2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3503-2814-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/Allerton58177.2023.10313387

Bayesian Observational Learning in Mean-Field Games with Imperfect Observations

Pawan Poojary and Randall Berry

***Abstract*— There are many settings in which agents learn from observing the actions of other agents. Bayesian observational learning models provide a framework for studying such situations and have been well-studied in settings where agents sequentially choose Bayes' optimal actions by learning from the actions of previous agents. Here, we consider such observational learning in a mean-field game setting, in which agents repeatedly choose actions over time to maximize an infinite horizon discounted pay-off. This pay-off depends on the underlying mean-field population state, which agents do not know and only have a prior common belief over it. At the end of each time-step, agents observe a common signal which is an imperfect observation of the mean-field action profile played in that time-step and use this to update their beliefs. We give a sequential decomposition of this game that enables one to characterize Markov perfect equilibria of the game. We then focus on a particular sub-class of these games which can be viewed as a mix of coordination/anti-coordination players. Using the sequential decomposition, we characterize the impact of varying the observation quality on the outcome of the game and show that this can exhibit non-monotonic behaviour, where in many instances, poorer observations lead to better expected total discounted pay-offs.**

I. Introduction

Bayesian observational learning models provide a framework for studying situations in which Bayesian agents seek to learn from observing the actions of other agents. Early models for such settings include [1]–[3], which considered a setting in which a homogeneous set of Bayesian rational agents sequentially take binary actions to optimize a pay-off that depends on an unknown binary state of the world. Each agent privately receives an independent binary signal about this state of the world and also observes the actions of previous agents, which may provide additional information about the private signals those agents received. A key result in these models is that a *herding* behavior can occur, where at some point in time, it is optimal for an agent to ignore its private signal and follow the action taken by the majority of previous agents. Subsequent agents then follow suit due to their homogeneity; hence learning stops. There have been many variations of this basic model subsequently studied including [4]–[12]. In particular, we highlight work on models in which the observations available to agents are imperfect due to the presence of observation noise [13] or "fake" agents that do not correctly report their actions [14].

This work was supported in part by the NSF under grants CNS-1908807 and ECCS-2216970.

P. Poojary and R. Berry are with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA. pawanpoojary2018@u.northwestern.edu, rberry@ece.northwestern.edu

The aforementioned papers all consider settings in which agents sequentially take an action, and each agent only acts once. Instead, in this paper, we consider observational learning in a mean-field game setting [15] in which players repeatedly take actions in a discrete-time mean-field game with an infinite horizon discounted pay-off. Each player has a private type and its pay-off depends on the underlying mean-field type distribution of the players, which is not known to the players. We assume that the players instead have a common prior belief over this type distribution. Further, after each time-slot, the players obtain an imperfect observation of the mean-field action profile, which they can use to update their beliefs. We study how these belief updates evolve. This type of imperfect observation of players' actions can also be viewed as a form of imperfect public monitoring [16]. To this end, we first show a sequential decomposition of this game as in [17], which also considers mean-field games with private types. However, the mean-field action profile is *perfectly* observed in [17], while in our work this observation is *imperfect*. Other related sequential decompositions for dynamic games include [18]–[20].

We next consider a specific class of games in which there are two types of players choosing binary valued actions: one type seeks to coordinate its action with the other players, while the other type seeks to anti-coordinate.[1] Further, we assume that the mean-field type distribution takes one of two possible values over which the players have a common prior belief. After each time-slot, the players receive a common binary valued signal regarding the resulting mean-field action profile. We use our sequential decomposition to characterize the possible Markov perfect equilibria of this game and for two particular choices of equilibria, we characterize how the players' beliefs evolve over time. Similar to the sequential models for observational learning, we show the possibility of "herding" behavior. Here, herding corresponds to all players choosing their actions independent of their private types, so that the resulting signals obtained do not convey any information about the mean-field type distribution, and thus learning stops. We also characterize how changing the observation quality impacts the players' expected discounted total pay-offs and show that, as in [13], [14], in several cases "better" information leads to lower pay-offs.

A. Notation

We use uppercase letters for random variables and lowercase for their realizations. We use notation $-i$ to

[1] This is related to models of mixed coordination and anti-coordination games, e.g. [21]

represent all players other than player i, i.e., $-i = \{1, 2, \ldots, i-1, i+1, \ldots, N\}$. Similarly, we use a_t^{-i} to mean $(a_t^1, a_t^2, \ldots, a_t^{i-1}, a_t^{i+1} \ldots, a_t^N)$. We use notation $a_{1:t}$ to represent the vector $(a_1, a_2, \ldots, a_t)$. We denote the indicator function of any set A by $\mathbb{1}\{A\}$. For any set $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ represents the space of probability measures on $\mathcal{S}$. We denote the probability measure generated by (or expectation with respect to) a strategy profile σ by P^σ (or $\mathbb{E}^\sigma$). For a probabilistic strategy profile of players $(\sigma_t^i)_{i\in[N]}$ where probability of action a_t^i conditioned on $(o_{1:t}, x^i)$ is given by $\sigma_t^i(a_t^i|o_{1:t}, x^i)$, we use the short-hand notation $\sigma_t^{-i}(a_t^{-i}|o_{1:t}, x^{-i})$ to represent the probablity $\prod_{j\neq i} \sigma_t^j(a_t^j|o_{1:t}, x^j)$ of action a_t^{-i} conditioned on $(o_{1:t}, x^{-i})$. All equalities and inequalities involving random variables are to be interpreted in *a.s.* sense.

II. Model and MPE solution

We consider an infinite-horizon discrete-time large population sequential game $\mathbb{G}$ as follows. There are a countably infinite number of homogenous players. Let $[N] = \{1, 2, \ldots\}$ denote this set of players and $[T] = \{1, 2, \ldots\}$ be the set of discrete time indices. To begin with, each player $i \in [N]$ observes its private type $x^i \in \mathcal{X} = \{1, 2, \cdots, N_x\}$. Let the mean-field of player types be $z = \big(z(1), z(2), \ldots, z(N_x)\big)$, where $z(x)$ denotes the fraction of population having type $x \in \mathcal{X}$, i.e.,

$$z(x) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{x^i = x\}, \tag{1}$$

and $\sum_{i=1}^{N_x} z(i) = 1$. Thus, $z \in \mathcal{Z}$, where $\mathcal{Z}$ is the $(N_x - 1)$-dimensional probability simplex. Further, we assume that at time $t = 1$, players do not know the mean-field z, but have a common prior belief $\pi_1 \in \mathcal{P}(\mathcal{Z})$ over it.

Now, at time $t \in [T]$, each player $i \in [N]$ takes an action $a_t^i \in \mathcal{A} = \{1, 2, \cdots, N_a\}$ which similarly generates a mean-field of players' actions: $y_t = (y_t(1), y_t(2), \ldots, y_t(N_a))$. Here $y_t(a)$ denotes the fraction of actions having type $a \in \mathcal{A}$ at time t, i.e.,

$$y_t(a) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{a_t^i = a\}, \tag{2}$$

and $\sum_{i=1}^{N_a} y_t(i) = 1$. Players do not observe other's actions or the resulting y_t, but instead observe a common signal $o_t \in \mathcal{O}$ which partially reveals y_t through a noisy channel represented through the kernel, $o_t \sim Q(\cdot|y_t)$. Here, conditioned on $\{y_t\}_{t=1}^\infty$, the sequence of observations $\{o_t\}_{t=1}^\infty$ are assumed to be mutually independent. Each player i then receives a reward $R(x^i, a_t^i, y_t)$ which is a function of its private type x^i, action a_t^i and the action mean-field y_t. The rewards are assumed to be bounded and unobservable to the players during the course of the game.

Remark 1: In this game, the mean-field of player types does not evolve over time; however, players' common belief about it will evolve based on their observations.

Let $\mathcal{O}^{t-1}$ denote the space of observations $o_{1:t-1}$ commonly observed before time t. At time t, player i's observation history is denoted by $h_t^i = (o_{1:t-1}, x^i)$. The player then takes action a_t^i according to a behavioral strategy $\sigma^i = (\sigma_t^i)_t$, where $\sigma_t^i : \mathcal{O}^{t-1} \times \mathcal{X} \to \mathcal{P}(\mathcal{A})$. We denote the space of such strategies as $\mathcal{K}^\sigma$. This implies $A_t^i \sim \sigma_t^i(\cdot|o_{1:t-1}, x^i)$. For the game $\mathbb{G}$, each player i wants to play a strategy $\sigma^i \in \mathcal{K}^\sigma$, that maximizes its expected total discounted reward over an infinite-time horizon, with discount factor $0 < \delta < 1$, which is given by

$$J^i := \mathbb{E}^{\sigma^i}\left[\sum_{t=1}^{\infty} \delta^{t-1} R(x^i, A_t^i, Y_t) \,\Big|\, \pi_1, x^i\right]. \tag{3}$$

A. Solution concept: MPE

The Nash equilibrium (NE) of $\mathbb{G}$ is defined as strategies $\tilde\sigma = (\tilde\sigma^i)_{i\in[N]}$ that satisfy, for all $i \in [N]$,

$$\begin{aligned}\mathbb{E}^{(\tilde\sigma^i, \tilde\sigma^{-i})}\Big[\sum_{t=1}^{\infty} \delta^{t-1} R(x^i, A_t^i, Y_t)\Big] \\ \geq\ \mathbb{E}^{(\sigma^i, \tilde\sigma^{-i})}\Big[\sum_{t=1}^{\infty} \delta^{t-1} R(x^i, A_t^i, Y_t)\Big],\end{aligned} \tag{4}$$

For sequential games, however, a more appropriate equilibrium concept is Markov perfect equilibrium (MPE) [22], which we use in this paper. We note that an MPE is also a Nash equilibrium of the game, although not every Nash equilibrium is an MPE. An MPE $(\tilde\sigma)$ satisfies sequential rationality such that, for any history $h_t^i = (o_{1:t-1}, x^i)$, and for any $i \in [N], t \in [T]$,

$$\begin{aligned}\mathbb{E}^{(\tilde\sigma^i \tilde\sigma^{-i})}\Big[\sum_{n=t}^{\infty} \delta^{n-t} R(x^i, A_n^i, Y_n) \,\Big|\, o_{1:t-1}, x^i\Big] \\ \geq \mathbb{E}^{(\sigma^i \tilde\sigma^{-i})}\Big[\sum_{n=t}^{\infty} \delta^{n-t} R(x^i, A_n^i, Y_n) \,\Big|\, o_{1:t-1}, x^i\Big].\end{aligned} \tag{5}$$

We now provide an MPE solution for $\mathbb{G}$. We consider a Markovian equilibrium strategy for each player i, which at time t, depends on the common information $o_{1:t-1}$ through a belief π_t (defined below in (7)) on the mean-field of player types, and on the player's private type x^i. Equivalently, player i takes action of the form $A_t^i \sim \sigma_t^i(\cdot|\pi_t, x^i)$. Similar to the common agent approach in [23], an alternate and equivalent way of defining the strategies of the players is as follows. We first generate a prescription $\gamma_t^i : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ as a function of the belief π_t through an equilibrium generating function $\theta_t^i : \mathcal{P}(\mathcal{Z}) \to \big(\mathcal{X} \to \mathcal{P}(\mathcal{A})\big)$ such that $\gamma_t^i = \theta_t^i[\pi_t]$. Then, action A_t^i is generated by applying the prescription γ_t^i on player i's private type x^i, i.e. $A_t^i \sim \gamma_t^i(\cdot|x^i)$. Thus, $A_t^i \sim \sigma_t^i(\cdot|\pi_t, x^i) = \theta_t^i[\pi_t](\cdot|x^i)$. We only consider symmetric equilibria of such games. We can thereby drop the dependence of i on the functions θ_t^i and γ_t^i, and have $A_t^i \sim \gamma_t(\cdot|x^i) = \theta_t[\pi_t](\cdot|x^i)$.

We now define some pre-requisites for the MPE solution. At time t, given the mean-field of player-types z, the action mean-field y_t that results from the players applying a symmetric prescription function γ_t can be computed as:

$$y_t(a) = \sum_{x\in\mathcal{X}} \gamma_t(a|x)\, z(x), \quad \forall\, a \in \mathcal{A},$$

which we represent by function f such that

$$y_t = f(z, \gamma_t). \tag{6}$$

Further, let the common belief on $z \in \mathcal{Z}$ at time t, denoted by π_t be defined as:

$$\pi_t(z) \triangleq \mathbb{P}(Z = z | \gamma_{1:t-1}, o_{1:t-1}), \tag{7}$$

with π_1 being the initial common belief of the players. Then, π_t can be updated according to:

$$\pi_{t+1} = \phi(\pi_t, \gamma_t, o_t), \tag{8}$$

where the update function ϕ is described as follows:

$$\pi_{t+1}(z) = \frac{Q(o_t|y_t)\pi_t(z)}{\sum_{\tilde{z}\in\mathcal{Z}} Q(o_t|\tilde{y}_t)\pi_t(\tilde{z})}, \quad \forall\, z \in \mathcal{Z}, \tag{9}$$

with $y_t = f(z, \gamma_t)$ and $\tilde{y}_t = f(\tilde{z}, \gamma_t)$. Lastly, note that through (6), the belief π_t on the mean-field of player-types translates to the belief $P_Y(\cdot|\pi_t, \gamma_t)$ on the action mean-field y_t,

$$Y_t \sim P_Y(y|\pi_t, \gamma_t) := \sum_{z\in\mathcal{Z}} \mathbb{1}_{\{y = f(z,\gamma_t)\}} \pi_t(z). \tag{10}$$

With the above pre-requisites, we can now provide a symmetric MPE of $\mathbb{G}$. Note that since $\mathbb{G}$ has an infinite-time horizon, there exist MPE solution functions $\tilde{\theta}_t$ and $\tilde{\gamma}_t$ that are stationary, thus the dependence on t is dropped. Then, the equilibrium strategy is defined as $A_t^i \sim \tilde{\gamma}(\cdot|x^i) = \tilde{\theta}[\pi_t](\cdot|x^i)$. In addition, we generate a reward-to-go function $V : \mathcal{P}(\mathcal{Z}) \times \mathcal{X} \to \mathbb{R}$. Then $(\tilde{\gamma}, V)$ are obtained as solutions of the following fixed-point equations, for all $x^i \in \mathcal{X}, \pi_t \in \mathcal{P}(\mathcal{Z})$:

$$\tilde{\gamma}(\cdot|x^i) \in \underset{\gamma(\cdot|x^i)}{\arg\max}\, \mathbb{E}^{\gamma(\cdot|x^i)} \Big[R(x^i, A_t^i, Y_t) + \delta V(\phi(\pi_t, \tilde{\gamma}_t, O_t), x^i) \,\Big|\, \pi_t, x^i \Big], \tag{11a}$$

$$V(\pi_t, x^i) \triangleq \mathbb{E}^{\tilde{\gamma}(\cdot|x^i)} \Big[R(x^i, A_t^i, Y_t) + \delta V(\phi(\pi_t, \tilde{\gamma}_t, O_t), x^i) \,\Big|\, \pi_t, x^i \Big], \tag{11b}$$

where the expectations in (11) are with respect to the random variable (r.v.) (A_t^i, Y_t, O_t) through the measure $\gamma(a_t^i|x^i) P_Y(y_t|\pi_t, \tilde{\gamma})\; Q(o_t|y_t)$.

Note that within the expectation in (11a), the second term does not have any dependence of $\gamma(\cdot|x^i)$. This implies that $\tilde{\gamma}(\cdot|x^i)$ can be obtained without knowing the V function defined in (11b), by simply solving the fixed point equation:

$$\tilde{\gamma}(\cdot|x^i) \in \underset{\gamma(\cdot|x^i)}{\arg\max}\, \mathbb{E}^{\gamma(\cdot|x^i)} \left[R(x^i, A_t^i, Y_t) \,\middle|\, \pi_t, x^i \right]. \tag{12}$$

Remark 2: A solution of eq. (12) is in fact a NE of the static version of game $\mathbb{G}$, with a common mean-field belief π_t on the player types.

Thus, with $\tilde{\gamma}(\cdot|x^i) = \tilde{\theta}[\pi_t](\cdot|x^i)$ defined as per (12) for each $\pi_t \in \mathcal{P}(\mathcal{Z})$, the MPE strategy at time t is defined as

$$\tilde{\sigma}_t^i(a_t^i|\pi_t, x^i) = \tilde{\gamma}(a_t^i|x^i), \tag{13}$$

where $\tilde{\gamma} = \tilde{\theta}[\pi_t]$. The next theorem shows that the strategy obtained by (12) and (13) is an MPE of the game.

Theorem 1: A strategy $(\tilde{\sigma})$ obtained by solving (12) and defined as per (13) is an MPE of game $\mathbb{G}$, i.e., it satisfies the sequential rationality property stated in (5), for any history $h_t^i = (o_{1:t-1}, x^i)$, for all $i \in [N], t \in [T]$.

Refer to the Appendix for a detailed proof. We now define an important phenomenon that may occur during the course of the game.

Definition 1: Herding is said to occur when the common mean-field belief π_t becomes such that the players' MPE strategy $\tilde{\gamma} = \tilde{\theta}[\pi_t]$ is independent of their private types.

A consequence of Definition 1 is as follows. If players herd at time t under belief π_t, then the signal o_t observed by the players does not convey any new information about the true value of the mean-field z. As a result, belief π_t stops updating and players continue to play the strategy $\tilde{\gamma} = \tilde{\theta}[\pi_t]$ for all $\tau \geq t$. Therefore, from the onset of herding, the players' actions do not reveal any information about the mean-field; hence learning stops.

Property 1: Once herding of players occurs, it lasts forever.

III. An example game

In this section, we consider a particular example $\mathbb{G}_E$ of the dynamic game $\mathbb{G}$, as follows. Let each player $i \in [N]$ have a private type $x^i \in \mathcal{X} = \{-1, 1\}$. Then, the mean-field of player types is the probability vector $z = \big(z(-1), z(1)\big)$. We assume that this mean-field takes only two possible values, that is $z \in \mathcal{Z} = \{(1-p_0, p_0), (1-p_1, p_1)\}$, where $0 < p_0 < p_1 < 1$. We further assume that this is common knowledge, i.e., every player knows that the mean-field will take on only one of these values. In this case, players' prior common belief on z, i.e., $\pi_1 \in \mathcal{P}(\mathcal{Z})$ can be parameterized by $d_1 \in (0, 1)$, where $d_1 := \pi_1[z = (1-p_1, p_1)]$. Now, at any time $t \in [T]$, each player i takes action $a_t^i \in \mathcal{A} = \{-1, 1\}$. This then generates the action mean-field $y_t = (y_t(-1), y_t(1))$ which is not observable to players. Instead, the players observe a common signal $o_t \in \{H\,(\text{high}), L\,(\text{low})\}$. This signal, as shown in Figure 1, partially reveals whether $y_t(1) \geq 0.5$ through a binary symmetric channel (BSC) with crossover probability $1-u$, where $u \in (0.5, 1]$ is the signal quality. Each player i then receives a reward:

$$R(x^i, a_t^i, y_t) := \big(2y_t(1) - 1\big)\, x^i a_t^i, \tag{14}$$

which implies that a player of type $x^i = 1$ prefers the action that will be taken by the greater fraction of the population. Whereas, a player of type $x^i = -1$ prefers to take the action

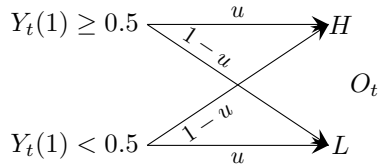

Fig. 1: The BSC through which players receive the common signal o_t at each t, where $u \in (0.5, 1]$ denotes the signal quality.

that will be taken by the lesser fraction of the population. Games with such a reward function relate to the work on mixed coordination and anti-coordination games [21].

A. MPE solution

Now, to provide the MPE solution for this game, first let the prescription $\gamma : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ be represented as $\gamma = [q, r] := [\gamma(1|-1), \gamma(1|1)]$, where $q, r \in [0, 1]$. In other words, q (r) is the probability that that a type -1 (1) player chooses action 1. If the symmetric prescription $[q, r]$ acts upon the belief d_t of the mean-field z, then as per (10), the action mean-field is a r.v. Y_t such that

$$Y_t(1) = \begin{cases} (1-p_1)q + p_1 r, & \text{w.p. } d_t, \\ (1-p_0)q + p_0 r, & \text{w.p. } (1-d_t). \end{cases} \tag{15}$$

The expectation of Y_t is then

$$\mathbb{E}^{[q,r]}\big[\, Y_t(1) \,\big|\, d_t \,\big] = B_t r + (1 - B_t)q,$$
$$B_t := p_1 d_t + p_0(1 - d_t) \in (p_0, p_1). \tag{16}$$

Here, B_t is the expected value of $z(1)$ under belief d_t. Also, let $A_t := 1 - B_t$. By applying this expectation to (14), each player i, which chooses action a_t^i as per strategy $[q, r]$, then receives the expected reward:

$$\mathbb{E}^{[q,r]}\big[R(x^i, a_t^i, y_t) \,\big|\, d_t\big] = \begin{cases} \big\{2\mathbb{E}^{[q,r]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\big\}(2r - 1), & \text{for } x^i = 1, \\ \big\{2\mathbb{E}^{[q,r]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\big\}(1 - 2q), & \text{for } x^i = -1. \end{cases} \tag{17}$$

The MPE solution of this game, $\tilde{\gamma} := [\tilde{q}, \tilde{r}]$ for a given belief d_t can then be obtained by applying the fixed point equation in (12) for $x^i = 1$ and -1, which are respectively given by

$$\tilde{r} = \arg\max_r \Big\{2\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\Big\}(2r - 1), \tag{18a}$$

$$\tilde{q} = \arg\max_q \Big\{2\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\Big\}(1 - 2q). \tag{18b}$$

When $B_t < 0.5$, it can be shown that the only possible solutions to these fixed point equations are for $\tilde{q}$ and $\tilde{r}$ to be chosen such that the term $\mathbb{E}^{[\tilde{q},\tilde{r}]}[Y_t(1) \,|\, d_t] = 1/2$, which results in all players' expected rewards, given in (17), to be zero. Any such choice of $[\tilde{q}, \tilde{r}]$ is a MPE. When $\tilde{B}_t \geq 0.5$, such a choice of $[\tilde{q}, \tilde{r}]$ is again a MPE as well as setting $[\tilde{q}, \tilde{r}]$ to be either $[0, 1]$ or $[1, 0]$. We can then summarize the set of MPEs as:

$$\text{MPE} = \begin{cases} \big\{\, [\tilde{q}, \tilde{r}] : \tilde{q} = \frac{0.5 - \tilde{r}B_t}{A_t},\ \tilde{r} \in [0, 1] \,\big\}, & \text{for } B_t < 0.5, \\ \big\{\, [\tilde{q}, \tilde{r}] : \tilde{r} = \frac{0.5 - \tilde{q}A_t}{B_t},\ \tilde{q} \in [0, 1] \,\big\} & \text{for } B_t \geq 0.5. \\ \qquad\qquad \cup \{[1, 0], [0, 1]\}, \end{cases} \tag{19}$$

The corresponding fixed point equation for the reward-to-go function V, given by (11b), simplifies for $x^i = 1$ and -1 as given below.

$$V(d_t, 1) = \Big\{2\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\Big\}(2\tilde{r} - 1), + \delta\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, V(d_{t+1}, 1) \,\big|\, d_t\big], \tag{20a}$$

$$V(d_t, -1) = \Big\{2\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, Y_t(1) \,\big|\, d_t \,\big] - 1\Big\}(1 - 2\tilde{q}), + \delta\mathbb{E}^{[\tilde{q},\tilde{r}]}\big[\, V(d_{t+1}, -1) \,\big|\, d_t\big]. \tag{20b}$$

Here, $d_{t+1} = \phi(d_t, [\tilde{q}, \tilde{r}], o_t)$, which is a simplified representation of the belief update equation in (8).

In the next section, we analyze the game $\mathbb{G}_E$ under an MPE strategy, that exhibits herding. We denote this strategy by $[\tilde{q}, \tilde{r}]^H$, where the superscript indicates that players could herd under this strategy. From this analysis, we aim to understand how herding, which stifles learning, impacts the aggregate reward accrued by each player and how is it affected by varying the quality of the observation signals.

IV. Learning Dynamics under $[\tilde{q}, \tilde{r}]^H$

We now analyze the dynamics of the sequence of mean-field belief states $\{d_t\}_{t=1}^{\infty}$ for the game $\mathbb{G}_E$, where all players play strategy $[\tilde{q}, \tilde{r}]^H$, which we define as follows:

$$[\tilde{q}, \tilde{r}]^H = \begin{cases} [1/2, 1/2], & \text{for} \quad B_t < 0.5, \\ [0, 1], & \text{for} \quad B_t \geq 0.5. \end{cases} \tag{21}$$

Now, as per the definition of B_t in (16), the inequality $B_t < (\geq)\ 0.5$ translates to $d_t < (\geq)\ d_{\text{th}}$. Here $d_{\text{th}} := (0.5 - p_0)/(p_1 - p_0)$ refers to a *threshold* belief, which when crossed results in players switching their strategies, as per the MPE strategy profile in (21), and more generally in (19). We then make the following assumption on the parameters of the game to ensure that such a threshold exists, which thereby ensures that both herding and non-herding belief states exist.

Assumption 1: Let $p_0 < 1/2 < p_1$ for the set of possible mean-field values: $\mathcal{Z} = \{(1 - p_0, p_0), (1 - p_1, p_1)\}$, as it ensures that the threshold belief d_{th} exists.

We now compute the updates on d_t, brought about by observing the common signal o_t. First, consider the case $d_t \geq d_{\text{th}}$, which is when players play the strategy $[\tilde{q}, \tilde{r}]^H = [0, 1]$. Applying this strategy in (15) generates a random $Y_t(1)$ such that $Y_t(1) = p_1$ when $z = (1 - p_1, p_1)$, which occurs w.p. d_t; otherwise $Y_t(1) = p_0$. As $p_0 < 1/2 < p_1$ (Assumption 1), this implies that $Y_t(1) \geq 0.5$ w.p. d_t; otherwise $Y_t(1) < 0.5$. Now, applying $Y_t(1)$ as input to the BSC shown in Figure 1 and then observing the output O_t yields the following updates for d_t.

$$d_{t+1} = \begin{cases} \dfrac{u d_t}{u d_t + (1-u)(1-d_t)}, & \text{if } O_t = H, \\ \dfrac{(1-u) d_t}{(1-u) d_t + u(1-d_t)}, & \text{if } O_t = L, \end{cases} \tag{22}$$

where $\mathbb{P}(O_t = H \,|\, z, d_t \geq d_{\text{th}}) = u$ if the true value of the mean-field, $z = (1 - p_1, p_1)$, and equals $1 - u$ if $z = (1 - p_0, p_0)$. Next, consider the case: $d_t < d_{\text{th}}$. In this case, as per (21), players play the strategy $[\tilde{q}, \tilde{r}]^H = [1/2, 1/2]$, which is evidently independent of the players' private types, i.e., players are herding (Definition 1). This implies that no information about the true mean-field value z gets conveyed by O_t. Therefore, d_t stops updating and players continue herding to the stratgey $[\tilde{q}, \tilde{r}]^H = [1/2, 1/2]$ (Property 1).

Property 2: In game $\mathbb{G}_E$, under MPE strategy $[\tilde{q},\tilde{r}]^H$ given in (21), herding begins only when $d_t < d_{\text{th}}$ (or $B_t < 0.5$).

To better understand the dynamics of mean-field belief d_t, we instead consider its *likelihood ratio* function: $l_t(d_t) := \mathbb{P}[(1-p_1,p_1)|d_t) \,/\, \mathbb{P}((1-p_0,p_0)|d_t] = d_t/(1-d_t)$. Also, let $l_{\text{th}} := d_{\text{th}}/(1-d_{\text{th}})$ be the the threshold likelihood ratio. Then, the updates for d_t stated in (22) when $d_t \geq d_{\text{th}}$ translate to the following updates on l_t when $l_t \geq l_{\text{th}}$:

$$l_{t+1} = \begin{cases} \left(\frac{u}{1-u}\right) l_t, & \text{if } O_t = H, \\ \left(\frac{1-u}{u}\right) l_t, & \text{if } O_t = L. \end{cases} \tag{23}$$

Now, if the players have not herded at time t, then they also did not herd at any prior time (Property 1), i.e., $l_n \geq l_{\text{th}}$ for all $n \leq t$. Then, as a result of the above updates, l_t can be shown to depend only on the number of H's (denoted by n_H) and L's (denoted by n_L) in the observation history $o_{1:t-1}$. Specifically, $l_t = l_1\left(\frac{u}{1-u}\right)^{h_t}$ where h_t is the difference between the number of H's and the number of L's,

$$h_t = n_H - n_L. \tag{24}$$

Property 3: Until herding occurs, h_t defined in (24) is a sufficient statistic of the information contained in the past observations $o_{1:t-1}$.

We now make the following assumption on the players' prior belief d_1 on the mean-field of their private types.

Assumption 2: To understand the dynamics of learning, we assume that at the start of the game, players are not herding, i.e., $d_1 \geq d_{\text{th}}$, where d_1 is the prior mean-field belief.

Herding of players begins at the first instance when $d_t < d_{\text{th}}$ (Property 2), which implies $l_t < l_{\text{th}}$. Applying the expression for l_t in this inequality yields the equivalent condition on the integer-valued sufficient statistic h_t, which is $h_t < -k$, where

$$k = \left\lfloor \left[\log\left(\frac{d_1}{1-d_1}\right) - \log\left(\frac{d_{\text{th}}}{1-d_{\text{th}}}\right)\right] / \log\left(\frac{u}{1-u}\right) \right\rfloor. \tag{25}$$

It follows that until herding occurs, $h_t \in \{-k,\ldots,-1,0,1,\ldots\}$ for all such times t, and the update rule for h_t is given by

$$h_{t+1} = \begin{cases} h_t + 1, & \text{if } O_t = H, \\ h_t - 1, & \text{if } O_t = L. \end{cases} \tag{26}$$

Whereas, once $h_t = -(k+1)$, herding begins and h_t stops updating (Property 2). More specifically, equation (26) shows that, until herding occurs, $\{h_t\}$ is a random walk (r.w.) that moves to the right by 1 w.p. $\mathbb{P}(O_t = H|z)$ or to the left by 1 w.p. $\mathbb{P}(O_t = L|z)$, with the walk starting from State 0. The random walk is depicted in Figure 2 where $p_z \triangleq \mathbb{P}(O_t = H|z)$ denotes the probability of a H being observed given the true value of the mean-field, z. Depending on the mean-field's true value, $p_z = u$ for $z = (1-p_1,p_1)$, whereas $p_z = 1-u$ for $z = (1-p_0,p_0)$. Note that for $z = (1-p_1,p_1)$, this random walk will have a drift to the right and so there will be a non-zero probability that it never reaches the absorbing (herding) state.

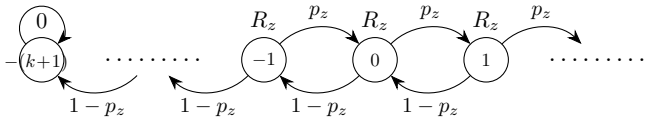

Fig. 2: Transition diagram with state rewards for random walk $\{h_t\}$, under strategy $[\tilde{q},\tilde{r}]^H$ and mean-field z. Absorption in State: $-(k+1)$, with k defined in (25)*.*

A. Aggregate Reward for each player

We now look at the total reward accrued by each player-type x^i under strategy $[\tilde{q},\tilde{r}]^H$ and a prior belief d_1, which we denote by $V^H(d_1,x^i)$ and is defined as per (20). Homogeneity within players of the same type implies that they would receive equal total rewards. To obtain this, we first assign a per-time reward to each state in Figure 2, which could be occupied by r.w. $\{h_t\}$. Note that for all transient states, i.e., $\{-k,\ldots,0,\ldots\}$, we have $B_t \geq 0.5$, which as per (21) means $[\tilde{q},\tilde{r}]^H = [0,1]$ is being played. Then as per (15), $y_t(1) = p_1$ if the mean-field of players' types z has the true value $z = (1-p_1,p_1)$, whereas $y_t(1) = p_0$ if $z = (1-p_0,p_0)$. By applying these values of $y_t(1)$ in (14), it follows that every transient state provides a reward R_z for every player (of both types), where

$$R_z = \begin{cases} (2p_1 - 1), & \text{if } z = (1-p_1,p_1), \\ (2p_0 - 1), & \text{if } z = (1-p_0,p_0). \end{cases} \tag{27}$$

However, for the sole absorption state $-(k+1)$, the MPE strategy changes to $[\tilde{q},\tilde{r}]^H = [1/2,1/2]$. This gives $y_t(1) = 1/2$ and a 0 reward for both player-types, for any z. The reward R_z for the transient states and reward 0 for the absorption state are indicated above their respective state nodes in Figure 2. Note that as the rewards are identical for both player types, we have $V^H(d_1,1) = V^H(d_1,-1)$, and henceforth, we refer to this common value by $V^H(d_1,x^i)$.

Now, to obtain the aggregate reward $V^H(d_1,x^i)$, we first evaluate its conditional value $[V^H(d_1,x^i)|z]$ under the true mean-field z and then average this value over the prior belief d_1 on z.

$$\begin{aligned} V^H(d_1,x^i) = {} & [V^H(d_1,x^i)|z = (1-p_1,p_1)]d_1 \\ & + [V^H(d_1,x^i)|z = (1-p_0,p_0)](1-d_1). \end{aligned} \tag{28}$$

To obtain $[V^H(d_1,x^i)|z]$, consider the random walk $\{h_t\}$ with state rewards, shown in Figure 2. Here, starting from state 0, at each time t, reward R_z is accrued with a discount factor of δ^{t-1}. Therefore, with r.v. T defined as the discounted time to absorption (herding) into state $-(k+1)$, we have:

$$[V^H(d_1,x^i)|z] = R_z \mathbb{E}^z[T], \tag{29}$$

where $\mathbb{E}^z[T]$ is the expected value of T under z. Note that the r.v. T satisfies: $1 \leq T < (1-\delta)^{-1}$ as it would take at least 1 time-step and at most an infinite number of time-steps for players to herd. Applying (29) in (28) then yields

$$\begin{aligned} V^H(d_1,x^i) = {} & (2p_1 - 1)\mathbb{E}^{(1-p_1,p_1)}[T]d_1 \\ & + (2p_0 - 1)\mathbb{E}^{(1-p_0,p_0)}[T](1-d_1). \end{aligned} \tag{30}$$

Property 4: Equation (30) explicitly relates the expected discounted time for herding to the players' aggregate reward, given a prior belief d_1 on the mean-field.

The expected value of T given true mean-field z can be found by solving the following system of linear equations:

$$\begin{cases} g_s = 0, & \text{for } s = -(k+1), \\ g_s = 1 + \delta[p_z\, g_{s+1} + (1-p_z) g_{s-1}], & \text{for } s \geq -k, \end{cases} \tag{31}$$

with variables $\{g_s\}_{s=-(k+1)}^{\infty}$. Here, g_s refers to the expected discounted time to herd, starting from State s, given z. We omit the dependence of g_s on z for notational convenience and emphasize that the system in (31) depends on z through the transition probability p_z. Now, as the r.w. $\{h_t\}$ starts from state 0, we have

$$\mathbb{E}^z[T] = g_0, \quad z \in \mathcal{Z}.$$

For the sake of numerical computations, we restrict (31) to a finite number of equations (and variables) by converting the transient state M, for some integer $M \gg 0$, into an absorption state. We then assign the value $1/(1-\delta)$ to variable g_M, which is the discounted time elapsed if absorption (herding) to state $-(k+1)$ never happens. This is indeed the case starting from state M, in the limit as $M \to \infty$.

B. Comparison with an "Oracle" game, $\mathbb{G}_o$

Now, to understand the effects that an uncertainty on the true value of the mean-field has on the players' aggregate reward, we compare $\mathbb{G}_E$ with a reference game $\mathbb{G}_o$, as follows. Game $\mathbb{G}_o$ shares the same prior belief d_1 as in $\mathbb{G}_E$, except that here, an *oracle* informs players of the true value of the realized mean-field z, at time $t = 1$. Once this is revealed, there is no uncertainty in the repeated game and so players can simply adopt a fixed strategy for all time. Note that B_t in (21) gets replaced by either p_0 or p_1 as per the realization of z. Then, as $p_0 < 1/2 < p_1$ (Assumption 1), players play the MPE strategy: $[\tilde{q}, \tilde{r}]^H = [1/2, 1/2]$ if $z = (1-p_0, p_0)$ which is when they receive 0 reward per time. Otherwise, players play $[\tilde{q}, \tilde{r}]^H = [0, 1]$ if $z = (1-p_1, p_1)$ and receive reward $(2p_1 - 1)$ per time. Then the aggregate reward in this scenario, denoted by $V_o(d_1, x^i)$, for both $x^i \in \{-1, 1\}$ is given by:

$$V_o(d_1, x^i) = (2p_1 - 1)(1-\delta)^{-1} d_1 + 0 \cdot (1 - d_1). \tag{32}$$

By comparing eq. (32) with eq. (30), it follows that when $z = (1-p_1, p_1)$, players would benefit from herding as late as possible (ideally never), since each time-step until absorption yields a positive reward of $(2p_1 - 1)$, as $p_1 > 1/2$. Whereas, when $z = (1-p_0, p_0)$, players would benefit from herding as soon as possible (ideally in an instant), since each time-step until absorption yields a negative reward of $(2p_0 - 1)$, as $p_0 < 1/2$.

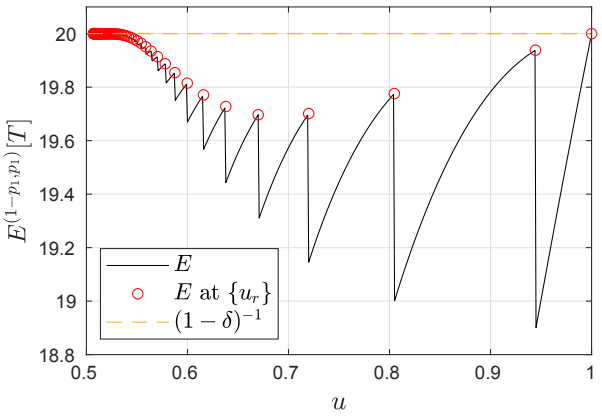

Fig. 3: Expected discounted time given $z = (1-p_1, p_1)$ as a function of signal quality u for $p_0 = 0.4$, $p_1 = 0.8$, $d_1 = 0.85$, $\delta = 0.95$. Values at thresholds $\{u_r\}$ are marked by $\circ$.

C. Effects of varying the signal quality

In this section, we consider the effects of varying the quality $u \in (0.5, 1]$ of the signals $\{O_t\}$, which the players observe through the BSC in Figure 1. First, we define a decreasing sequence of signal quality thresholds: $\{u_r\}_{r=0}^{\infty}$, which is characterized in the following lemma.

Lemma 1: For $r = 0, 1, \ldots$ define the decreasing sequence of thresholds $\{u_r\}_{r=0}^{\infty}$, where the r^{th} threshold u_r is given as:

$$u_r = \frac{\alpha^{\frac{1}{r}}}{1 + \alpha^{\frac{1}{r}}}, \quad \text{where } \alpha := \frac{d_1(1-d_{\text{th}})}{d_{\text{th}}(1-d_1)}. \tag{33}$$

Define $\mathcal{I}_r := (u_{r+1}, u_r]$ as the r^{th} u–interval. Then for all $u \in \mathcal{I}_r$, the index k in Figure 2 equals r. Thus, at least $r+1$ consecutive L's are necessary for a herd to begin.

Lemma 1 implies that when u marginally exceeds threshold u_r, the absorption state index in Figure 2 abruptly increases from $-(r+1)$ to $-r$. The proof of Lemma 1 follows by noting that integer k in (25) increases in steps from $k = 0$ at $u = 1$ to $k = \infty$ at $u = 0.5$. Then, u_r is the value of u at which $k = \lfloor r \rfloor = r$. Solving this equality for u yields u_r in (33).

We now observe the effects of varying u. Figures 3 and 4 show the plots of $\mathbb{E}^z[T]$ with respect to u for $z = (1-p_1, p_1)$ and $z = (1-p_0, p_0)$, respectively. The game parameters are: $p_0 = 0.4$, $p_1 = 0.8$, $d_1 = 0.85$, $\delta = 0.95$. Observe that the abrupt drops in $\mathbb{E}^z[T]$ (marked by $\circ$) in both figures occur exactly at the threshold values $\{u_r\}_{r=1}^{\infty}$, and are an effect of the abrupt change in the number of consecutive L's required for herding at these values (Lemma 1). Further, the constant level of $(1-\delta)^{-1}$ in these figures indicates the upper bound on T and $\mathbb{E}^z[T]$, i.e., the discounted time elapsed when herding never occurs. Interestingly, $\mathbb{E}^z[T]$ for both z's tends to $(1-\delta)^{-1}$ as $u \to 0.5$. This is to be expected as in this case, the herding state $-(k+1)$ tends to $-\infty$ and the drift of the r.w. tends to 0. Varying u also effects the dynamics of $\{h_t\}$ (see Fig. 2). Under $z = (1-p_1, p_1)$, an increase in u has two opposing effects on the r.w., namely, an increase

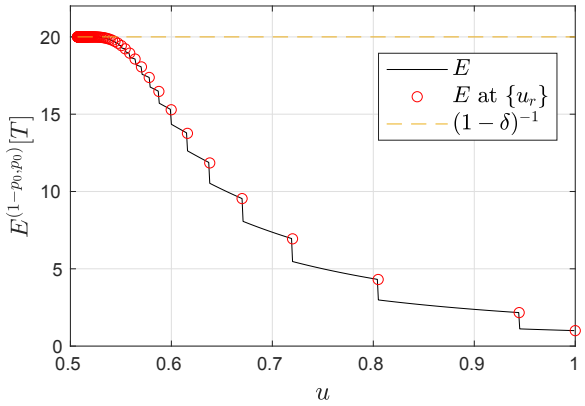

Fig. 4: Expected discounted time given $z = (1 - p_0, p_0)$ as a function of signal quality u for $p_0 = 0.4$, $p_1 = 0.8$, $d_1 = 0.85$, $\delta = 0.95$. Values at thresholds $\{u_r\}$ are marked by $\circ$.

in the drift away from the herding state and a reduced number of conseutive L's required for herding. For this reason, $\mathbb{E}^{(1-p_1,p_)}[T]$ in Fig. 3 increases with u, over every interval $(u_{r+1}, u_r]$, but abruptly drops at thresholds $\{u_r\}$. On the contrary, under $z = (1 - p_0, p_0)$ the corresponding effects do not oppose but align with each other. This causes $\mathbb{E}^{(1-p_0,p_0)}[T]$ in Fig. 4 to monotonically decay to 1 despite the discontinuities, as u increases.

Lastly, Fig. 5 plots the aggregate reward $V^H(d_1, x^i)$ with respect to u and compares it with the baseline $V_o(d_1, x^i)$, which is the aggregate reward under game $\mathbb{G}_o$ (with the oracle). We see that, within each interval $(u_{r+1}, u_r]$, the aggregate reward is increasing in the signal quality, u. But, counter to expectation, a slight increase in u beyond u_r causes an abrupt and significant decrease in the aggregate reward. Also observe that as $u \to 0.5$, $V^H(d_1, x^i)$ tends to a limiting value, indicated as a constant level, which can be computed to be $(2B_1 - 1)(1 - \delta)^{-1}$. This limiting value of $V^H(d_1, x^i)$ is obtained by putting $\lim_{u\to 0.5} \mathbb{E}^z[T] = (1-\delta)^{-1}$, for both z's, in the corresponding limit of eq. (30).

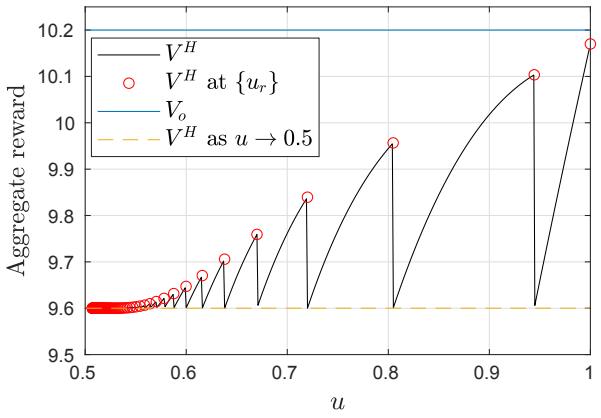

Fig. 5: Aggregate rewards for the indicated cases as a function of signal quality u for $p_0 = 0.4$, $p_1 = 0.8$, $d_1 = 0.85$, $\delta = 0.95$. Values at thresholds $\{u_r\}$ are marked by $\circ$.

V. Conclusions and Future Work

We considered a mean-field game model in which Bayesian agents, each having a private type, update beliefs about the underlying mean-field of population types from imperfect observations of the mean-field action profile. We gave a sequential decomposition of this type of game and used this to study a particular class of games with a mix of coordinating and anti-coordinating players. We showed that for a particular choice of equilibrium strategies, a type of *herding* behavior can emerge in which players no longer learn from their observations. We found that each player's expected total discounted reward over the infinite horizon relates to the expected discounted time taken for herding to occur. Lastly, we illustrated how "better" observations may in several cases, lead to worse expected outcomes for the players, similar to what has been observed in sequential Bayesian observational learning models.

Possible future directions include considering other games that fit within this framework and other types of observation models.

References

[1] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.

[2] A. V. Banerjee, "A simple model of herd behavior," *The quarterly journal of economics*, vol. 107, no. 3, pp. 797–817, 1992.

[3] I. Welch, "Sequential sales, learning, and cascades," *The Journal of finance*, vol. 47, no. 2, pp. 695–732, 1992.

[4] L. Smith and P. Sørensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000.

[5] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[6] I. H. Lee, "On the convergence of informational cascades," *Journal of Economic theory*, vol. 61, no. 2, pp. 395–411, 1993.

[7] Y. Wang and P. M. Djurić, "Social learning with bayesian agents and random decision making," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3241–3250, 2015.

[8] T. Le, V. Subramanian, and R. Berry, "Quantifying the utility of noisy reviews in stopping information cascades," in *IEEE CDC*, 2016.

[9] T. N. Le, V. G. Subramanian, and R. A. Berry, "Bayesian learning with random arrivals," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 926–930.

[10] P. Poojary and R. Berry, "Observational learning with negative externalities," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1495–1496.

[11] I. Bistritz and A. Anastasopoulos, "Characterizing non-myopic information cascades in bayesian learning," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 2716–2721.

[12] G. Schoenebeck, S. Su, and V. G. Subramanian, "Social learning with questions," *CoRR*, vol. abs/1811.00226, 2018. [Online]. Available: http://arxiv.org/abs/1811.00226

[13] T. N. Le, V. G. Subramanian, and R. A. Berry, "Information cascades with noise," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 239–251, 2017.

[14] P. Poojary and R. Berry, "Observational learning with fake agents," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1373–1378.

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

[16] D. Fudenberg, D. K. Levine, and E. Maskin, "The folk theorem with imperfect public information." *Econometrica,*, vol. 62, 1994.

[17] D. Vasal, "Sequential decomposition of mean-field games," in *2020 American Control Conference (ACC),*, 2020, pp. 5388–5393.

[18] Y. Ouyang, H. Tavafoghi, and D. Teneketzis, "Dynamic games with asymmetric information: Common information based perfect bayesian equilibria and sequential decomposition,," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 222–237, Jan. 2017.

[19] D. Tang, H. Tavafoghi, V. Subramanian, A. Nayyar, and D. Teneketzis, "Dynamic games among teams with delayed intra-team information sharing," *Dynamic Games and Applications*, vol. 13, no. 1, pp. 353–411, 2023.
[20] D. Vasal, A. Sinha, and A. Anastasopoulos, "A systematic process for evaluating structured perfect bayesian equilibria in dynamic games with asymmetric information," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 81–96, 2018.
[21] L. Arditti, G. Como, F. Fagnani, and M. Vanelli, "Equilibria and learning dynamics in mixed network coordination/anti-coordination games," in *60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 4982–4987.
[22] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. observable actions," *Journal of Economic Theory*, vol. 100, pp. 191–219, 2001.
[23] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1644–1658, 2013.
[24] P. Poojary and R. Berry, "Supplementary material with detailed proofs," 2023. [Online]. Available: https://drive.google.com/file/d/1-kZM3PR--i6hdlbwzwKsyUACJ6l2-dve/view?usp=sharing

APPENDIX

We provide the following lemma which is required for proving Theorem 1 and defer the detailed proof to [24].

Lemma 2: For any $t \in [T]$, $i \in [N]$, $h_t^i \in \mathcal{H}_t^i$ and σ^i,

$$V(\pi_t, x^i) \geq W_t^{\sigma^i,T}(h_t^i), \tag{34}$$

where the reward-to-go function $W_t^{\sigma^i,T}(\cdot)$ is defined later in (35b) and $\mathcal{H}_t^i$ denotes the space of history $(o_{1:t-1}, x^i)$.

Proof of Theorem 1:

Proof: The proof comprises of two parts. First, we show that the function V obtained in (11b) is at least as big as any reward-to-go function. Second, we show that V is in fact the reward-to-go under the MPE strategy $\tilde{\sigma}$ defined in (13). Note that $h_t^i := (o_{1:t-1}, x^i)$.

Part 1: For any $i \in [N]$ and strategy σ^i, define the following reward-to-go functions at time t:

$$W_t^{\sigma^i}(h_t^i) = \mathbb{E}^{\sigma^i,\tilde{\sigma}^{-i}}\left\{\sum_{n=t}^{\infty} \delta^{n-t} R(x^i, A_n^i, Y_n) \mid h_t^i\right\} \tag{35a}$$

$$W_t^{\sigma^i,T}(h_t^i) = \mathbb{E}^{\sigma^i,\tilde{\sigma}^{-i}}\Big\{\sum_{n=t}^{T} \delta^{n-t} R(x^i, A_n^i, Y_n) + \delta^{T+1-t} V(\Pi_{T+1}, x^i) \mid h_t^i\Big\} \tag{35b}$$

Since $\mathcal{X}, \mathcal{A}$ are finite sets, the reward R is absolutely bounded, the reward-to-go $W_t^{\sigma^i}(h_t^i)$ is finite $\forall\ i, t, \sigma^i, h_t^i$.

For any $i \in [N]$, $h_t^i \in \mathcal{H}_t^i$,

$$V\big(\pi_t, x^i\big) - W_t^{\sigma^i}(h_t^i) = \Big[V\big(\pi_t, x^i\big) - W_t^{\sigma^i,T}(h_t^i)\Big] + \Big[W_t^{\sigma^i,T}(h_t^i) - W_t^{\sigma^i}(h_t^i)\Big] \tag{36}$$

By applying Lemma 2 the term in the first bracket in RHS of (36) is non-negative. Using (35), the term in the second bracket is

$$\big(\delta^{T+1-t}\big)\, \mathbb{E}^{\sigma^i,\tilde{\sigma}^{-i}}\Big\{ - \sum_{n=T+1}^{\infty} \delta^{n-(T+1)} R(x^i, A_n^i, Y_n) + V(\Pi_{T+1}, x^i) \mid h_t^i\Big\}.$$

The summation in the expression above is bounded by a convergent geometric series. Also, V is bounded. Hence, the above quantity can be made arbitrarily small by choosing T appropriately large. Now, since the LHS of (36) does not depend on T, we have

$$V\big(\pi_t, x_t^i\big) \geq W_t^{\sigma^i}(h_t^i). \tag{37}$$

Part 2: Since the equilibrium strategy $\tilde{\sigma}$ generated in (13) is such that $\tilde{\sigma}_t^i$ depends on h_t^i only through π_t and x^i, the reward-to-go $W_t^{\tilde{\sigma}^i}$, at strategy $\tilde{\sigma}$, can be written (with abuse of notation) as

$$W_t^{\tilde{\sigma}^i}(h_t^i) = W_t^{\tilde{\sigma}^i}(\pi_t, x^i) \tag{38}$$

$$= \mathbb{E}^{\tilde{\sigma}}\left\{\sum_{n=t}^{\infty} \delta^{n-t} R(x^i, A_n^i, Y_n) \mid \pi_t, x^i\right\}. \tag{39}$$

For any $h_t^i \in \mathcal{H}_t^i$,

$$W_t^{\tilde{\sigma}^i}(\pi_t, x_t^i) = \mathbb{E}^{\tilde{\sigma}}\Big\{R(x^i, A_t^i, Y_t) + \delta W_{t+1}^{\tilde{\sigma}^i}\big(\phi(\pi_t, \theta[\pi_t], O_t)\big), x^i\big) \mid \pi_t, x^i\Big\} \tag{40a}$$

$$V(\pi_t, x^i) = \mathbb{E}^{\tilde{\sigma}}\Big\{R(x^i, A_t^i, Y_t) + \delta V\big(\phi(\pi_t, \theta[\pi_t], O_t)\big), x^i\big) \mid \pi_t, x^i\Big\}. \tag{40b}$$

Repeatedly applying the above for the next $(n-1)$ successive time periods gives:

$$W_t^{\tilde{\sigma}^i}(\pi_t, x^i) = \mathbb{E}^{\tilde{\sigma}}\Big\{\sum_{m=t}^{t+n-1} \delta^{m-t} R(x^i, A_t^i, Y_t) + \delta^n W_{t+n}^{\tilde{\sigma}^i}\big(\Pi_{t+n}, x^i\big) \mid \pi_t, x^i\Big\}, \tag{41a}$$

$$V(\pi_t, x^i) = \mathbb{E}^{\tilde{\sigma}}\Big\{\sum_{m=t}^{t+n-1} \delta^{m-t} R(x^i, A_t^i, Y_t) + \delta^n V\big(\Pi_{t+n}, x^i\big) \mid \pi_t, x^i\Big\}. \tag{41b}$$

Next, we take the difference as follows:

$$W_t^{\tilde{\sigma}^i}(\pi_t, x^i) - V(\pi_t, x^i) = \delta^n \mathbb{E}^{\tilde{\sigma}}\Big\{W_{t+n}^{\tilde{\sigma}^i}\big(\Pi_{t+n}, x^i\big) - V\big(\Pi_{t+n}, x^i\big) \mid \pi_t, x^i\Big\}. \tag{42}$$

Taking absolute value of both sides, then using Jensen's inequality for $f(x) = |x|$ and finally taking supremum over h_t^i reduces (42) to

$$\sup_{h_t^i} \big|W_t^{\tilde{\sigma}^i}(\pi_t, x^i) - V(\pi_t, x^i)\big| \leq \delta^n \sup_{h_t^i} \mathbb{E}^{\tilde{\sigma}}\Big\{\big|W_{t+n}^{\tilde{\sigma}^i}(\Pi_{t+n}, x^i) - V(\Pi_{t+n}, x^i)\big| \mid \pi_t, x^i\Big\}.$$

Now using the fact that W_{t+n} and V are bounded and that we can choose n arbitrarily large, we get $\sup_{h_t^i} |W_t^{\tilde{\sigma}^i}(\pi_t, x^i) - V(\pi_t, x^i)| = 0$, which implies $V(\pi_t, x^i) = W_t^{\tilde{\sigma}^i}(\pi_t, x^i)$. ∎