \forall \lambda_0 \in \Delta(\mathcal{X}).$$

In words, this assumption says that the law of the hidden state at time t, X_t, converges, as $t \to \infty$, to some invariant ergodic distribution $\lambda_\pi \in \Delta(\mathcal{X})$. The invariant distribution will depend on the policy π, but convergence to this distribution holds for any distribution $\lambda_0 \in \Delta(\mathcal{X})$ of the initial state X_0.

Remark: Assumption 8 is made for clarity of presentation, and in fact can be relaxed. For the results below, one can alternatively assume the following.

Assumption 8* *For any policy $\pi \in \Pi_{\mathcal{M},S}$, there exists unique $\lambda_\pi \in \Delta(\mathcal{X})$ such that*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \Pr{}^\pi_{\lambda_0}(X_t \in \cdot) = \lambda_\pi(\cdot), \quad \forall \lambda_0 \in \Delta(\mathcal{X}).$$

Under either Assumption 8 or Assumption 8*, each stationary policy is assigned a corresponding backward channel, implied transition kernel, and implied cost function: for each $\pi \in \Pi_{\mathcal{M},S}$, we say that B^{λ_π} is the backward channel associated with π, $\mathcal{T}_{\lambda_\pi}$ is the implied transition kernel associated with π, and C_{λ_π} is the implied cost unction associated with π.

For each $\pi \in \Pi_{\mathcal{M},S}$, we define an operator $f_\pi : \mathbb{R}^{\mathcal{Y}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{Y}\times\mathcal{A}}$, whose (y,a)-component mapping is given by

$$[f_\pi Q](y,a) := C_{\lambda_\pi}(y,a) + \gamma \sum_{y'\in\mathcal{Y}} \min_{a'\in\mathcal{A}} Q(y',a')\mathcal{T}_{\lambda_\pi}(y'|y,a), \quad \forall\, Q \in \mathbb{R}^{\mathcal{Y}\times\mathcal{A}}.$$

It is easily verified that f_π is a γ-contraction on $\mathbb{R}^{\mathcal{Y}\times\mathcal{A}}$ and therefore admits a unique fixed point, which we denote by $\mathcal{Q}^*_{\mathcal{M},\pi} \in \mathbb{R}^{\mathcal{Y}\times\mathcal{A}}$.

Definition 36 *The* implied Q-function *associated with π is defined as $\mathcal{Q}^*_{\mathcal{M},\pi}$.*

Structurally, one can see that the implied Q-function $\mathcal{Q}^*_{\mathcal{M},\pi}$ is the optimal Q-function for a suitably defined *implied MDP*, namely that with state space $\mathcal{Y}$, action space $\mathcal{A}$, stage cost function C_{λ_π}, and transition kernel $\mathcal{T}_{\lambda_\pi}$.

Remark: To be clear, we do not claim that using the greedy policy with respect to $\mathcal{Q}^*_{\mathcal{M},\pi}$ will yield meaningful performance guarantees for the POMDP $\mathcal{M}$.

In addition to the operator f_π, we define an operator $g_\pi : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}^{\mathcal{Y}}$, whose y-component mapping is given by

$$[g_\pi J](y) := \sum_{a\in\mathcal{A}} \pi(a|y) \left\{ C_{\lambda_\pi}(y,a) + \gamma \sum_{y'\in\mathcal{Y}} J(y')\mathcal{T}_{\lambda_\pi}(y'|y,a) \right\}, \quad \forall J \in \mathbb{R}^{\mathcal{Y}}.$$

As with f_π, one can verify that g_π is a contraction and admits a unique fixed point, which we denote by $\mathcal{J}^*_{\mathcal{M},\pi} \in \mathbb{R}^{\mathcal{Y}}$.

Definition 37 *The* implied value function *associated with π is defined as $\mathcal{J}^*_{\mathcal{M},\pi}$.*

One can interpret the function $\mathcal{J}^*_{\mathcal{M},\pi}$ as the state value function of the policy $\pi \in \Pi_{\mathcal{M},S}$ in the same implied MDP as before, with state space $\mathcal{Y}$, action space $\mathcal{A}$, stage cost function C_{λ_π}, and transition kernel $\mathcal{T}_{\lambda_\pi}$.

B.2 Implied MDPs and Subjective Functions in Partially Observed n-Player Mean-Field Games

Let $\mathbf{G} = (\mathcal{N}, \mathbb{X}, \mathbb{Y}, \mathbb{A}, \{\varphi^i\}_{i\in\mathcal{N}}, c, \gamma, P_{\mathrm{loc}}, \nu_0)$ be a partially observed n-player mean-field game, with the symbols inheriting their meanings from (1). By Lemma 6, we have that if $\boldsymbol{\pi}^{-i} \in \mathbf{\Pi}_S^{-i}$ is a stationary policy for the remaining players, then player i faces a POMDP $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$, with partially observed state process $\{\mathbf{x}_t\}_{t\geq 0}$ and observation process $\{y^i_t = \varphi^i(\mathbf{x}_t)\}_{t\geq 0}$.

Under Assumption 5, if $(\pi^i, \boldsymbol{\pi}^{-i}) \in \Pi^i_S \times \mathbf{\Pi}_S^{-i}$ is stationary, then the ergodicity condition of Assumption 8 holds for the POMDP $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$. Thus, for any stationary policy $\pi^i \in \Pi^i_S$, we may define the objects of the preceding section: that is, we may define backward channels, implied transition kernels, implied cost functions, implied Q-functions, and implied value functions associated with π in the POMDP $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$.

Definition 38 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumption 5. For each player* $i \in \mathcal{N}$ *and policy* $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Pi}_S^{-i}$*, let* $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$ *denote the POMDP faced by player* i*.*

For any $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i}) \in \Pi_S^i \times \boldsymbol{\Pi}_S^{-i}$*, the* subjective value function $V_{\boldsymbol{\pi}}^{*i}$ *is defined as the implied value function associated with* π^i *in the POMDP* $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$*. That is,* $V_{\boldsymbol{\pi}}^{*i} := \mathcal{J}^\star_{\mathcal{M}_{\boldsymbol{\pi}^{-i}}, \pi^i}$*.*

For any $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i}) \in \Pi_S^i \times \boldsymbol{\Pi}_S^{-i}$*, the* subjective Q-function $W_{\boldsymbol{\pi}}^{*i}$ *is defined as the implied Q-function associated with* π^i *in the POMDP* $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$*. That is,* $W_{\boldsymbol{\pi}}^{*i} := \mathcal{Q}^\star_{\mathcal{M}_{\boldsymbol{\pi}^{-i}}, \pi^i}$*.*

Appendix C. Proof of Theorem 13

By Lemma 6, player i faces a POMDP with observation sequence $\{y_t^i\}_{t\geq 0}$ and underlying state process $\{\mathbf{x}_t\}_{t\geq 0}$. By Assumption 5 and the assumed softness of $\boldsymbol{\pi}$, we have that all pairs $(\mathbf{s}, a^i) \in \mathbf{X} \times \mathbb{A}$ are visited infinitely often $\Pr_\nu^{\boldsymbol{\pi}}$-almost surely for any $\nu \in \Delta(\mathbf{X})$. We can therefore invoke (Kara and Yüksel, 2022, Theorem 4(i)) to establish the almost sure convergence of Q-factor iterates $\{\bar{Q}_t^i\}_{t\geq 0}$. The same analysis can be used to establish the convergence of the value function iterates $\{\bar{J}_t^i\}_{t\geq 0}$. This proves the first part.

Under Assumption 1, player i faces an MDP with state process $\{\mathbf{x}_t\}_{t\geq 0}$. Again by the softness of $\boldsymbol{\pi}$ and Assumption 5, each $(\mathbf{s}, a^i) \in \mathbf{X} \times \mathbb{A}$ is visited infinitely often almost surely. The convergence of Q-factors under these conditions is well-known; see for instance Tsitsiklis (1994). The same analysis can be used to prove that $\lim_{t\to\infty} \bar{J}_t^i(\mathbf{s}) = J^i(\boldsymbol{\pi}, \mathbf{s})$ almost surely for each $\mathbf{s} \in \mathbf{X}$, proving the second part.

The proof of the third part parallels that of the second part, replacing $\mathbf{x}_t$ by $y_t^i = (x_t^i, \mu(\mathbf{x}_t))$ and using Theorem 8 to see that player i faces an MDP in $\{y_t^i\}_{t\geq 0}$.

Appendix D. Proof of Lemma 17

For any $i \in \mathcal{N}$, the sets $\boldsymbol{\Pi}_{S,\mathrm{sym}}$ and $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$ are compact under the topologies induced by the metrics $\mathbf{d}_{\mathrm{sym}}$ and $\mathbf{d}_{\mathrm{sym}}^{-i}$, respectively. By compactness, Lemma 3, and Lemma 10, we see that the functions of Lemmas 3 and 10 are in fact uniformly continuous on $\boldsymbol{\Pi}_{S,\mathrm{sym}}$ and $\boldsymbol{\Pi}_{S,\mathrm{sym}}^{-i}$. It follows that there exists $\xi > 0$ such that if two joint policies $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \boldsymbol{\Pi}_{S,\mathrm{sym}}$ satisfy $\mathbf{d}_{\mathrm{sym}}(\boldsymbol{\pi}, \boldsymbol{\pi}') < \xi$, then

$$\left|J^i(\boldsymbol{\pi}, \mathbf{s}) - J^i(\boldsymbol{\pi}', \mathbf{s})\right| < \frac{\epsilon}{2} \quad \text{and} \quad \left|\min_{a^i\in\mathbb{A}} Q_{\boldsymbol{\pi}^{-i}}^{*i}(\varphi^i(\mathbf{s}), a^i) - \min_{a^i\in\mathbb{A}} Q_{\boldsymbol{\pi}'^{-i}}^{*i}(\varphi^i(\mathbf{s}), a^i)\right| < \frac{\epsilon}{2}, \quad (15)$$

for any $\mathbf{s} \in \mathbf{X}$, where $\varphi^i(\mathbf{s}) = (s^i, \mu(\mathbf{s}))$ due to Assumption 2. We fix $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}_S^{0\text{-eq}} \cap \boldsymbol{\Pi}_{\mathrm{sym}}$ to be a symmetric perfect equilibrium policy, which exists by Theorem 12. (That a *symmetric* perfect equilibrium exists can be seen from the proof of Theorem 12.) Let $\boldsymbol{\pi}_{\mathrm{soft}} \in \boldsymbol{\Pi}_{S,\mathrm{sym}}$ be a soft, symmetric joint policy satisfying $\mathbf{d}_{\mathrm{sym}}(\boldsymbol{\pi}^*, \boldsymbol{\pi}_{\mathrm{soft}}) < \xi$.

By part 3 of Theorem 13, since $\boldsymbol{\pi}_{\mathrm{soft}}$ is soft, we have that

$$W_{\boldsymbol{\pi}_{\mathrm{soft}}}^{*i} = Q_{\boldsymbol{\pi}_{\mathrm{soft}}^{-i}}^{*i} \quad \text{and} \quad V_{\boldsymbol{\pi}_{\mathrm{soft}}}^{*i}(\varphi^i(\mathbf{s})) = J^i(\boldsymbol{\pi}_{\mathrm{soft}}, \mathbf{s}), \quad \forall i \in \mathcal{N}, \mathbf{s} \in \mathbf{X}.$$

Combining this with (15), it follows that, for any $\mathbf{s} \in \mathbf{X}$ and $i \in \mathcal{N}$, we have $V_{\boldsymbol{\pi}_{\mathrm{soft}}}^{*i}(\varphi^i(\mathbf{s})) \leq \min_{a^i\in\mathbb{A}} W_{\boldsymbol{\pi}_{\mathrm{soft}}}^{*i}(\varphi^i(\mathbf{s}), a^i) + \epsilon$, which shows that $\boldsymbol{\pi}_{\mathrm{soft}} \in \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$. □

Appendix E. Proof of Lemma 24

To prove Lemma 24, we first need the following auxiliary result, which also appears in the proof of Lemma 23.

Lemma 39 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumptions 2 and 5 hold, and let* $\epsilon \geq 0$*. Let* $\boldsymbol{\pi} \in \boldsymbol{\Pi}_S$ *be soft. For* $i, j \in \mathcal{n}$*, suppose* π^i *and* π^j *are symmetric. Then, we have*

$$\pi^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*) \iff \pi^j \in \text{Subj-BR}^j_\epsilon(\boldsymbol{\pi}^{-j}, \mathcal{V}^*, \mathcal{W}^*).$$

Proof To prove this result, we argue that the subjective functions of players i and j satisfy $V^{*i}_{\boldsymbol{\pi}} = V^{*j}_{\boldsymbol{\pi}}$ and $W^{*i}_{\boldsymbol{\pi}} = W^{*j}_{\boldsymbol{\pi}}$ whenever π^i and π^j are symmetric and soft. In Section 4, we observed that, under Assumptions 2 and 5, the learned values $V^{*i}_{\boldsymbol{\pi}}$ and $W^{*i}_{\boldsymbol{\pi}}$ were in fact the state value and action value functions, respectively, for an MDP with state space $\{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\} \subseteq \mathbb{Y}$. An analogous remark holds for $V^{*j}_{\boldsymbol{\pi}}$ and $W^{*j}_{\boldsymbol{\pi}}$ and player j. The MDP in question was called an *approximate belief MDP* by Kara and Yüksel (2022), and can be characterized in terms of the implied MDP construction of Appendix B.

We will argue that the approximate belief MDP on $\{\varphi^i(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$ faced by player i is equivalent to the approximate belief MDP on $\{\varphi^j(\mathbf{s}) : \mathbf{s} \in \mathbf{X}\}$ faced by player j. Then, since player i and j's policies are symmetric, they correspond to the same stationary policy for this approximate belief MDP, and are therefore both either ϵ-optimal for that MDP or not; the result will follow.

To see that the approximate belief MDP facing player i is equivalent to that facing player j, we observe that the construction of Kara and Yüksel (2022) depends only on the stage cost function, the observation channel, and the unique invariant distribution on the underlying state space. The stage cost and observation channel are symmetric and shared by players i and j, so it suffices to show that the unique invariant distribution on $\mathbf{X}$, say $\nu_{\boldsymbol{\pi}}$, is symmetric in the following sense:

$$\nu_{\boldsymbol{\pi}}(\text{swap}_{ij}(\mathbf{s})) = \nu_{\boldsymbol{\pi}}(\mathbf{s}), \quad \forall \mathbf{s} \in \mathbf{X}, \tag{16}$$

where, for any $\mathbf{s} \in \mathbf{X}$, $\text{swap}_{ij}(\mathbf{s}) \in \mathbf{X}$ is the global state satisfying (1) $\text{swap}_{ij}(\mathbf{s})^p = s^p$ for each $p \in \mathcal{n} \setminus \{i, j\}$, and (2) we have $\text{swap}_{ij}(\mathbf{s})^j = s^i$ and $\text{swap}_{ij}(\mathbf{s})^i = s^j$.

To see that (16) holds, note that by Assumption 5, $\{\mathbf{x}_t\}_{t \geq 0}$ is an irreducible, aperiodic Markov chain on $\mathbf{X}$ under the soft policy $\boldsymbol{\pi}$. Thus, we have for any $\tilde{\nu} \in \Delta(\mathbf{X})$.

$$\Pr^{\boldsymbol{\pi}}_{\tilde{\nu}}(\mathbf{x}_t \in \cdot) \to \nu_{\boldsymbol{\pi}}(\cdot),$$

as $t \to \infty$, where $\nu_{\boldsymbol{\pi}}$ is the (unique) invariant measure on $\mathbf{X}$ induced by $\boldsymbol{\pi}$. In particular, putting the initial measure $\tilde{\nu} = \text{Unif}(\mathbf{X})$, we have that for each $t \geq 0$ and each $\mathbf{s} \in \mathbf{X}$,

$$\Pr^{\boldsymbol{\pi}}_{\tilde{\nu}}(\mathbf{x}_t = \mathbf{s}) = \Pr^{\boldsymbol{\pi}}_{\tilde{\nu}}(\mathbf{x}_t = \text{swap}_{ij}(\mathbf{s})).$$

It follows that

$$\nu_{\boldsymbol{\pi}}(\mathbf{s}) := \lim_{t \to \infty} \Pr^{\boldsymbol{\pi}}_{\tilde{\nu}}(\mathbf{x}_t = \mathbf{s}) = \lim_{t \to \infty} \Pr^{\boldsymbol{\pi}}_{\tilde{\nu}}(\mathbf{x}_t = \text{swap}_{ij}(\mathbf{s})) =: \nu_{\boldsymbol{\pi}}(\text{swap}_{ij}(\mathbf{s})).$$

From this, it follows that the approximate belief MDP on $\mathbb{Y}$ faced by player i is the same as the approximate belief MDP on $\mathbb{Y}$ faced by player j. ■

Proof of Lemma 24

We have that $\{\boldsymbol{\pi}_k\}_{k\geq 0}$ is a time homogeneous Markov chain on $\widehat{\boldsymbol{\Pi}}$. For any subjective ϵ-equilibrium $\boldsymbol{\pi}^* \in \widehat{\boldsymbol{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$, we have that the singleton $\{\boldsymbol{\pi}^*\}$ is an absorbing set for this Markov chain. By part 2 of Lemma 23, the game $\mathbf{G}$ has the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing paths property in $\widehat{\boldsymbol{\Pi}}$. For any $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$, let $L_{\boldsymbol{\pi}} < \infty$ denote the length of a shortest $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing path within $\widehat{\boldsymbol{\Pi}}$ that starts at $\boldsymbol{\pi}$ and terminates at a policy in $\widehat{\boldsymbol{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$. Call such a policy $\boldsymbol{\pi}^*$ and note it depends on $\boldsymbol{\pi}$. We also define $p_{\boldsymbol{\pi}} > 0$ as the probability the Markov chain follows this path when starting at $\boldsymbol{\pi}$, i.e.

$$p_{\boldsymbol{\pi}} := \Pr(\boldsymbol{\pi}_{L_{\boldsymbol{\pi}}} = \boldsymbol{\pi}^* | \boldsymbol{\pi}_0 = \boldsymbol{\pi}) > 0, \quad \forall \boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}.$$

Define $L := \max\{L_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}\}$ and $\hat{p} := \min\{p_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}\} > 0$. For any $m \geq 0$, we have

$$\Pr\left(\bigcap_{j=1}^{m} \{\boldsymbol{\pi}_{jL} \notin \widehat{\boldsymbol{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)\}\right) \leq (1-\hat{p})^m.$$

Taking $m \to \infty$ gives the result.

Appendix F. Approximation Results on the Sequences of Learning Iterates

Remark: The contents of this and the next section closely resembles that of (Yongacoglu et al., 2023, Appendix A). The proof technique used here parallels the proof technique of (Yongacoglu et al., 2023, Theorem 5.1).

In the coming sections, we prove that Algorithm 3 leads to the convergence of joint policies as described in Theorems 25, 57, and 58. Since the evolution of the policy process $\{\boldsymbol{\pi}_k\}_{k\geq 0}$ depends on the evolution of the learning iterates $\{\widehat{J}_t^i, \widehat{Q}_t^i\}_{t\geq 0, i\in\mathcal{n}}$, we begin by studying the convergence behaviour of these iterates. We argue that if parameters are suitably selected, then these learning iterates sampled at the end of each exploration phase will closely approximate the subjective functions for that exploration phase, and consequently the policy process of Algorithm 3 approximates the policy process of the Markov chain resulting from Algorithm 2.

We note that when each agent $i \in \mathcal{n}$ uses Algorithm 3, it is actually following a particular randomized, non-stationary policy. When all agents use Algorithm 3, we use Pr (with no policy index in the superscript and optional initial distribution in the subscript, e.g. $\Pr_\nu$ for some $\nu \in \Delta(\mathbf{X})$) to denote the resulting probability measure on trajectories of states and actions. For all other policies $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}$, we continue to use $\Pr_\nu^{\tilde{\boldsymbol{\pi}}}$, as before.

The policy process $\{\boldsymbol{\pi}_k\}_{k\geq 0}$ depends on the sequences $\{\widehat{J}_t^i, \widehat{Q}_t^i\}_{t\geq 0, i\in\mathcal{n}}$ only through these sequences sampled at the end of exploration phases; that is, the iterate sequences are relevant to the updating of policies only along the subsequence of times $\{t_k\}_{k\geq 0}$. Recall that we used $\{\bar{Q}_t^i\}_{t\geq 0}$ and $\{\bar{J}_t^i\}_{t\geq 0}$ to denote the naively learned stochastic iterates obtained when player $i \in \mathcal{n}$ employed Algorithm 1 and all players followed a soft stationary policy. We now analyze the sequences $\{\widehat{Q}_t^i\}_{t\geq 0}$ and $\{\widehat{J}_t^i\}_{t\geq 0}$ by comparison to the sequences $\{\bar{Q}_t^i\}_{t\geq 0}$ and $\{\bar{J}_t^i\}_{t\geq 0}$.

The sequences $\{\widehat{Q}_t^i\}_{t\geq 0}$ and $\{\bar{Q}_t^i\}_{t\geq 0}$ are related through the Q-factor update. There are, however, two major differences. First, Algorithm 3 instructs player i to reset its counters

at the end of the k^{th} exploration phase (i.e. after the update at time t_{k+1}, before the update at time $t_{k+1}+1$), meaning the step sizes differ for the two iterate sequences $\{\bar{Q}^i_t\}_{t\geq 0}$ and $\{\widehat{Q}^i_t\}_{t\geq 0}$ even when the state-action-cost trajectories observed by player i are identical. Second, Algorithm 3 instructs player i to reset its Q-factors at the end of the k^{th} exploration phase, while Algorithm 1 does not involve any resetting.

Consequently, one sees that the process $\{\widehat{Q}^i_t\}_{t\geq 0}$ depends on the initial condition $\widehat{Q}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{A}}$, the state-action trajectory, and the resetting times $\{t_k\}_{k\geq 0}$. In contrast, the process $\{\bar{Q}^i_t\}_{t\geq 0}$ depends only on the initial value $\bar{Q}^i_0 = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{A}}$ and the state-action trajectory. Analogous remarks hold relating $\{\widehat{J}^i_t\}_{t\geq 0}$ and $\{\bar{J}^i_t\}_{t\geq 0}$.

Recall: exploration phase k begins with the stage game at t_k and ends before the stage game at $t_{k+1} = t_k + T_k$. During exploration phase k, the sequences $\{\widehat{Q}^i_t\}_{t=t_k}^{t_k+T_k}$ and $\{\widehat{J}^i_t\}_{t=t_k}^{t_k+T_k}$ depend only on the state-action trajectory $\mathbf{x}_{t_k}, \mathbf{a}_{t_k}, \cdots, \mathbf{a}_{t_k+T_k-1}, \mathbf{x}_{t_k+T_k}$. This leads to the following useful observation:

$$\begin{aligned}
&\Pr\left(\{\mathbf{x}_{t_k+T_k} = \mathbf{s}_{T_k}\} \bigcap_{j=0}^{T_k-1} \{\mathbf{x}_{t_k+j} = \mathbf{s}_j, \mathbf{a}_{t_k+j} = \tilde{\mathbf{a}}_j\} \middle| \mathbf{x}_{t_k} = \mathbf{s}, \boldsymbol{\pi}_k = \boldsymbol{\pi}\right) \\
=&\Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\{\mathbf{x}_{T_k} = \mathbf{s}_{T_k}\} \bigcap_{j=0}^{T_k-1} \{\mathbf{x}_j = \mathbf{s}_j, \mathbf{a}_j = \tilde{\mathbf{a}}_j\}\right),
\end{aligned}$$

holds for any $(\mathbf{s}_0, \mathbf{s}_1, \cdots, \mathbf{s}_{T_k}) \in \mathbf{X}^{T_k+1}$ and $(\tilde{\mathbf{a}}_0, \cdots, \tilde{\mathbf{a}}_{T_k-1}) \in \mathbf{A}^{T_k}$, where $\mathbf{A} = \times_{i\in\mathcal{N}}\mathbb{A}$.

In words, once players following Algorithm 3 select a policy $\boldsymbol{\pi}$ for the k^{th} exploration phase, then the conditional probabilities of the trajectories restricted to time indices in that exploration phase can be described by $\Pr^{\boldsymbol{\pi}}$, with the indices of the events suitably shifted to start at time 0. This leads to a series of useful lemmas, which we include below for completeness.

Lemma 40 *For any $i \in \mathcal{N}, \boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}, k \geq 0$ and global state $\mathbf{s} \in \mathbf{X}$, we have*

$$\Pr\left(\widehat{Q}^i_{t_{k+1}} \in \cdot \middle| \boldsymbol{\pi}_k = \boldsymbol{\pi}, \mathbf{x}_{t_k} = \mathbf{s}\right) = \Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\bar{Q}^i_{T_k} \in \cdot\right)$$

and

$$\Pr\left(\widehat{J}^i_{t_{k+1}} \in \cdot \middle| \boldsymbol{\pi}_k = \boldsymbol{\pi}, \mathbf{x}_{t_k} = \mathbf{s}\right) = \Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\bar{J}^i_{T_k} \in \cdot\right).$$

Combining Lemma 40 with Theorem 13, we get the following result.

Lemma 41 *For any joint policy $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$, global state $\mathbf{s} \in \mathbf{X}$, and player $i \in \mathcal{N}$, we have*

1. $\Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\lim_{t\to\infty} \bar{Q}^i_t = W^{*i}_{\boldsymbol{\pi}}\right) = 1.$
2. $\Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\lim_{t\to\infty} \bar{J}^i_t = V^{*i}_{\boldsymbol{\pi}}\right) = 1.$
3. *For any $\xi > 0$, there exists $T = T(i, \boldsymbol{\pi}, \xi) \in \mathbb{N}$ such that*

$$\Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\sup_{t\geq T} \left\|\bar{Q}^i_t - W^*_{\boldsymbol{\pi}}\right\|_\infty < \xi\right) \geq 1-\xi, \text{ and } \Pr^{\boldsymbol{\pi}}_{\mathbf{s}}\left(\sup_{t\geq T} \left\|\bar{J}^i_t - V^*_{\boldsymbol{\pi}}\right\|_\infty < \xi\right) \geq 1-\xi$$

Finally, we combine Lemmas 40 and 41 to obtain the following useful result on conditional probabilities.

Lemma 42 *Let $k, \ell \in \mathbb{Z}_{\geq 0}$ and suppose $k \leq \ell$. Let $\mathcal{F}_k$ denote the σ-algebra generated by the random variables $\boldsymbol{\pi}_k$ and $\mathbf{x}_{t_k}$. For any $\xi > 0$, there exists $T = T(\xi) \in \mathbb{N}$ such that if $T_\ell \geq T$, then* Pr*-almost surely, we have*

$$\Pr\left(\bigcap_{i\in\mathcal{n}} \left\{\left\|\widehat{Q}^i_{t_{\ell+1}} - W^{*i}_{\boldsymbol{\pi}_\ell}\right\|_\infty < \xi\right\} \cap \left\{\left\|\widehat{J}^i_{t_{\ell+1}} - V^{*i}_{\boldsymbol{\pi}_\ell}\right\|_\infty < \xi\right\} \middle| \mathcal{F}_k\right) \geq 1 - \xi.$$

Appendix G. Proof of Theorem 25

We begin by introducing the quantity $\bar{d}_{\mathrm{MF}}$, which will serve as the upper bound for the tolerance parameters $d^i, i \in \mathcal{n}$. The quantity $\bar{d}_{\mathrm{MF}}$, depends on both $\epsilon > 0$ and the subset of policies $\widehat{\boldsymbol{\Pi}} \subset \boldsymbol{\Pi}_S$ as follows: $\bar{d}_{\mathrm{MF}} := \min \mathcal{O}_{\mathrm{MF}}$, where $\mathcal{O}_{\mathrm{MF}} := S_{\mathrm{MF}} \setminus \{0\}$, and S_{MF} is given by

$$S_{\mathrm{MF}} := \left\{\left|\epsilon - \left(V^{*i}_{\boldsymbol{\pi}}(y) - \min_{a^i\in\mathbb{A}} W^{*i}_{\boldsymbol{\pi}}(y, a^i)\right)\right| : i \in \mathcal{n}, \boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}, y \in \mathbb{Y}\right\}.$$

That is, S_{MF} is the collection of (subjective) ϵ-optimality gaps of the various joint policies in $\widehat{\boldsymbol{\Pi}}$, and $\bar{d}_{\mathrm{MF}}$ is the minimum non-zero separation between ϵ and the (subjective) suboptimality gap of some player i's performance.

In Assumption 7, we assumed that each player's d^i parameter, which represents tolerance for suboptimality in their policy evaluation step and is included to account for noise in the learning iterates, is positive and small: $d^i \in (0, \bar{d}_{\mathrm{MF}})$ for each $i \in \mathcal{n}$.

We define $\Xi := \frac{1}{2}\min_{i\in\mathcal{n}}\{d^i, \bar{d}_{\mathrm{MF}} - d^i\}$. The quantity Ξ will serve as a desirable upper bound on learning error: if players jointly follow a joint policy $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$ and engage in Q-learning and value function estimation, then convergence to within Ξ of the limiting values ensures that each player correctly assesses whether it is $(\mathcal{V}^*, \mathcal{W}^*)$-subjectively ϵ-best-responding by using the comparison of Line 17 of Algorithm 3. We formalize this below.

For $k \geq 0$, we define Event(Ξ, k) as

$$\text{Event}(\Xi, k) := \left\{\max\left\{\|\widehat{J}^i_{t_{k+1}} - V^{*i}_{\boldsymbol{\pi}_k}\|_\infty, \|\widehat{Q}^i_{t_{k+1}} - W^{*i}_{\boldsymbol{\pi}_k}\|_\infty : i \in \mathcal{n}\right\} < \Xi\right\}.$$

Given Event(Ξ, k), we have that for any player $i \in \mathcal{n}$,

$$\pi^i_k \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}_k, \mathcal{V}^*, \mathcal{W}^*) \iff \widehat{J}^i_{t_{k+1}}(\varphi^i(\mathbf{x})) \leq \min_{a^i\in\mathbb{A}} \widehat{Q}^i_{t_{k+1}}(\varphi^i(\mathbf{x}), a^i) + \epsilon + d^i \quad \forall \mathbf{x} \in \mathbf{X}.$$

For convenience, we also define the following intersection of events. For any $k, \ell \in \mathbb{Z}_{\geq 0}$, let

$$E_{k:k+\ell} := \cap_{j=0}^{\ell} \text{Event}(\Xi, k + j).$$

That is, $E_{k:k+\ell}$ is the event where all agents obtain Ξ-accurate learning estimates in each of the exploration phases $k, k+1, \cdots, k+\ell$. For $k \in \mathbb{Z}_{\geq 0}$, we let G_k denote the event that the policy $\boldsymbol{\pi}_k$ is an ϵ-equilibrium: $G_k := \{\boldsymbol{\pi}_k \in \boldsymbol{\Pi} \cap \text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)\}$.

From the preceding discussion on $\bar{d}_{\mathrm{MF}}, \{d^i\}_{i\in\mathcal{n}}$ and the choice of Ξ, we have that for any $\ell \geq 0$,

$$\Pr\left(G_{k+\ell} | G_k \cap E_{k:k+\ell}\right) = 1. \tag{17}$$

Recall the quantity $L := \max\{L_{\boldsymbol{\pi}_0} : \boldsymbol{\pi}_0 \in \widehat{\boldsymbol{\Pi}}\}$, where for each $\boldsymbol{\pi}_0 \in \widehat{\boldsymbol{\Pi}}$, the number $L_{\boldsymbol{\pi}_0} < \infty$ is defined as the shortest $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satisficing path within $\widehat{\boldsymbol{\Pi}}$ starting at $\boldsymbol{\pi}_0$ and ending in $\widehat{\boldsymbol{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*)$. From our assumptions on $\widehat{\boldsymbol{\Pi}}$, such a path exists for every $\boldsymbol{\pi}_0 \in \widehat{\boldsymbol{\Pi}}$. If $L_{\boldsymbol{\pi}_0} < L$ for a particular initial policy $\boldsymbol{\pi}_0$, we may extend this path to have length L by repeating the final term. Thus, for every $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$, we obtain the inequality

$$\Pr\left(G_{k+L} | \{\boldsymbol{\pi}_k = \boldsymbol{\pi}\} \cap E_{k:k+L}\right) \geq p_{\min} > 0, \tag{18}$$

where $p_{\min} := \prod_{i \in \mathcal{N}} \left(\frac{e^i}{|\widehat{\Pi}^i|}\right)^L > 0$. The bound $p_{\min}$ is obtained through the following loose lower bounding argument: beginning at $\boldsymbol{\pi}_k = \boldsymbol{\pi}$, the joint policy process $\boldsymbol{\pi}_k, \ldots, \boldsymbol{\pi}_{k+L}$ follows the $(\mathcal{V}^*, \mathcal{W}^*)$-subjective ϵ-satificing path of length L described above with probability no less than the event that—at each step—the "correct" unsatisfied player updates to the "correct" policy uniformly at random, which occurs with probability no less than the probability given by the ratio in the product.

Fix $u^* \in (0, 1)$ such that

$$\frac{u^* p_{\min}}{1 - u^* + u^* p_{\min}} > 1 - \frac{\xi}{2}.$$

Combining Lemma 42 with a union bound, we have that there exists $\tilde{T} \in \mathbb{N}$ such that if $T_l > \tilde{T}$ for all $l \geq 0$, then $\Pr\left(E_{k:k+L} | \boldsymbol{\pi}_k = \boldsymbol{\pi}\right) \geq u^*$ for all $k \geq 0$ and any $\boldsymbol{\pi} \in \widehat{\boldsymbol{\Pi}}$. Thus, we have $\Pr\left(E_{k:k+L} | G_k\right) \geq u^*$ and $\Pr\left(E_{k:k+L} | G_k^c\right) \geq u^*$ for each $k \geq 0$.

We now lower bound $\Pr(G_{k+L})$ by conditioning on G_k and G_k^c as follows.

$$\Pr\left(G_{k+L}\right) = \Pr(G_{k+L} | G_k)\Pr(G_k) + \Pr(G_{k+L} | G_k^c)(1 - \Pr(G_k)).$$

We then lower bound each of the terms above by conditioning on $E_{k:k+L}$ and invoking inequalities (17) and (18):

$$\Pr(G_{k+L}) \geq 1 \cdot \Pr(E_{k:k+L} | G_k) \cdot \Pr(G_k) + p_{\min} \Pr(E_{k:k+L} | G_k^c)(1 - \Pr(G_k)).$$

Assume now that $T_l > \tilde{T}$ for all $l \geq 0$. We have

$$\Pr(G_{k+L}) \geq u^* \Pr(G_k) + u^* p_{\min}(1 - \Pr(G_k)), \quad \forall k \geq 0.$$

For each $k \in \{0, 1, \ldots, L-1\}$, define $y_0^{(k)} := \Pr(G_k)$, and for $m \geq 0$ define $y_{m+1}^{(k)} := u^* y_m^{(k)} + u^* p_{\min}\left(1 - y_m^{(k)}\right)$. By induction, one can show that

$$\Pr(G_{k+mL}) \geq y_m^{(k)}, \quad \forall m \geq 0. \tag{19}$$

Observe that $y_{m+1}^{(k)}$ can be written as

$$y_{m+1}^{(k)} = \left(u^* - u^* p_{\min}\right)^{m+1} y_0^{(k)} + u^* p_{\min} \sum_{j=0}^{m} \left(u^* - u^* p_{\min}\right)^j.$$

Since $0 < u^* - u^* p_{\min} < 1$, we have that $\lim_{m \to \infty} y_m^{(k)} = \frac{u^* p_{\min}}{1 - u^* + u^* p_{\min}} > 1 - \frac{\xi}{2}$. Then, by (19), we have that $\Pr(G_{k+mL}) \geq 1 - \xi/2$ holds for all sufficiently large m, which completes the proof.

Appendix H. An Approximation Result Relating to Mean-Field Equilibrium

In this section, we present results relating various solution concepts in mean-field games. The aim of this section is to suggest how one may use approximation results to establish the existence of subjective ϵ-equilibrium in partially observed n-player mean-field games with local observability. We begin by introducing the concept of *stationary mean-field equilibrium* (SMFE) and giving sufficient conditions for the existence of SMFE. We then argue that, under some additional conditions on the transition probabilities, a SMFE can be used to obtain an (objective) ϵ-equilibrium for a particular partially observed n-player mean-field game with sufficiently large n. We conclude this section with high-level discussion on how these results can be used to establish existence of *subjective* ϵ-equilibrium.

H.1 Mean-Field System

To facilitate passing from a model with finitely many players to a model with infinitely many players, we now introduce the terminology of *mean-field systems*, upon which the games will be defined or re-defined. The mean-field system underlying the mean-field game is a partial list of problem data, G, given by

$$G = (\mathbb{X}, \mathbb{A}, c, P_{\text{loc}}, \gamma).$$

The symbols in G retain their earlier meanings: $\mathbb{X}$ is a finite set of states, $\mathbb{A}$ is a finite set of actions, $c : \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A} \to \mathbb{R}$ is a cost function, $P_{\text{loc}} \in \mathscr{P}(\mathbb{X}|\mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A})$ is a local state transition kernel, and $\gamma \in (0, 1)$ is a discount factor.

H.2 Mean-Field MDPs

We now introduce a family of single-agent MDPs, denoted $\mathbf{M}$, whose elements are indexed by mean-field terms in $\Delta(\mathbb{X})$. For each $\nu \in \Delta(\mathbb{X})$, we let

$$\mathcal{M}_\nu = (\mathbb{X}, \mathbb{A}, c_\nu, P_\nu, \gamma)$$

be the MDP with state space $\mathbb{X}$, action set $\mathbb{A}$, discount factor γ, and where the stage cost function c_ν and transition kernel P_ν are given by

$$P_\nu(\cdot|x, a) := P_{\text{loc}}(\cdot|x, \nu, a), \quad c_\nu(x, a) := c(x, \nu, a), \quad \forall (x, a) \in \mathbb{X} \times \mathbb{A}.$$

To be clear, $\nu \in \Delta(\mathbb{X})$ need not be the initial distribution for the MDP $\mathcal{M}_\nu$: it is used merely to define the cost function and state transition probabilities by fixing the mean-field term at ν. In this section, any MDP of this form is called a *mean-field MDP*.

We let $\mathbf{M} := \{\mathcal{M}_\nu : \nu \in \Delta(\mathbb{X})\}$. For each MDP $\mathcal{M}_\nu \in \mathbf{M}$, the history sets are defined as $H_0 := \mathbb{X}$ and $H_{t+1} := H_t \times \mathbb{A} \times \mathbb{X}$ for each $t \geq 0$. An admissible policy is defined as a sequence of transition kernels $\pi = (\pi_t)_{t \geq 0}$ such that $\pi_t \in \mathscr{P}(\mathbb{A}|H_t)$ for each $t \geq 0$. The set of all policies for any MDP $\mathcal{M}_\nu \in \mathbf{M}$ is denoted by $\Pi_{\mathbf{M}}$, and we note that this set of admissible policies is common to every MDP in $\mathbf{M}$. As usual, $\Pi_{\mathbf{M},S}$ denotes the set of stationary policies. Here, we identify stationary policies with transition kernels in $\mathscr{P}(\mathbb{A}|\mathbb{X})$.

In accordance with the notation of Section A.1, the objective function for the MDP $\mathcal{M}_\nu$ controlled by policy $\pi \in \Pi_{\mathbf{M}}$ is denoted

$$\mathcal{G}_{\mathcal{M}_\nu}(\pi, \nu_0) = E^{\pi}_{\nu_0}\left[\sum_{t=0}^{\infty} \gamma^t c_\nu(x_t, a_t)\right],$$

where the expectation $E^{\pi}_{\nu_0}$ denotes that $x_0 \sim \nu_0$, and that for each $t \geq 0$ we have that $x_{t+1} \sim P_\nu(\cdot|x_t, a_t)$ and a_t is selected according to π. When $\nu_0 = \delta_s$ for some $s \in \mathbb{X}$, we simply write $\mathcal{G}_{\mathcal{M}_\nu}(\pi, s)$. Later, it will be necessary to add further indices to distinguish limiting objects from a sequence indexed by n. For this reason, we will also denote the preceding expectation $E^{\pi}_{\nu_0}$ as $E^{\pi,\infty}_{\nu_0}$.

We observe that the optimization problem faced by an agent controlling the single-agent MDP $\mathcal{M}_\nu$ is equivalent to the optimization problem faced by an agent in a partially observed n-player mean-field game wherein the mean-field sequence $\{\mu_t\}_{t\geq 0}$ is constant and given by $\mu_t = \nu$ for each $t \geq 0$.

Optimality Multi-Function: Since $\mathcal{M}_\nu$ is an MDP for each $\nu \in \Delta(\mathbb{X})$, the set of stationary optimal policies for $\mathcal{M}_\nu$ is non-empty. We define $\mathrm{opt} : \Delta(\mathbb{X}) \to 2^{\Pi_{\mathbf{M},S}}$ as

$$\mathrm{opt}(\nu) := \left\{\pi^* \in \Pi_{\mathbf{M},S} : \mathcal{G}_{\mathcal{M}_\nu}(\pi^*, x) = \min_{\pi \in \Pi_{\mathbf{M}}} \mathcal{G}_{\mathcal{M}_\nu}(\pi, x)\ \forall x \in \mathbb{X}\right\}, \quad \forall \nu \in \Delta(\mathbb{X}).$$

Note that the set $\mathrm{opt}(\nu)$ is compact and convex for each $\nu \in \Delta(\mathbb{X})$.

Population Update Function: Lastly, we define a function $\Phi : \Delta(\mathbb{X}) \times \Pi_{\mathbf{M},S} \to \Delta(\mathbb{X})$ as follows:

$$\Phi(\nu, \pi)(s') := \sum_{s \in \mathbb{X}} \sum_{a \in \mathbb{A}} P_{\mathrm{loc}}(s'|s, a)\pi(a|s)\nu(s), \quad \forall s' \in \mathbb{X}.$$

For each $s' \in \mathbb{X}$, $\Phi(\nu, \pi)(s')$ may be interpreted as the probability of $x_1 = s'$, where $x_0 \sim \nu$, $a_0 \sim \pi(\cdot|x_0)$, and $x_1 \sim P_{\mathrm{loc}}(\cdot|x_0, \nu, u_0)$.

H.3 Stationary Mean-Field Equilibrium: Definition and Existence

We now formally introduce the notion of stationary mean-field equilibrium (c.f. (Subramanian and Mahajan, 2019, Definition 2.1)) and give sufficient conditions for its existence.

Definition 43 *A pair $(\pi^*, \nu^*) \in \Pi_{\mathbf{M},S} \times \Delta(\mathbb{X})$ is called a* stationary mean-field equilibrium *(SMFE) for the mean-field system $G = (\mathbb{X}, \mathbb{A}, c, P_{\mathrm{loc}}, \gamma)$ if (i) $\Phi(\nu^*, \pi^*) = \nu^*$, and (ii) $\pi^* \in \mathrm{opt}(\nu^*)$.*

Assumption 9 *For each stationary policy $\pi \in \Pi_{\mathbf{M},S}$, there exists a unique invariant probability measure $\nu_\pi \in \Delta(\mathbb{X})$ satisfying $\Phi(\nu_\pi, \pi) = \nu_\pi$. Furthermore, the mapping $\pi \mapsto \nu_\pi$ is continuous on $\Pi_{\mathbf{M},S}$.*

Assumption 10 *For each $(s, a) \in \mathbb{X} \times \mathbb{A}$, the mapping $\nu \mapsto c(s, \nu, a)$ is continuous on $\Delta(\mathbb{X})$.*

Lemma 44 *For the mean-field system $G = (\mathbb{X}, \mathbb{A}, c, P_{\mathrm{loc}}, \gamma)$ defined above, if Assumptions 9 and 10 hold, then there exists a stationary mean-field equilibrium $(\pi^*, \nu_{\pi^*}) \in \Pi_{\mathbf{M},S} \times \Delta(\mathbb{X})$.*

Proof We define a point-to-set mapping $\mathcal{B} : \Pi_{\mathbf{M},S} \to 2^{\Pi_{\mathbf{M},S}}$ as follows: for each $\pi \in \Pi_{\mathbf{M},S}$, $\mathcal{B}(\pi) := \mathrm{opt}(\nu_\pi)$, where ν_π is the unique solution to $\Phi(\nu, \pi) = \nu$.

As previously noted, the correspondence $\nu \mapsto \mathrm{opt}(\nu)$ is convex-valued and compact-valued for each measure $\nu \in \Delta(\mathbb{X})$, and therefore so is $\mathrm{opt}(\nu_\pi)$ for each $\pi \in \Pi_{\mathbf{M},S}$. The continuity conditions imposed by Assumptions 9 and 10 ensure that $\mathcal{B}$ is upper hemicontinuous. We then invoke Kakutani's fixed point theorem and obtain our result. ■

For the remainder of this section, we suppose Assumptions 9 and 10 hold, and we fix $(\pi^*, \nu_{\pi^*}) \in \Pi_{\mathbf{M},S} \times \Delta(\mathbb{X})$ to be a stationary mean-field equilibrium.

H.4 A Family of Partially-Observed n-Player Mean-Field Games

To state the approximation result of this section, in which SMFE is used to obtain an objective ϵ-equilibrium in an n-player game when n is sufficiently large, we now introduce a family $\mathscr{G}$ of partially observed n-player mean-field games with local state observability. This family of games will be indexed both by the number of players, $n \in \mathbb{N}$, as well as mean-field terms relevant to the initial distribution of the local states, $\nu \in \Delta(\mathbb{X})$.

For the analysis, it will also be necessary to introduce a second family of partially observed n-player mean-field games with local state observability. This second family of n-player games is denoted by $\bar{\mathscr{G}}$.

For each $n \geq 1$ and $\nu \in \Delta(\mathbb{X})$, we define a partially observed n-player mean-field game

$$\mathcal{G}_n(\nu) = \left(\mathcal{n}_{(n)}, \mathbb{X}_{(n)}, \mathbb{Y}_{(n)}, \mathbb{A}_{(n)}, \{\varphi^i_{(n)}\}_{i \in \mathcal{n}_{(n)}}, c_{(n)}, P^{(n)}_{\mathrm{loc}}, \gamma_{(n)}, \boldsymbol{\nu}^{(n)}\right).$$

Here, $\mathcal{n}_{(n)} = \{1, \ldots, n\}$ is a set of n players, and we use (n) to note that the objects in the definition of $\mathcal{G}_n(\nu)$ depend on the number of players. We put $\mathbb{X}_{(n)} = \mathbb{X}$, $\mathbb{Y}_{(n)} = \mathbb{X}$, $\mathbb{A}_{(n)} = \mathbb{A}$, $c_{(n)} = c$, $P^{(n)}_{\mathrm{loc}} = P_{\mathrm{loc}}$, $\gamma_{(n)} = \gamma$, and note that $\mathbb{X}$, $\mathbb{A}$, c, P_{loc}, and γ retain their meanings from the mean-field system G. The set of global states is therefore $\mathbb{X}^n$, and the observation functions $\varphi^i_{(n)} : \mathbb{X}^n \to \mathbb{Y}_{(n)}$ are given by $\varphi^i_{(n)}(x^1, \ldots, x^n) = x^i$ for each $i \in \mathcal{n}_{(n)}$ and $(x^1, \ldots, x^n) \in \mathbb{X}^n$.

For the approximation results to follow, it will not be necessary to discuss n-player games with arbitrary initial distributions of global states. Our claims and analysis will be restricted to the special case where the local states of all agents are independently and identically distributed. We therefore let $\boldsymbol{\nu}^{(n)} \in \Delta(\mathbb{X}^n)$ denote the product distribution with marginals given by ν.

For any $n \in \mathbb{N}$ and player $i \in \mathcal{n}_n$, we let $\Pi^i_{(n)}$ denote player i's set of policies for the game $\mathcal{G}_n(\nu)$. We note that $\Pi^i_{(n)}$ is in natural bijection with $\Pi_{\mathbf{M}}$, the set of admissible policies for any mean-field MDPs in the family $\mathbf{M}$. The set of stationary policies for player $i \in \mathcal{n}_n$ is denoted by $\Pi^i_{(n),S}$ and is in bijection with the set $\Pi_{\mathbf{M},S}$ of stationary policies for the mean-field MDPs in $\mathbf{M}$. We then denote the joint policy set for $\mathcal{G}_n(\nu)$ by $\boldsymbol{\Pi}^{(n)}$, and we let $\boldsymbol{\Pi}^{(n)}_S$ denote the set of joint stationary policies for $\mathcal{G}_n(\nu)$.

The family of n-player mean-field games $\mathscr{G}$ is defined as $\mathscr{G} = \{\mathcal{G}_n(\nu) : n \in \mathbb{N}, \nu \in \Delta(\mathbb{X})\}$.

To distinguish state and action variables associated with different games in the family $\mathscr{G}$, we index local states, local actions, and mean-field states with (n): for instance, the local state of player $i \in \mathcal{N}_n$ at time t in the game $\mathcal{G}_n(\nu)$ is denoted $x_t^{i,(n)}$. Actions and mean-field states are similarly denoted $a_t^{i,(n)}$ and $\mu_t^{(n)}$. Dependence on ν will be reflected in the expectation operator and objective function, and is not included in the random variables.

The objective function for player $i \in \mathcal{N}_n$ in the game $\mathcal{G}_n(\nu)$ is denoted by

$$J_{(n)}^i\left(\pi_{(n)}^i, \boldsymbol{\pi}_{(n)}^{-i}, \boldsymbol{\nu}^{(n)}\right) = E_{\boldsymbol{\nu}^{(n)}}^{\boldsymbol{\pi}_{(n)}}\left[\sum_{t=0}^{\infty} \gamma^t c\left(x_t^{i,(n)}, a_t^{i,(n)}, \mu_t^{(n)}\right)\right],$$

for every $n \in \mathbb{N}$, $i \in \mathcal{N}_n$, $\boldsymbol{\pi}_{(n)} \in \boldsymbol{\Pi}^{(n)}$, and $\nu \in \Delta(\mathbb{X})$.

Finally, we introduce special notation to capture the performance of a policy $\pi \in \Pi_{\mathbf{M}}$ used in the game $\mathcal{G}_n(\nu)$ while all other agents in $\mathcal{N}_n$ follow a fixed stationary policy. Let $(\pi^*, \nu_{\pi^*}) \in \Pi_{\mathbf{M},S} \times \Delta(\mathbb{X})$ be the stationary mean-field equilibrium fixed after Lemma 44. For each $n \geq 1$, we fix player $1 \in \mathcal{N}_n$ and let $\boldsymbol{\pi}_{(n)}^{*-1}$ denote the $(n-1)$-player joint policy according to which all agents $j \neq 1$ use policy π^*. Then, we define

$$\mathscr{J}_{\pi}^{(n)}(\nu_{\pi^*}) := J_{(n)}^1\left(\pi, \boldsymbol{\pi}_{(n)}^{*-1}, \boldsymbol{\nu}_{\pi^*}^{(n)}\right), \quad \forall n \geq 1, \pi \in \Pi_{\mathbf{M}}.$$

H.5 A Second Family of Partially-Observed n-Player Mean-Field Games

For approximation purposes, we must introduce a second family of partially observed n-player mean-field games with local observability. The second family is denoted by $\bar{\mathscr{G}}$, and game in this family will serve to approximate games in the family $\mathscr{G}$.

First, recall that for each $\nu \in \Delta(\mathbb{X})$, we defined the transition kernel $P_\nu \in \mathscr{P}(\mathbb{X}|\mathbb{X}\times\mathbb{A})$ by $P_\nu(\cdot|s,a) = P_{\text{loc}}(\cdot|s,\nu,a)$ for every $(s,a) \in \mathbb{X}\times\mathbb{A}$. Now, we extend P_ν to obtain a transition kernel $\bar{P}_\nu \in \mathscr{P}(\mathbb{X}|\mathbb{X}\times\Delta(\mathbb{X})\times\mathbb{A})$: for each $\nu \in \Delta(\mathbb{X})$, define

$$\bar{P}_\nu(\cdot|s,\tilde{\nu},a) = P_\nu(\cdot|s,a), \quad \forall \tilde{\nu} \in \Delta(\mathbb{X}).$$

For each $n \geq 1$ and $\nu \in \Delta(\mathbb{X})$, we define a partially observed n-player mean-field game $\bar{\mathcal{G}}_n(\nu)$ with local observability:

$$\bar{\mathcal{G}}_n(\nu) := \left(\mathcal{N}_{(n)}, \mathbb{X}_{(n)}, \mathbb{Y}_{(n)}, \mathbb{A}_{(n)}, \{\varphi_{(n)}^i\}_{i\in\mathcal{N}_{(n)}}, c_{(n)}, \bar{P}_\nu, \gamma_{(n)}, \boldsymbol{\nu}^{(n)}\right).$$

In the list above, the objects $\mathcal{N}_{(n)}, \mathbb{X}_{(n)}, \mathbb{Y}_{(n)}, \mathbb{A}_{(n)}, \{\varphi_{(n)}^i\}_{i\in\mathcal{N}_{(n)}}, c_{(n)}, \gamma_{(n)}$, and $\boldsymbol{\nu}^{(n)}$ retain their meaning from the definition of $\mathcal{G}_n(\nu)$. The major difference between the definition of $\mathcal{G}_n(\nu)$ and $\bar{\mathcal{G}}_n(\nu)$ is the transition kernel: in $\bar{\mathcal{G}}_n(\nu)$, the transition kernel has ν fixed as its mean-field term, and local state transitions do not depend on the empirical distribution of the population. We return to this point below.

As in the case of $\mathcal{G}_n(\nu)$, the game $\bar{\mathcal{G}}_n(\nu)$ has local observability, and the global state space is $\mathbb{X}^n$. The joint action set is $\mathbb{A}^n$. As before, the set of policies for a given player

$i \in \mathcal{n}_n$ coincide with the set of policies $\Pi_{\mathbf{M}}$ for the mean-field MDPs in $\mathbf{M}$. Due to this natural bijection, we use the symbols $\Pi^i_{(n)}$, $\Pi^i_{(n),S}$, $\mathbf{\Pi}^{(n)}$, and $\mathbf{\Pi}^{(n)}_S$, retaining their meanings from before.

The family of games $\bar{\mathscr{G}}$ is given by $\bar{\mathscr{G}} := \left\{\bar{\mathcal{G}}_n(\nu) : n \in \mathbb{N}, \nu \in \Delta(\mathbb{X})\right\}$.

We now describe the construction of probability measures on sequences of global states and joint actions that describe probabilities for games in $\bar{\mathscr{G}} := \{\bar{\mathcal{G}}_n(\nu)\}_{n,\nu}$. To avoid mixing notation between analogous games $\mathcal{G}_n(\nu)$ and $\bar{\mathcal{G}}_n(\nu)$, we employ the symbol "$^-$" to identify that a quantity refers to $\bar{\mathcal{G}}_n(\nu)$. That is, a sequence in $\mathbb{X}^n \times \mathbb{A}^n$ of global states and joint actions obtained from a game in $\bar{\mathscr{G}}$ is expressed as

$$\left(\bar{\mathbf{x}}^{(n)}_t, \bar{\mathbf{a}}^{(n)}_t\right)_{t\geq 0}, \text{ where } \left(\bar{\mathbf{x}}^{(n)}_t, \bar{\mathbf{a}}^{(n)}_t\right) = \left(\bar{x}^{i,(n)}_t, \bar{a}^{i,(n)}_t\right)_{i\in\mathcal{n}_n}, \forall\, t \geq 0.$$

We define the empirical distribution of local states at time t in the n-player game $\bar{\mathcal{G}}_n(\nu)$ in the natural way and denote it $\{\bar{\mu}^{(n)}_t : t \geq 0\}$.

For $t \geq 0$, we use

$$\bar{\mathbf{h}}^{(n)}_t = \left(\bar{\mathbf{x}}^{(n}_0, \bar{\mathbf{a}}^{(n)}_0, \ldots, \bar{\mathbf{x}}^{(n)}_{t-1}, \bar{\mathbf{a}}^{(n)}_{t-1}, \bar{\mathbf{x}}^{(n)}_t\right)$$

to denote the system history at time t, and for each $i \in \mathcal{n}_n$ we let $\bar{h}^{i,(n)}_t$ denote player i's suitably defined locally observable history. As before, the initial distribution $\boldsymbol{\nu}^{(n)} \in \Delta(\mathbb{X}^n)$ is defined such that $\{\bar{x}^{i,(n)}_0\}_{i\in\mathcal{n}_n}$ is an i.i.d. family of random variables with common distribution $\nu \in \Delta(\mathbb{X})$.

For each $\boldsymbol{\pi} \in \mathbf{\Pi}^{(n)}$, $\overline{\Pr}^{\boldsymbol{\pi}_{(n)}}_{\nu}$ is the unique canonically defined probability measure on $(\mathbb{X}^n \times \mathbb{A}^n)^\infty$ such that:

- $\overline{\Pr}^{\boldsymbol{\pi}_{(n)}}_{\nu}\left(\bar{\mathbf{x}}^{(n)}_0 \in \cdot\right) = \boldsymbol{\nu}^{(n)}(\cdot)$;
- For every $i \in \mathcal{n}_n$ and $t \geq 0$, $\overline{\Pr}^{\boldsymbol{\pi}_{(n)}}_{\nu}\left(\bar{a}^{i,(n)}_t \in \cdot \middle| \bar{h}^{i,(n)}_t\right) = \pi^i_t(\cdot|\bar{h}^{i,(n)}_t)$, and the collection $\{\bar{a}^{j,(n)}_t\}_{j\in\mathcal{n}_n}$ is jointly independent given $\bar{\mathbf{h}}^{(n)}_t$;
- For every $i \in \mathcal{n}_n$ and $t \geq 0$, $\bar{x}^{i,(n)}_{t+1} \sim \bar{P}_\nu\left(\cdot \middle| \bar{x}^{i,(n)}_t, \bar{\mu}^{(n)}_t, \bar{a}^{i,(n)}_t\right)$, and the collection $\{\bar{x}^j_{t+1}\}_{j\in\mathcal{n}_n}$ is jointly independent given $(\bar{\mathbf{h}}^{(n)}_t, \bar{\mathbf{a}}^{(n)}_t)$.

For each $n \geq 1$, $\nu \in \Delta(\mathbb{X})$, player $i \in \mathcal{n}_n$, and joint policy $\boldsymbol{\pi} \in \mathbf{\Pi}^{(n)}$, we define player i's objective function in the game $\bar{\mathcal{G}}_n(\nu)$ under joint policy $\boldsymbol{\pi}$ as

$$\bar{J}^i_{(n)}\left(\boldsymbol{\pi}, \boldsymbol{\nu}^{(n)}\right) := \bar{E}^{\boldsymbol{\pi}_{(n)}}_{\nu}\left[\sum_{t\geq 0}\gamma^t c\left(\bar{x}^{i,(n)}_t, \bar{\mu}^{(n)}_t, \bar{a}^{i,(n)}_t\right)\right].$$

where $\bar{E}^{\boldsymbol{\pi}_{(n)}}_{\nu}$ is the expectation associated with $\overline{\Pr}^{\boldsymbol{\pi}_{(n)}}_{\nu}$.

Remark: Due to our definition of $\bar{P}_\nu$, we have that $\bar{x}_{t+1}^{i,(n)} \sim \bar{P}_\nu\left(\cdot\middle|\bar{x}_t^{i,(n)}, \bar{\mu}_t^{(n)}, \bar{a}_t^{i,(n)}\right)$ is equivalent to $\bar{x}_{t+1}^{i,(n)} \sim P_{\mathrm{loc}}\left(\cdot\middle|\bar{x}_t^{i,(n)}, \nu, \bar{a}_t^{i,(n)}\right)$. This distinction is crucial to the analysis below. In the game $\bar{\mathcal{G}}_n(\nu)$, the local state transitions for each player $i \in \mathcal{n}_n$ are totally uncoupled from the mean-field state. Although each player's local state evolution is uncoupled from the mean-field state in the game $\bar{\mathcal{G}}_n(\nu)$, player i's objective nevertheless depends on the mean-field sequence $\{\bar{\mu}_t^{(n)}\}_{t\geq 0}$, which appears in the stage costs.

Finally, for each $n \geq 1$ and policy $\pi \in \Pi_{\mathbf{M}}$, we fix agent $1 \in \mathcal{n}_n$, we define

$$\bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*}) := \bar{J}_{(n)}^1\left(\pi, \boldsymbol{\pi}_{(n)}^{*-1}, \boldsymbol{\nu}_{\pi^*}^{(n)}\right), \quad \forall n \geq 1, \pi \in \Pi_{\mathbf{M}},$$

where (π^*, ν_{π^*}) is the same SMFE as was used in the definition of $\mathscr{J}_\pi^{(n)}(\nu_{\pi^*})$, and $\boldsymbol{\pi}_{(n)}^{*-1}$ is the same joint policy as appeared in that definition.

H.6 From Stationary Mean-Field Equilibrium to Objective Equilibrium in an n-Player Game

We now proceed to the main result of this section, in which we show that for fixed $\epsilon > 0$, if n is sufficiently large, then the joint policy $\boldsymbol{\pi}_{(n)}^* := (\pi^*, \cdots, \pi^*) \in \boldsymbol{\Pi}^{(n)}$ is an ϵ-equilibrium with respect to $\boldsymbol{\nu}_{\pi^*}^{(n)}$ for the game $\mathcal{G}_n(\nu_{\pi^*})$. This result requires an additional assumption on the transition kernel P_{loc}, as well as two intermediate lemmas to facilitate the analysis.

In what follows, we suppose Assumptions 9 and 10 hold, and we let (π^*, ν_{π^*}) denote the stationary mean-field equilibrium used in the preceding definitions (e.g. the definition of $\mathscr{J}_\pi^{(n)}(\nu_{\pi^*})$).

Assumption 11 *There exists a function $F : \mathbb{X} \times \Delta(\mathbb{X}) \times \mathbb{A} \times [0,1] \to \mathbb{X}$ and a probability measure $\mathcal{L}$ on $[0,1]$ such that*

$$P_{\mathrm{loc}}(s'|s,\nu,a) = \int_{[0,1]} \mathbf{1}_{\{\xi\in[0,1]:F(s,\nu,a,\xi)=s'\}}(\tau)\, d\mathcal{L}(\tau),$$

for each $s' \in \mathbb{X}$ and every $(s,\nu,a) \in \mathbb{X}\times\Delta(\mathbb{X})\times\mathbb{A}$. Furthermore, there exists $\rho > 0$ such that if $\|\nu - \nu_{\pi^*}\|_\infty < \rho$, then $F(s,\nu,a,\xi) = F(s,\nu_{\pi^*},a,\xi)$ holds for any $(s,a,\xi) \in \mathbb{X}\times\mathbb{A}\times[0,1]$.

Lemma 45 *Let $\epsilon > 0$ and suppose Assumptions 9, 10, and 11 hold. There exists $N(\epsilon) \in \mathbb{N}$ such that if $n \geq N(\epsilon)$, then for any policy $\pi \in \Pi_{\mathbf{M}}$, we have*

$$|\mathscr{J}_\pi^{(n)}(\nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*})| < \epsilon/4.$$

Proof We use a coupling argument: for each $\pi \in \Pi_{\mathbf{M}}$ and $n \geq 1$, we construct a probability space $(\Omega_{n,\pi}, \mathcal{F}_{n,\pi}, \mathrm{Prob}^{n,\pi})$ on which both $\mathcal{G}_n(\nu_{\pi^*})$ and $\bar{\mathcal{G}}_n(\nu_{\pi^*})$ are defined, where agent $1 \in \mathcal{n}_n$ uses policy π and the remaining agents all use the stationary policy π^*. The component projections give rise to the appropriate marginal probability measures, i.e. $\mathrm{Pr}_{\boldsymbol{\nu}_{\pi^*}}^{\boldsymbol{\pi}_{(n)}}$ and $\overline{\mathrm{Pr}}_{\boldsymbol{\nu}_{\pi^*}}^{\boldsymbol{\pi}_{(n)}}$ with suitably defined policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{(n)}$, but are otherwise closely coupled.

Let $\Omega_{n,\pi} := (\mathbb{X}^n \times \mathbb{X}^n \times [0,1]^n \times \mathbb{A}^n \times \mathbb{A}^n \times [0,1]^n)^\infty$, and let $\mathcal{F}_{n,\pi}$ be the σ-algebra generated by finite dimensional cylinders in $\Omega_{n,\pi}$. Components of $\Omega_{n,\pi}$ will be denoted by sequences $(\bar{\mathbf{x}}_t^{(n)}, \mathbf{x}_t^{(n)}, \boldsymbol{\eta}_t^{(n)}, \bar{\mathbf{a}}_t^{(n)}, \mathbf{a}_t^{(n)}, \boldsymbol{\xi}_t^{(n)})_{t\geq 0}$. The probability measure $\text{Prob}^{n,\pi}$ is constructed as follows:

- For each $j \in \mathcal{n}_n$, $\text{Prob}^{n,\pi}\left(\bar{x}_0^{j,(n)} \in \cdot\right) = \nu_{\pi^*}(\cdot)$, and the collection $\left\{\bar{x}_0^{j,(n)}\right\}_{j\in\mathcal{n}_n}$ is jointly independent.
- $\text{Prob}^{n,\pi}\left(\mathbf{x}_0^{(n)} = \bar{\mathbf{x}}_0^{(n)}\right) = 1$.
- Let $\mathcal{K}$ be a probability measure on $[0,1]$, let $g : (\cup_{t\geq 0} H_t) \times [0,1] \to \mathbb{A}$, and let $g^* : \mathbb{X} \times [0,1] \to \mathbb{A}$ satisfy

  $$\pi_t(a|h_t) = \int_{[0,1]} \mathbf{1}_{\{\xi\in[0,1]: g(h_t,\xi)=a\}}(\tau)\, d\mathcal{K}(\tau), \quad \forall t \geq 0, h_t \in H_t, a \in \mathbb{A},$$

  where $\pi = (\pi_t)_{t\geq 0}$, and

  $$\pi^*(a|s) = \int_{[0,1]} \mathbf{1}_{\{\xi\in[0,1]: g^*(s,\xi)=a\}}(\tau)\, d\mathcal{K}(\tau), \quad \forall s \in \mathbb{X}, a \in \mathbb{A}.$$

- For any $t \geq 0$, we have $\text{Prob}^{n,\pi}\left(\eta_t^{j,(n)} \in \cdot\right) = \mathcal{K}(\cdot)$ for all $j \in \mathcal{n}_n$, the collection $\{\eta_t^{j,(n)}\}_{j\in\mathcal{n}_n}$ is jointly independent, and $\boldsymbol{\eta}_t^{(n)}$ is independent of the history up to $\mathbf{x}_t^{(n)}$.
- For $j \neq 1$ and $t \geq 0$, we have $\bar{a}_t^j = g^*\left(\bar{x}_t^j, \eta_t^{j,(n)}\right)$ and $a_t^j = g^*\left(x_t^j, \eta_t^{j,(n)}\right)$.
- For agent $1 \in \mathcal{n}_n$ and $t \geq 0$, we have $\bar{a}_t^{1,(n)} = g\left(\bar{h}_t^{1,(n)}, \eta_t^{1,(n)}\right)$ and $a_t^{1,(n)} = g\left(h_t^{1,(n)}, \eta_t^{1,(n)}\right)$.
- For any $t \geq 0$ and agent $j \in \mathcal{n}_n$, we have $\text{Prob}^{n,\pi}\left(\xi_t^{j,(n)} \in \cdot\right) = \mathcal{L}(\cdot)$, where $\mathcal{L}$ is the probability measure introduced in Assumption 11. Furthermore, the collection $\{\xi_t^{j,(n)}\}_{j\in\mathcal{n}_n}$ is jointly independent, and $\boldsymbol{\xi}_t^{(n)}$ is independent of the history up to $\mathbf{a}_t^{(n)}$;
- For every $t \geq 0$ and agent $j \in \mathcal{n}_n$, we have

  $$\bar{x}_{t+1}^{j,(n)} = F(\bar{x}_t^j, \nu_{\pi^*}, \bar{a}_t^{j,(n)}, \xi_t^{j,(n)}) \text{ and } x_{t+1}^{j,(n)} = F(x_t^{j,(n)}, \mu_t^{(n)}, a_t^{j,(n)}, \xi_t^{j,(n)}),$$

  where $\mu_t^{(n)}$ is defined as the empirical measure of $\mathbf{x}_t^{(n)}$.

The measure $\text{Prob}^{n,\pi}$ is then the unique extension of the conditional probabilities above to $(\Omega_{n,\pi}, \mathcal{F}_{n,\pi})$. We now observe some consequences of this construction.

First, observe that the state transitions resulting in $\{\bar{x}_{t+1}^{j,(n)}\}_{j\in\mathcal{n}_n}$ are uncoupled. For any $\pi \in \Pi_{\mathbf{M}}$, $n \geq 1$, and $j \neq 1 \in \mathcal{n}_n$, since player j uses policy π^*, since $\bar{x}_0^{j,(n)} \sim \nu_{\pi^*}$, and

since $\Phi(\nu_{\pi^*}, \pi^*) = \nu_{\pi^*}$, we have that (1) $\bar{x}_t^{j,(n)} \sim \nu_{\pi^*}$ for every $t \geq 0$; (2) the collection $\{\bar{x}_t^{j,(n)}\}_{j \in \mathcal{n}_n}$ is independent; (3) the distribution of $\{\bar{x}_t^{j,(n)}\}_{j \in \mathcal{n}_n}$ does not depend on the policy π for player $1 \in \mathcal{n}_n$. Together, this implies that for any $t \geq 0$ and $\epsilon' > 0$, we have

$$\lim_{n \to \infty} \text{Prob}^{n,\pi} \left(\|\bar{\mu}_t^{(n)} - \nu_{\pi^*}\|_\infty < \epsilon' \right) = 1, \quad \forall \pi \in \Pi_{\mathbf{M}}.$$

In particular, this holds for $\epsilon' = \rho$, where $\rho > 0$ is the radius introduced in Assumption 11.

We fix $T \in \mathbb{N}$ such that $\frac{\gamma^T \|c\|_\infty}{1-\gamma} < \frac{\epsilon}{32}$. By the preceding remarks, we have that for any $\epsilon_1 > 0$, there exists $N_1(\epsilon_1) \in \mathbb{N}$ such that if $n \geq N_1(\epsilon_1)$, we have

$$\text{Prob}^{n,\pi} \left(\bigcap_{t=0}^{T-1} \{ \|\bar{\mu}_t^{(n)} - \nu_{\pi^*}\|_\infty < \rho \} \right) \geq 1 - \epsilon_1, \quad \forall \pi \in \Pi_{\mathbf{M}}. \tag{20}$$

We now argue that the preceding inequalities enable the following claim: for any $\epsilon_2 > 0$, there exists $N_2(\epsilon_2) \in \mathbb{N}$ such that if $n \geq N_2(\epsilon_2)$, then

$$\text{Prob}^{n,\pi} \left(\bigcap_{t=0}^{T-1} \left\{ \bar{\mathbf{x}}_t^{(n)} = \mathbf{x}_t^{(n)}, \bar{\mathbf{a}}_t^{(n)} = \mathbf{a}_t^{(n)} \right\} \right) \geq 1 - \epsilon_2, \quad \forall \pi \in \Pi_{\mathbf{M}}. \tag{21}$$

To see this, we take $\tau > 0$ such that $1 - \tau > \left(\max\{\frac{1}{2}, 1 - \epsilon_2\} \right)^{1/T}$. From the discussion above, there exists $N_1(\tau/2) \in \mathbb{N}$ such that if $n \geq N_1(\tau/2)$, then (20) holds with $\epsilon_1 = \tau/2$. For each $k \in \{0, 1, \ldots, T-1\}$, let $\text{Event}(k, n) := \{\bar{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)}, \bar{\mathbf{a}}_k^{(n)} = \mathbf{a}_k^{(n)}\}$. Observe the following three facts, which hold for any $t \in \{0, 1, \ldots, T-1\}$:

1. By construction,

$$\text{Prob}^{n,\pi} \left(\text{Event}(t+1, n) \middle| \cap_{k=0}^t \text{Event}(k, n), \left\| \bar{\mu}_t^{(n)} - \nu_{\pi^*} \right\|_\infty < \rho \right) = 1.$$

2. If $\text{Prob}^{n,\pi} \left(\cap_{k=0}^t \text{Event}(k, n) \right) \geq (1-\tau)^t$, then by our choice of τ and $t \leq T$, this implies $\text{Prob}^{n,\pi} \left(\cap_{k=0}^t \text{Event}(k, n) \right) \geq \frac{1}{2}$. Then, by our choice of $n \geq N_1(\tau/2)$, we have $\text{Prob}^{n,\pi} \left(\left\| \bar{\mu}_t^{(n)} - \nu_{\pi^*} \right\|_\infty < \rho \right) \geq 1 - \tau/2$. Putting the two together,[6], we obtain

$$\text{Prob}^{n,\pi} \left(\left\| \bar{\mu}_t^{(n)} - \nu_{\pi^*} \right\|_\infty < \rho \middle| \bigcap_{k=0}^t \text{Event}(k, n) \right) \geq 1 - \tau.$$

3. If $\text{Prob}^{n,\pi} \left(\cap_{k=0}^t \text{Event}(k, n) \right) \geq (1-\tau)^t$, then by the previous two items, we have

$$\text{Prob}^{n,\pi} \left(\text{Event}(t+1, n) \middle| \bigcap_{k=0}^t \text{Event}(k, n) \right) \geq (1-\tau)^{t+1}.$$

6. Here, we have used the fact that if events A and B satisfy $\Pr(B) \geq b > 0$, $\Pr(A) \geq 1 - ab$ with $a > 0$, then $\Pr(A|B) \geq 1 - a$.

4. By construction, we have that the base case for item 2 holds with $t = 0$. Thus, we repeatedly apply item 3 to see that, indeed, the inequality of (21) holds for $n \geq N_2(\epsilon_2) := N_1(\tau/2)$.

Now, we argue that the existence result summarized in (21) implies the result. For each $n \geq 1$ and $\pi \in \Pi_{\mathbf{M}}$, let $\mathbb{E}^{n,\pi}$ be the expectation associated with $\text{Prob}^{n,\pi}$. Observe that, for any $n \geq 1$ and $\pi \in \Pi_{\mathbf{M}}$, we have

$$\mathscr{J}_\pi^{(n)}(\nu_{\pi^*}) = \mathbb{E}^{n,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c\left(x_t^{1,(n)}, \mu_t^{(n)}, a_t^{1,(n)}\right)\right],$$

and

$$\bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*}) = \mathbb{E}^{n,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c\left(\bar{x}_t^{1,(n)}, \bar{\mu}_t^{(n)}, \bar{a}_t^{1,(n)}\right)\right].$$

Then, for each $n \geq 1, \pi \in \Pi_{\mathbf{M}}$, we let $\mathscr{J}_\pi^{(n),T}(\nu_{\pi^*})$ and $\bar{\mathscr{J}}_\pi^{(n),T}(\nu_{\pi^*})$ denote the associated series truncated after the summand with index $T-1$. That is,

$$\mathscr{J}_\pi^{(n),T}(\nu_{\pi^*}) := \mathbb{E}^{n,\pi}\left[\sum_{t=0}^{T-1} \gamma^t c(x_t^{1,(n)}, \mu_t^{(n)}, a_t^{1,(n)})\right],$$

and $\bar{\mathscr{J}}_\pi^{(n),T}(\nu_{\pi^*})$ is defined analogously.

By our choice of T to satisfy $\frac{\gamma^T \|c\|_\infty}{1-\gamma} < \frac{\epsilon}{32}$, we have the following for any $n \geq 1, \pi \in \Pi_{\mathbf{M}}$:

$$\left|\mathscr{J}_\pi^{(n)}(\nu_{\pi^*}) - \mathscr{J}_\pi^{(n),T}(\nu_{\pi^*})\right| < \frac{\epsilon}{16} \text{ and } \left|\bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n),T}(\nu_{\pi^*})\right| < \frac{\epsilon}{16}.$$

Applying the triangle inequality twice, we then obtain

$$\left|\mathscr{J}_\pi^{(n)}(\nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*})\right| \leq \frac{\epsilon}{8} + \left|\mathscr{J}_\pi^{(n),T}(\nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n),T}(\nu_{\pi^*})\right|.$$

By the deductions above culminating in (21), we have that for any $\epsilon_3 > 0$, there exists $N_3(\epsilon_3)$ such that if $n \geq N_3(\epsilon_3)$, we have

$$\left|\mathscr{J}_\pi^{(n),T}(\nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n),T}(\nu_{\pi^*})\right| < \epsilon_3.$$

In particular, this is true for $\epsilon_3 = \frac{\epsilon}{8}$. This proves the result, with $N(\epsilon) = N_3(\epsilon/8)$. ■

For the next approximation result, recall that for $\nu \in \Delta(\mathbb{X})$, $\mathcal{J}_{\mathcal{M}_\nu}(\pi, \nu')$ is the objective performance of policy $\pi \in \Pi_{\mathbf{M}}$ when controlling the MDP $\mathcal{M}_\nu$ with initial state distributed according to ν'. In the lemma below, we fix both $\nu = \nu' = \nu_{\pi^*}$ to be the probability distribution appearing in the stationary mean-field equilibrium (π^*, ν_{π^*}).

Lemma 46 *Let $\epsilon > 0$ and suppose Assumptions 9, 10, and 11 hold. There exists $N(\epsilon) \in \mathbb{N}$ such that if $n \geq N(\epsilon)$, then for any policy $\pi \in \Pi_{\mathbf{M}}$, we have*

$$|\mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*}) - \bar{\mathscr{J}}_\pi^{(n)}(\nu_{\pi^*})| < \epsilon/4.$$

Proof For each $n \geq 1$ and $\pi \in \Pi_{\mathbf{M}}$, recall the probability spaces $(\Omega_{n,\pi}, \mathscr{F}_{n,\pi}, \mathrm{Prob}^{n,\pi})$ used in the coupling argument of the proof of Lemma 45 and recall the definition of $T \in \mathbb{N}$. For the MDP $\mathcal{M}_{\nu_{\pi^*}}$, we let $\Pr^{\pi,\infty}_{\nu_{\pi^*}}$ denote the probability measure on sequences of local states and actions when the MDP is controlled by π.

We note that, by construction of $(\Omega_{n,\pi}, \mathscr{F}_{n,\pi}, \mathrm{Prob}^{n,\pi})$, the marginal distributions on (local) state-action trajectories for player $1 \in \mathcal{N}_n$ are equal to those in the MDP $\mathcal{M}_{\nu_{\pi^*}}$ controlled by policy π. That is,

$$\mathrm{Prob}^{n,\pi}\left(\{\bar{x}_t^{1,(n)}, \bar{a}_t^{1,(n)}\}_{t=0}^{T-1} \in \cdot\right) = \Pr^{\pi,\infty}_{\nu_{\pi^*}}\left(\{x_t, a_t\}_{t=0}^{T-1} \in \cdot\right).$$

Thus, the expected costs $\mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})$ and $\bar{\mathscr{J}}_{\pi}^{(n)}(\nu_{\pi^*})$ differ in the mean-field term, which is posited to be constant at ν_{π^*} in the former, while in the latter the mean-field sequence $\{\bar{\mu}_t^{(n)}\}_{t\geq 0}$ is random.

For each $n \geq 1$ and $\pi \in \Pi_{\mathbf{M}}$, we recall the object $\bar{\mathscr{J}}_{\pi}^{(n),T}(\nu_{\pi^*})$ from the proof of Lemma 45. We then define an analogous quantity involving the MDP $\mathcal{M}_{\nu_{\pi^*}}$ controlled by policy π:

$$\mathcal{J}^T_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*}) := E^{\pi,\infty}_{\nu_{\pi^*}}\left[\sum_{t=0}^{T-1} \gamma^t c(x_t, \nu_{\pi^*}, a_t)\right].$$

By our choice of T, $\left|\bar{\mathscr{J}}_{\pi}^{(n)}(\nu_{\pi^*}) - \bar{\mathscr{J}}_{\pi}^{(n),T}(\nu_{\pi^*})\right| < \frac{\epsilon}{16}$ and $\left|\mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*}) - \mathcal{J}^T_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})\right| < \frac{\epsilon}{16}$, thus

$$\left|\bar{\mathscr{J}}_{\pi}^{(n)}(\nu_{\pi^*}) - \mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})\right| \leq \frac{\epsilon}{8} + \left|\bar{\mathscr{J}}_{\pi}^{(n),T}(\nu_{\pi^*}) - \mathcal{J}^T_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})\right|.$$

As we argued in the proof of Lemma 45, for any $\xi_1, \xi_2 > 0$, there exists $N = N(\xi_1, \xi_2) \in \mathbb{N}$ such that if $n \geq N$, we have

$$\mathrm{Prob}^{n,\pi}\left(\bigcap_{t=0}^{T-1}\left\{\left\|\bar{\mu}_t^{(n)} - \nu_{\pi^*}\right\|_{\infty} < \xi_1\right\}\right) \geq 1 - \xi_2, \quad \forall \pi \in \Pi_{\mathbf{M}}.$$

By Assumption 10, the mapping $\nu \mapsto c(s, \nu, a)$ is continuous on $\Delta(\mathbb{X})$ for all $(s, a) \in \mathbb{X}\times\mathbb{A}$. Thus, there exists $\xi > 0$ such that $\|\nu - \nu_{\pi^*}\|_{\infty} < \xi$ implies $|c(s, \nu, a) - c(s, \nu_{\pi^*}, a)| \leq \frac{\epsilon}{8T}$ for any $(s, a) \in \mathbb{X} \times \mathbb{A}$. The result then follows another mechanical computation. ∎

Theorem 47 *Let (π^*, ν_{π^*}) be a stationary mean-field equilibrium for $(\mathbb{X}, \mathbb{A}, c, P_{\mathrm{loc}}, \gamma)$, and suppose Assumptions 9, 10, and 11 hold. Let $\epsilon > 0$ be given. There exists $N(\epsilon) \in \mathbb{N}$ such that if $n \geq N(\epsilon)$, then, in the game $\mathcal{G}_n(\nu_{\pi^*})$, the policy $\boldsymbol{\pi}^*_{(n)} \in \boldsymbol{\Pi}_S^{(n)}$, in which every agent uses π^*, is an ϵ-equilibrium with respect to $\boldsymbol{\nu}_{\pi^*}$.*

Proof By Lemmas 45 and 46, there exists $\bar{N}(\epsilon) \in \mathbb{N}$ such that if $n \geq \bar{N}(\epsilon)$, then both

$$\left|\mathscr{J}_{\pi}^{(n)}(\nu_{\pi^*}) - \bar{\mathscr{J}}_{\pi}^{(n)}(\nu_{\pi^*})\right| < \frac{\epsilon}{4} \text{ and } \left|\bar{\mathscr{J}}_{\pi}^{(n)}(\nu_{\pi^*}) - \mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})\right| < \frac{\epsilon}{4} \quad \forall \pi \in \Pi_{\mathbf{M}}. \quad (22)$$

Since (π^*, ν_{π^*}) is a SMFE for the system, we know $\mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi^*, \nu_{\pi^*}) \leq \mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*})$ for any $\pi \in \Pi_{\mathbf{M}}$. Then, by (22), we have

$$\mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi, \nu_{\pi^*}) \leq \mathscr{J}_{\pi}^{(n)}(\nu_{\pi^*}) + \frac{\epsilon}{2}, \quad \forall \pi \in \Pi_{\mathbf{M}}.$$

Also by (22), we have $\mathscr{J}_{\pi^*}^{(n)}(\nu_{\pi^*}) \leq \mathcal{J}_{\mathcal{M}_{\nu_{\pi^*}}}(\pi^*, \nu_{\pi^*}) + \epsilon/2$. Combining the previous observations, we have that

$$\mathscr{J}_{\pi^*}^{(n)}(\nu_{\pi^*}) \leq \mathscr{J}_{\pi}^{(n)}(\nu_{\pi^*}) + \epsilon, \quad \forall \pi \in \Pi_{\mathbf{M}}.$$

This shows that π^* is an ϵ-best-response to $\boldsymbol{\pi}_{(n)}^{*-1}$ with respect to the initial measure $\boldsymbol{\nu}_{\pi^*}^{(n)}$ in the game $\mathcal{G}_n(\nu_{\pi^*})$. ■

Remarks

As we have observed in the preceding analysis, under some regularity conditions on the partially observed n-player mean-field game with local state observability, if the number of agents is sufficiently large and the agents happen to be using the policy π^* given by a stationary mean-field equilibrium, *and* if the initial global state is distributed according to the associated invariant measure ν_{π^*}, then the environment faced by a learning agent is approximated by the MDP $\mathcal{M}_{\nu_{\pi^*}}$.

Although the summary above is highly qualified, we observe that if every agent uses policy π^* during a given exploration phase during the running of Algorithm 3 and (for one reason or another) does not switch policies at the end of that exploration phase, then the initial state of the subsequent exploration phase will be distributed according to a measure that is close to ν_{π^*}, due to the assumed ergodicity of the system. Thus, it is conceivable that the naive learning iterates of an agent in the n-player game will be approximated by the Q-function and state value function for the MDP $\mathcal{M}_{\nu_{\pi^*}}$, leading to a possible avenue for establishing subjective ϵ-equilibrium in the truly decentralized problem.

Appendix I. Satisficing Under Other Observation Channels

Subjective Satisficing Paths under Global Observability

Definition 48 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumption 1 is satisfied, and let* $i \in \mathcal{N}$*. A stationary policy* $\pi^i \in \Pi_S^i$ *is said to be* of the mean-field type *if there exists* $f^i \in \mathscr{P}(\mathbb{A}|\mathbb{X} \times \mathrm{Emp}_N)$ *such that* $\pi^i(\cdot|\mathbf{s}) = f^i\left(\cdot\middle|s^i, \mu(\mathbf{s})\right)$ *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 $\mathscr{P}(\mathbb{A}|\mathbb{X} \times \mathrm{Emp}_N)$.

Definition 49 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumption 1 is satisfied. For* $i, j \in \mathcal{N}$*, let* $\pi^i \in \Pi_S^i$ *and* $\pi^j \in \Pi_S^j$ *both be of the mean-field type. We say that* π^i *and* π^j *are* mean-field symmetric *if they are identified with the same transition kernel in* $\mathscr{P}(\mathbb{A}|\mathbb{X} \times \mathrm{Emp}_N)$*.*

Lemma 50 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumptions 1 and 5 are satisfied, and let* $\epsilon \geq 0$*. Let* $\boldsymbol{\pi} \in \mathbf{\Pi}_S$ *be a soft stationary joint policy for which* π^p *is of the mean-field type for each player* $p \in \mathcal{n}$*. Suppose that, for some* $i, j \in \mathcal{n}$*, we have that* π^i *and* π^j *are mean-field symmetric, then we have*

$$\pi^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*) \iff \pi^j \in \text{Subj-BR}^j_\epsilon(\boldsymbol{\pi}^{-j}, \mathcal{V}^*, \mathcal{W}^*).$$

Proof Let $i, j \in \mathcal{n}$. For each $\mathbf{s} \in \mathbf{X}$, we define $\text{swap}_{ij}(\mathbf{s}) \in \mathbf{X}$ to be the global state such that (1) $\text{swap}_{ij}(\mathbf{s})^p = s^p$ for each $p \in \mathcal{n} \setminus \{i, j\}$, and (2) we have $\text{swap}_{ij}(\mathbf{s})^j = s^i$ and $\text{swap}_{ij}(\mathbf{s})^i = s^j$.

By Theorem 13, we have that $V^{*i}_{\boldsymbol{\pi}}(\mathbf{s}) = J^i(\boldsymbol{\pi}, \mathbf{s})$ for each $\mathbf{s} \in \mathbf{X}$, $W^{*i}_{\boldsymbol{\pi}} = Q^{*i}_{\boldsymbol{\pi}^{-i}}$, $V^{*j}_{\boldsymbol{\pi}}(\mathbf{s}) = J^j(\boldsymbol{\pi}, \mathbf{s})$ for each $\mathbf{s} \in \mathbf{X}$, and $W^{*j}_{\boldsymbol{\pi}} = Q^{*j}_{\boldsymbol{\pi}^{-j}}$. Thus, we have that $\pi^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*)$ if and only if $\pi^i \in \text{BR}^i_\epsilon(\boldsymbol{\pi}^{-i})$, and analogously for j.

Toward obtaining a contradiction, assume that $\pi^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*)$ while $\pi^j \notin \text{Subj-BR}^j_\epsilon(\boldsymbol{\pi}^{-j}, \mathcal{V}^*, \mathcal{W}^*)$. Equivalently, assume that $\pi^i \in \text{BR}^i_\epsilon(\pi^{-i})$ while $\pi^j \notin \text{BR}^i_\epsilon(\pi^{-j})$. That is, we have that

$$J^i(\boldsymbol{\pi}, \mathbf{s}) \leq \inf_{\tilde{\pi}^i \in \Pi^i_S} J^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}, \mathbf{s}) + \epsilon, \quad \forall \mathbf{s} \in \mathbf{X},$$

while there exists $\mathbf{s}^* \in \mathbf{X}$ such that $J^j(\boldsymbol{\pi}, \mathbf{s}^*) > \inf_{\tilde{\pi}^j \in \Pi^j_S} J^j(\tilde{\pi}^j, \boldsymbol{\pi}^{-j}, \mathbf{s}^*) + \epsilon$.

Observe that if $\boldsymbol{\pi} \in \mathbf{\Pi}_S$ is of the mean-field type and $\tilde{\pi}^i \in \Pi^i_S$ is any stationary policy for player i (not necessarily of the mean-field type), then defining a policy $\tilde{\pi}^j \in \Pi^j_S$ state-wise by $\tilde{\pi}^j(\cdot|\mathbf{s}) = \tilde{\pi}^i(\cdot|\text{swap}_{ij}(\mathbf{s}))$ for all $\mathbf{s}$, we have that

$$J^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}, \mathbf{s}) = J^j(\tilde{\pi}^j, \boldsymbol{\pi}^{-j}, \text{swap}_{ij}(\mathbf{s})), \quad \forall \mathbf{s} \in \mathbf{X}.$$

It follows that, first, we have $J^i(\boldsymbol{\pi}, \mathbf{s}) = J^j(\boldsymbol{\pi}, \text{swap}_{ij}(\mathbf{s}))$ for each $\mathbf{s} \in \mathbf{X}$ and furthermore

$$\inf_{\tilde{\pi}^i \in \Pi^i_S} J^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}, \mathbf{s}) = \inf_{\tilde{\pi}^j \in \Pi^j_S} J^j(\tilde{\pi}^j, \boldsymbol{\pi}^{-j}, \text{swap}_{ij}(\mathbf{s})), \quad \forall \mathbf{s} \in \mathbf{X}.$$

In particular, this holds for $\text{swap}_{ij}(\mathbf{s}^*)$, which yields

$$\begin{aligned} J^j(\boldsymbol{\pi}, \mathbf{s}^*) > \inf_{\tilde{\pi}^j \in \Pi^j_S} J^j(\tilde{\pi}^j, \boldsymbol{\pi}^{-j}, \mathbf{s}^*) + \epsilon &= \inf_{\tilde{\pi}^i \in \Pi^i_S} J^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}, \text{swap}_{ij}(\mathbf{s}^*)) + \epsilon \\ &\geq J^i(\boldsymbol{\pi}, \text{swap}_{ij}(\mathbf{s}^*)) = J^j(\boldsymbol{\pi}, \mathbf{s}^*), \end{aligned}$$

a contradiction, which completes the proof. ■

Lemma 51 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumptions 1 and 5. Let* $\epsilon > 0$*. There exists* $\xi = \xi(\epsilon) > 0$ *such that if* $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ *is any soft* ξ*-quantization of* $\mathbf{\Pi}_S$*, then we have*

1) $\mathbf{\Pi}^{\epsilon\text{-eq}} \cap \widehat{\mathbf{\Pi}} \neq \varnothing$ *and* $\text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \widehat{\mathbf{\Pi}} \neq \varnothing$*.*

Moreover, if $\tilde{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ 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 $\tilde{\mathbf{\Pi}}$, then we have

2) $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \cap \mathbf{\Pi}' \neq \varnothing$*;*

3) $\mathbf{G}$ *has the* $(\mathcal{V}^*, \mathcal{W}^*)$*-subjective* ϵ*-satisficing paths property within* $\mathbf{\Pi}'$*.*

Proposition 52 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumptions 1 and 5 hold. Let* $\epsilon > 0$*, and let* $(\mathcal{V}^*, \mathcal{W}^*)$ *be the subjective function family for* $\mathbf{G}$*. Suppose* $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ *is a subset of stationary joint policies satisfying the following properties: (i) Every* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$ *is of the mean-field type; (ii)* $\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*; (iii) the set* $\widehat{\mathbf{\Pi}}$ *is symmetric, i.e.* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for each* $i, j \in \mathcal{N}$*; (iv) every* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$ *is soft.*

Then, $\mathbf{G}$ *has the* $(\mathcal{V}^*, \mathcal{W}^*)$*-subjective* ϵ*-satisficing paths property within* $\widehat{\mathbf{\Pi}}$*.*

The proof of Proposition 52 is omitted, as it parallels that of (Yongacoglu et al., 2023, Theorem 3.6), which was in the context of symmetric, n-player stochastic games with global observability. One important difference, which explains the need for Lemma 50 and restriction to policies of the mean-field type, is that here the global state dynamics do not satisfy the conditions outlined in Yongacoglu et al. (2023). This occurs because of the added structure on the global state space present in mean-field games.

Subjective Satisficing Paths under Compressed Observability

In light of our previous discussion on the possible non-existence of subjective ϵ-equilibrium in games with compressed observability, we now give a qualified result analogous to Proposition 52 for the case with compressed observability. In contrast to the earlier results, here we must assume the existence of subjective equilibrium.

Definition 53 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game and let* $i \in \mathcal{N}$*. A policy* $\pi^i \in \Pi^i_S$ *is said to be* of the local type *if there exists a transition kernel* $g^i \in \mathcal{P}(\mathbb{A}|\mathbb{X})$ *such that*

$$\pi^i(\cdot|\varphi^i(\mathbf{s})) = g^i(\cdot|s^i), \quad \forall \mathbf{s} \in \mathbf{X}.$$

Each policy of the local type is identified with the corresponding transition kernel in $\mathcal{P}(\mathbb{A}|\mathbb{X})$. For players $i, j \in \mathcal{N}$, if the policies $\pi^i \in \Pi^i_S, \pi^j \in \Pi^j_S$ are both of the local type and are identified with the same transition kernel, we say π^i and π^j are *locally symmetric*.

Lemma 54 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumptions 3 and 5 hold. Let* $\boldsymbol{\pi} \in \mathbf{\Pi}_S$ *be soft. For* $i, j \in \mathcal{N}$*, suppose* π^i *and* π^j *are locally symmetric. Then, we have* $\pi^i \in \text{Subj-BR}^i_\epsilon(\boldsymbol{\pi}^{-i}, \mathcal{V}^*, \mathcal{W}^*)$ *if and only if* $\pi^j \in \text{Subj-BR}^j_\epsilon(\boldsymbol{\pi}^{-j}, \mathcal{V}^*, \mathcal{W}^*)$*.*

The proof of Lemma 54 mirrors the proof of Lemma 39 and is therefore omitted.

Proposition 55 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game for which Assumptions 3 and 5 hold. Let* $\epsilon > 0$*. Assume that* $\mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*. Suppose* $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ *is a subset of policies satisfying (i)* $\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*; (ii) the set* $\widehat{\mathbf{\Pi}}$ *is symmetric, i.e.* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for each* $i, j \in \mathcal{N}$*; (iii) each policy* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$ *is soft. Then,* $\mathbf{G}$ *has the* $(\mathcal{V}^*, \mathcal{W}^*)$*-subjective* ϵ*-satisficing paths property within* $\widehat{\mathbf{\Pi}}$*.*

As with Proposition 52, Proposition 55 can be proved using the argument of (Yongacoglu et al., 2023, Theorem 3.6), suitably modified to use Lemma 54 instead of (Yongacoglu et al., 2023, Corollary 2.9).

Convergence of the Oracle Algorithm under Global Observability

Lemma 56 *Let* $\mathbf{G}$ *be a partially observed n-player mean-field game satisfying Assumptions 1 and 5, and let* $\epsilon > 0$. *Let* $\widehat{\mathbf{\Pi}} \subset \mathbf{\Pi}_S$ *be a quantization of the subset of policies of the mean-field type,* $\{\boldsymbol{\pi} \in \mathbf{\Pi}_S : \pi^i$ *is of the mean-field type* $\forall i \in \mathcal{N}\}$. *Suppose* $\widehat{\mathbf{\Pi}}$ *satisfies (1)* $\widehat{\mathbf{\Pi}} \cap \mathrm{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$; *(2)* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for all* $i, j \in \mathcal{N}$; *and (3) every policy* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$ *is soft.*

Suppose that each agent $i \in \mathcal{N}$ *updates is policy sequence* $\{\pi_k^i\}_{k\geq 0}$ *according to Algorithm 2 and that, for each* $k \geq 0$, *the policy updates for* $\boldsymbol{\pi}_{k+1}$ *are conditionally independent across agents given* $\boldsymbol{\pi}_k$. *Then,* $\lim_{k\to\infty} \Pr(\boldsymbol{\pi}_k \in \widehat{\mathbf{\Pi}} \cap \mathbf{\Pi}^{\epsilon\text{-eq}}) = 1$.

The proof of Lemma 56 is essentially the same as that of Lemma 24. One important difference between Lemmas 56 and 24 is such: in the former, convergence to an objective equilibrium is guaranteed, while in the latter one only has convergence to a subjective equilibrium.

Appendix J. Learning Results Under Other Observation Channels

J.1 Learning with Global Observability

We now present convergence results for Algorithm 3 under global observability, the richest of the information structures that we consider. Under Assumption 1, strong convergence guarantees can be made. These are presented below in Theorem 57. In order to state this result, we now fix $\epsilon > 0$ and make the following assumptions on the various parameters of Algorithm 3.

Assumption 12 *Fix* $\epsilon > 0$. *For each* $i \in \mathcal{N}$ *define* $\Pi^i_{S,\mathrm{MF}}$ *as*

$$\Pi^i_{S,\mathrm{MF}} := \{\pi^i \in \Pi^i_S : \pi^i \text{ is of the mean-field type.}\}.$$

Assume that $\widehat{\mathbf{\Pi}} \subset \times_{i\in\mathcal{N}} \Pi^i_{S,\mathrm{MF}}$ *is a fine quantization of* $\times_{i\in\mathcal{N}} \Pi^i_{S,\mathrm{MF}}$ *satisfying: (1)* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for all* $i, j \in \mathcal{N}$; *(2)* $\widehat{\mathbf{\Pi}} \cap \mathbf{\Pi}^{\epsilon\text{-eq}} \neq \varnothing$; *(3) For any* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$, $\boldsymbol{\pi}$ *is soft.*

Next we present a restriction on the parameters $\{d^i\}_{i\in\mathcal{N}}$. 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 small, so that poorly performing policies are not mistaken for ϵ-best-responses. The bound $\bar{d}_G$ below is defined analogous to the term $\bar{d}$ in Yongacoglu et al. (2023) and depends on both ϵ and $\widehat{\mathbf{\Pi}}$.

Assumption 13 *For each player* $i \in \mathcal{N}$, $d^i \in (0, \bar{d}_G)$, *where* $\bar{d}_G = \bar{d}_G(\epsilon, \widehat{\mathbf{\Pi}})$ *is defined as* $\bar{d}_G := \min \mathcal{O}_G$, *where* $\mathcal{O}_G := S_G \setminus \{0\}$, *and* S_G *is given by*

$$S_G := \left\{ \left| \epsilon - \left(V^{*i}_{\boldsymbol{\pi}}(y) - \min_{a^i \in \mathbb{A}} W^{*i}_{\boldsymbol{\pi}}(y, a^i) \right) \right| : i \in \mathcal{N}, \boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}, y \in \mathbb{Y} \right\}.$$

Theorem 57 *Let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumptions 1 and 5, and let* $\epsilon > 0$*. Suppose the policy set* $\widehat{\mathbf{\Pi}}$ *and the tolerance parameters* $\{d^i\}_{i\in\mathcal{n}}$ *satisfy Assumptions 12 and 13, and suppose all players follow Algorithm 3. For any* $\xi > 0$*, there exists* $\tilde{T} = \tilde{T}(\xi, \epsilon, \widehat{\mathbf{\Pi}}, \{d^i\}_{i\in\mathcal{n}})$ *such that if* $T_k \geq \tilde{T}$ *for all* k*, then*

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

for all sufficiently large k*.*

The details of the proof of Theorem 57 resemble those of Theorem 25, with the following exceptions: here, $\bar{d}_G$ replaces $\bar{d}_{\text{MF}}$, and the term Ξ is now defined as $\Xi := \frac{1}{2}\min_{i\in\mathcal{n}}\{d^i, \bar{d}_G - d^i\}$. With these modifications, the mechanics of the proofs are identical.

J.2 Learning with Compressed Observability State

We conclude with a discussion of convergence guarantees for Algorithm 3 under compressed observability (Assumption 3). As we have discussed previously, there is no guarantee that the set $\mathbf{\Pi}_S$ contains ϵ-equilibrium policies—either in the objective sense or in the subjective sense using the naively learned subjective function family $(\mathcal{V}^*, \mathcal{W}^*)$. As a result, the convergence guarantees in this setting are highly qualified, and must be made with a potentially restrictive assumption that subjective ϵ-equilibrium exist within the set of policies $\widehat{\mathbf{\Pi}}$.

Assumption 14 *Fix* $\epsilon > 0$*. Assume that the set of* $(\mathcal{V}^*, \mathcal{W}^*)$*-subjective* ϵ*-equilibrium is non-empty. That is,* $\text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*.*

Assumption 15 *Fix* $\epsilon > 0$*. Assume that* $\widehat{\mathbf{\Pi}}$ *is a fine quantization of* $\mathbf{\Pi}_S$ *satisfying: (1)* $\widehat{\Pi}^i = \widehat{\Pi}^j$ *for all* $i, j \in \mathcal{n}$*; (2)* $\widehat{\mathbf{\Pi}} \cap \text{Subj}_\epsilon(\mathcal{V}^*, \mathcal{W}^*) \neq \varnothing$*; (3) For any* $\boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}$*,* $\boldsymbol{\pi}$ *is soft.*

Assumption 16 *For all* $i \in \mathcal{n}$*,* $d^i \in (0, \bar{d}_{\text{comp}})$*, where* $\bar{d}_{\text{comp}} = \bar{d}_{\text{comp}}(\epsilon, \widehat{\mathbf{\Pi}})$ *is defined as* $\bar{d}_{\text{comp}} := \min \mathcal{O}_{\text{comp}}$*, where* $\mathcal{O}_{\text{comp}} := S_{\text{comp}} \setminus \{0\}$ *and*

$$S_{\text{comp}} := \left\{ \left| \epsilon - \left(V^{*i}_{\boldsymbol{\pi}}(y) - \min_{a^i \in \mathbb{A}} W^{*i}_{\boldsymbol{\pi}}(y, a^i) \right) \right| : i \in \mathcal{n}, \boldsymbol{\pi} \in \widehat{\mathbf{\Pi}}, y \in \mathbb{Y} \right\}.$$

Theorem 58 *Let* $\epsilon > 0$ *and let* $\mathbf{G}$ *be a partially observed* n*-player mean-field game satisfying Assumptions 3, 5, and 14. Suppose the policy set* $\mathbf{\Pi}$ *and the tolerance parameters* $\{d^i\}_{i\in\mathcal{n}}$ *satisfy Assumptions 15 and 16, and suppose all players follow Algorithm 3. For any* $\xi > 0$*, there exists* $\tilde{T} = \tilde{T}(\xi, \epsilon, \widehat{\mathbf{\Pi}}, \{d^i\}_{i\in\mathcal{n}})$ *such that if* $T_k \geq \tilde{T}$ *for all* k*, then*

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

for all sufficiently large k*.*

The proof of Theorem 58 parallels the proof of Theorem 25, with $\bar{d}_{\text{comp}}$ replacing $\bar{d}_{\text{MF}}$ and Ξ redefined as $\Xi := \frac{1}{2}\min_{i\in\mathcal{n}}\{d^i, \bar{d}_{\text{comp}} - d^i\}$. The rest of the proof goes through unchanged. We note that the requisite satisficing paths structure holds by Assumption 15 and the usual line of argument following Proposition 55