Decentralized Bayesian learning in dynamic games: A framework to study informational cascades

Deepanshu Vasal and Achilleas Anastasopoulos

Abstract

We study the problem of Bayesian learning in a dynamical system involving strategic agents with asymmetric information. In a series of seminal papers in the literature, this problem has been studied under a simplifying model where selfish players appear sequentially and act once in the game, based on private noisy observations of the system state and public observation of past players’ actions. It is shown that there exist information cascades where users discard their private information and mimic the action of their predecessor. In this paper, we provide a framework for studying Bayesian learning dynamics in a more general setting than the one described above. In particular, our model incorporates cases where players participate for the whole duration of the game, and cases where an endogenous process selects which subset of players will act at each time instance. The proposed framework hinges on a sequential decomposition methodology for finding perfect Bayesian equilibria (PBE) of a general class of dynamic games with asymmetric information, where user-specific states evolve as conditionally independent Markov processes and users make independent noisy observations of their states. Using this methodology, we study a specific dynamic learning model where players make decisions about investing in the team, based on their estimates of everyone’s types. We characterize a set of informational cascades for this problem where learning stops for the team as a whole.

Deepanshu Vasal is with the Department of Electrical and Computer Engineering, and Wisconsin Institute of Discovery, University of Wisconsin, Madison, WI, 53706 USA e-mail: {dvasal} at wisc.edu.

Achilleas Anastasopoulos is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48105 USA e-mail: {anastas} at umich.edu.

The authors would like to thank Vijay Subramanian for his contribution to the paper.

Part of the paper was presented in [1].

This work is supported in part by NSF grant ECCS-1608361.
Index Terms

Bayesian learning, Social networks, Informational cascades, Dynamic games with asymmetric information, Perfect Bayesian equilibrium

I. INTRODUCTION

The problem of how information spreads in a social network is of profound importance in understanding how learning occurs in a group of people or in a society, and it is important even more so today with the ubiquitous presence of internet and social media. Some scenarios of interest include how people vote for a candidate, or make a decision to buy competing products for example apple vs android, or dynamics of mass protests and movements, fads, trends or cult behavior. In all such problems there exist a group of people who have access to certain private information available through say their friends or their own experience, and certain publicly available information such as actions of others say available through mass media. Based on this information they make a decision that affects their reward and further spread of information in the system.

Such problems have been addressed by academicians in various disciplines such as behavioral economics, statistics, engineering and computer science. These problems have the following key features: (a) there are multiple decision makers (henceforth referred to as players) who can be cooperative or strategic, based on if they have same or different objectives, (b) there is asymmetry of information such that players have private and common information, and (c) there is dynamic evolution of the system.

From the mathematical perspective, such problems have two characteristics: (i) decision theoretic: finding optimum or equilibrium or heuristic strategies of players and (ii) statistical/probabilistic/analytic: understanding the evolution and limiting behavior of the system dynamics under those strategies.

One such scenario was studied in the seminal papers [2], [3] where the authors investigated the occurrence of fads in a social network, which was later generalized in [4]. The authors in [2], [3] and [4] study a problem of learning over a social network with pure informational externalities (i.e. when a player’s reward does not directly depend on other players’ actions, however, those actions provide useful information about the state of the system). In this model, there is a product which is either good or bad and there are countably many buyers, i.e. different decision makers, that are chosen exogenously and act exactly once in the process. They make noisy observation about the value of the product and sequentially act strategically to either buy or not buy the product. Their actions are based on their own private observation and the actions of the previous users. It is shown that herding can occur in such a case where the publicly available information becomes powerful enough that user discards its own private information and follows the majority action of its predecessors (characterizing fads in social networks). As a result, the user’s action does not reveal any new information and all future users repeat this behavior. This phenomenon is defined as an informational cascade where learning stops
for the group as a whole. While a good cascade is desirable, there’s a positive probability of a bad cascade that hurts all future users in the community. There is a growing body of literature on alternative learning models that study cascades such as [5], [6]. Inspired by social networks, Acemoğlu et al in [5] consider a model where players only observe a random set of past actions. They show that under sufficient conditions of expanding observations and unbounded beliefs (i.e. log-likelihood ratio of their private beliefs is unbounded), players learn the true state asymptotically and thus cascading does not occur. Le et al [6] study a model where agents observe the past actions erroneously where again they show that cascading does not occur.

There are however more general scenarios, such as cases where players participate in the game more than once, deterministically or randomly, through an exogenous or even an endogenous process. Furthermore, there are practical scenarios where players may be adversarial to each others’ learning (with dynamic zero-sum games in the extreme). Studying such scenarios may reveal more interesting and richer equilibrium behaviors including cascading phenomena, not manifested in the models considered in the current literature. An indispensable tool for studying cascades is a framework for finding equilibria for these dynamical systems involving strategic players with different information sets, which are modeled as dynamic games with asymmetric information. Appropriate equilibrium concepts for such games include perfect Bayesian equilibrium (PBE), sequential equilibrium, trembling hand equilibrium [7], [8]. Each of these notions of equilibrium consists of a strategy and a belief profile of all players where the equilibrium strategies satisfy sequential rationality (i.e. no player as an advantage to unilaterally deviate at equilibrium) given the equilibrium beliefs and the equilibrium beliefs are derived from the equilibrium strategy profile using Bayes’ rule (whenever possible). For the games considered in the current literature including [3]–[6], since every buyer participates only for one time period and thus acts myopically, finding PBE reduces to solving a straightforward, one-shot optimization problem. However, for general dynamic games with asymmetric information, finding PBE is hard, since it requires solving a fixed point equation in the space of strategy and belief profiles across all users and all time periods (for a more elaborate discussion, see [8 Ch. 8]). In general, there is no known sequential decomposition methodology for finding PBE for such games.

Recently, we presented a forward-backward algorithm in [9] for finding structured PBE, also referred to as SPBE, for a general class of dynamic games where a finite number of players have different states associated to them that evolve as conditionally independent Markov processes, and are observed perfectly by the corresponding players. In this paper, in Theorem 1 we present a backward-forward algorithm for finding SPBE for a more general model where players’ do not perfectly observe their states; rather they make independent, noisy observations. The results in [9] vis a vis Theorem 1 in this paper could be interpreted with the analogy of dynamic programming methodology for Markov decision process (MDP) vs that for partially observed Markov decision process (POMDP), where in the former, the state
of the system if perfectly observed by the controller, and in the latter the state is imperfectly observed and thus a new belief state is introduced which then behaves like an MDP. We propose this framework to study Bayesian learning dynamics in dynamic games with asymmetric information. For such games in general, and a given PBE, we then define informational cascades as those histories of the game where players’ actions do not depend on their private information from that point on, and thus the system dynamics are governed only through the common information. Unlike other scenarios in the cascades literature discussed before, the proposed general framework can incorporate, as special cases, scenarios where players participate in the game more than once, deterministically or randomly through an exogenous or endogenous process, and scenarios where players may be adversarial to each others’ learning. Finally, we consider a specific dynamic learning model with pure informational externalities where each player makes a decision to invest (or not invest) in the team, depending on its estimate of the average of all players’ types. Thus learning players types is important aspect of the problem, however players are not adversarial to each others’ learning. Using the methodology presented, in Theorem 3 we characterize a set of informational cascades for this model where learning stops for the team. This occurs in contrast to the fact that players learn their own types asymptotically, however, once in a cascade, their estimates of others’ types freeze. This example provides analysis and intuition on the learning dynamics in games with asymmetric information, and also serves as motivation for exploring a vast landscape of the scenarios that can be studied through the proposed methodology.

A. Relevant literature

There is a growing body of literature on learning in social networks which can broadly be categorized as follows (1) Bayesian learning with myopic or bounded-rational selfish players, and (2) Non-Bayesian learning. These assumptions are argued as being motivated by the people’s behavior, which also simplifies the analysis. We mention some works in each of these categories as follows, where the list by no means is exhaustive of many different models considered in the literature.

1) Bayesian learning with myopic or bounded-rational players: The works in [3]–[6] mentioned before and other related work with the similar model, falls in this category where sequentially acting selfish players participate once in the game and are thus myopic by nature. Some other works where all players act in each period although are assumed to be myopic by design, include [10]–[14]. Mossel and Tamuz consider a repeated round of voting in [11], where in each round, a finite group of myopically selfish players make a binary decision on worthiness of a candidate, based on their Bayesian beliefs which are function of their private information about the candidate and previous actions of the players. They show that a consensus is always reached and probability of a wrong decision decays exponentially in time. Mossel et al in [12] consider general voting models and show that asymptotic learning holds such that as the number of voters goes to infinity the probability of convergence to the
correct outcome goes to one. The same authors in [13] study how topology of a network affect social learning where they identify “egalitarianism” condition where learning occurs in large finite networks. Harel et al. in [14] study the speed of learning with myopically selfish agents acting repeatedly and show that only a fraction of players private information is transmitted through their actions, where this fraction goes to zero as the number of players go to infinity, demonstrating groupthink behavior. Gale and Kariv in [15] consider a similar model as [10] (discussed below) with players on a connected social network where agents observe their neighbors actions. They assume continuum of players such that a player does not influence future play of the game and thus acts myopically. They show that agents converge to an action in finite time, although it may not be an optimal action. Thus there is aggregation of information but not necessarily efficiently.

2) Non-Bayesian learning: There are works on non-Bayesian learning models where players’ don’t update their beliefs in a Bayesian sense. Nedić et al provide a survey of such models in [16]. Some early work in this category includes the work by DeGroot in [17] where the $n$ players have different subjective beliefs about state of the world and in each time-period, they update their beliefs by taking linear average of everyone’s belief. The author finds sufficient conditions (based on recurrent class of a Markov chain) for all players to converge to same beliefs (i.e. the considered Markov chain has a steady-state distribution). Jadabaie et al consider a more general non-Bayesian model in [18], [19] where players have imperfect recall and they consider other players’ beliefs as sufficient statistics. Ellison and Fudenberg in [20], [21] asymptotic learning of true state using rule-of-thumb policies. Bala and Goyal in [10] consider a model of myopically selfish and non-Bayesian players on a connected social network where a player can only observe its neighbor’s actions and observations. They show that in this model players’ beliefs converge almost surely and all players receive the same payoff in the long run.

As mentioned before, in this paper we consider fully rational players in a truly dynamic set-up and is thus in contrast to the models discussed before. While there is some justification in the argument that if computing equilibria for such games is so intractable in theory, it is more likely that people are bound to act bounded-rationally or myopic rationally, and thus these models might be closer to the reality. However, we still see much value in studying the fully rational model, as apart from providing more thorough analysis, our framework does not preclude for simpler strategies to constitute an equilibrium, which could be the case in many situations. Also it can be even more applicable to firms who could employ high computational power. Moreover, our framework allows to scale down from completely rational to bounded-rational behavior by appropriately choosing the domain of the strategies.

The paper is structured as follows. In section III we provide a general methodology to find a class of PBEs for such games. In Section IV we formally define informational cascades and specialize our methodology to study a specific Bayesian learning game, for which we characterize its informational
cascades. We conclude in Section VII.

II. NOTATION

We use uppercase letters for random variables and lowercase for their realizations. For any variable, subscripts represent time indices and superscripts represent player identities. We use notation \(-i\) to represent all players other than player \(i\) i.e. \(-i = \{1, 2, \ldots, i-1, i+1, \ldots, N\}\). We use notation \(a_{t:t'}\) to represent vector \((a_t, a_{t+1}, \ldots, a_{t'})\) when \(t' \geq t\) or an empty vector if \(t' < t\). 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 remove superscripts or subscripts if we want to represent the whole vector, for example \(a_t\) represents \((a_t^1, \ldots, a_t^N)\). In a similar vein, for any collection of finite sets \((\mathcal{X}^i)_{i \in \mathcal{N}}\), we denote \(\times_{i = 1}^{N} \mathcal{X}^i\) by \(\mathcal{X}\). We denote the indicator function of any set \(A\) by \(I_A(\cdot)\). For any finite set \(\mathcal{S}\), \(\Delta(\mathcal{S})\) represents space of probability measures on \(\mathcal{S}\) and \(|\mathcal{S}|\) represents its cardinality. We denote by \(P^g\) (or \(E^g\)) the probability measure generated by (or expectation with respect to) strategy profile \(g\).

We denote the set of real numbers by \(\mathbb{R}\). For a probabilistic strategy profile of players \((\beta^i_{t})_{i \in \mathcal{N}}\) where probability of action \(a^i_t\) conditioned on \((a_{1:t-1}, x^i_{1:t})\) is given by \(\beta^i_t(a^i_t|a_{1:t-1}, x^i_{1:t})\), we use the short hand notation \(\beta^i_{t-1}(a^i_{1:t-1}, x^i_{1:t})\) to represent \(\prod_{j \neq i} \beta^j_t(a^j_t|a_{1:t-1}, x^j_{1:t})\). All equalities and inequalities involving random variables are to be interpreted in \(a.s\) sense.

III. GENERAL MODEL

A. Model

We consider a discrete-time dynamical system with \(N\) strategic players in the set \(\mathcal{N} := \{1, 2, \ldots, N\}\), over a finite time horizon \(T := \{1, 2, \ldots T\}\) and with perfect recall. The system state is \(x_t := (x^1_t, x^2_t, \ldots, x^N_t)\), where \(x^i_t \in \mathcal{X}^i\) is the state of player \(i\) at time \(t\). Players’ states evolve as conditionally independent, controlled Markov processes such that

\[
P(x_t|x_{1:t-1}, a_{1:t-1}) = P(x_t|x_{t-1}, a_{t-1})
\]

\[
= \prod_{i=1}^{N} Q^i_{x_t}(x^i_t|x_{t-1}, a_{t-1}),
\]

where \(a_t = (a^1_t, \ldots, a^N_t)\) and \(a^i_t\) is the action taken by player \(i\) at time \(t\). Player \(i\) does not observe its state perfectly, rather it makes a private observation \(w^i_t \in \mathcal{W}^i\) at time \(t\), where all observations are conditionally independent across time and across players given \(x_t\) and \(a_{t-1}\), in the following way, \(\forall t \in 1, \ldots, T\),

\[
P(w_{1:t}|x_{1:t}, a_{1:t-1}) = \prod_{n=1}^{t} \prod_{i=1}^{N} Q_{w,x}(w^i_t|x^i_t, a_{n-1}).
\]

Player \(i\) takes action \(a^i_t \in \mathcal{A}^i\) at time \(t\) upon observing \(a_{1:t-1}\), which is common information among players, and \(w^i_{1:t}\), which is player \(i\)’s private information. The sets \(\mathcal{A}^i, \mathcal{X}^i, \mathcal{W}^i\) are assumed to be finite. Let \(g^i = (g^i_t)\) be a probabilistic strategy of player \(i\) where \(g^i_t : (\times_{j=1}^{N} \mathcal{A}^j)^{t-1} \times (\mathcal{W}^i)^{t} \rightarrow \Delta(\mathcal{A}^i)\) such
that player $i$ plays action $a_i^t$ according to $A_i^t \sim g_i^t(\cdot|a_{1:t-1},w_{1:t}^i)$. Let $g := (g^i)_{i \in \mathcal{N}}$ be a strategy profile of all players. At the end of interval $t$, player $i$ gets an instantaneous reward $R^i(x_t, a_t)$. The objective of player $i$ is to maximize its total expected reward

$$J^{i,g} := \mathbb{E}_g \left[ \sum_{t=1}^{\mathcal{T}} R^i(X_t, A_t) \right].$$

(3)

With all players being strategic, this problem is modeled as a dynamic game $\mathcal{D}$ with asymmetric information and with simultaneous moves. Although this model considers all $N$ players acting in all periods of the game, it can accommodate cases where at each time $t$, players are chosen through an endogenously defined (controlled) Markov process. This can be done by introducing a nature player 0, who perfectly observes its state process $(X^0_t)_t$, has reward function zero, and plays actions $a_0^t = w_0^t = x_0^t$. Equivalently, all players publicly observe a controlled Markov process $(X^0_{t-1})_t$, and a player selection process could be defined through this process. For instance, let $X^0 = A^0 = \mathcal{N}$, $\forall i$, $R^i_t(x_t, a_t) = 0$ if $i \neq a_0^t$, and $Q_x(x_{t+1}^t|x_t^t, a_t) = Q_x(x_{t+1}^t|x_t^t, a_t^0)$. Here, in each period only one player acts in the game which is selected through an internal, controlled Markov process.

**B. Solution concept: PBE**

In this section, we introduce PBE as an appropriate equilibrium concept for the game considered. Any history of this game at which players take action is of the form $h_t = (a_{1:t-1}, x_{1:t}, w_{1:t})$. Let $\mathcal{H}_t$ be the set of such histories at time $t$. At any time $t$ player $i$ observes $h_i^t = (a_{1:t-1}, w_{1:t}^i)$ and all players together observe $h_i^t = a_{1:t-1}$ as common history. Let $\mathcal{H}_t^i$ be the set of observed histories of player $i$ at time $t$ and $\mathcal{H}^t$ be the set of common histories at time $t$. An appropriate concept of equilibrium for such games is the PBE [8] which consists of a pair $(\beta^*, \mu^*)$ of strategy profile $\beta^* = (\beta^*_t)_{t \in \mathcal{T}, i \in \mathcal{N}}$ where $\beta^*_t : \mathcal{H}_t \rightarrow \Delta(A^i)$ and a belief profile $\mu^* = (\mu^*_t)_t \in \mathcal{T}, i \in \mathcal{N}$ where $\mu^*_t : \mathcal{H}_t^i \rightarrow \Delta(\mathcal{H}_t)$ that satisfy sequential rationality so that $\forall i \in \mathcal{N}, t \in \mathcal{T}, h^i_t \in \mathcal{H}^i_t, \beta^*$

$$\mathbb{E}^{(\beta^*, \beta^*-i, \mu^*)\left\{ \sum_{n=t}^{\mathcal{T}} R^i(X_n, A_n) \middle| h^i_t \right\} \geq \mathbb{E}^{(\beta^*, \beta^*-i, \mu^*)\left\{ \sum_{n=t}^{\mathcal{T}} R^i(X_n, A_n) \middle| h^i_t \right\}},$$

(4)

and the beliefs satisfy consistency conditions as described in [8] p. 331).

In general, $\mu^*_t$ is defined as the belief of player $i$ at time $t$ on the history $h_t = (a_{1:t-1}, x_{1:t}, w_{1:t})$, conditioned on its observed history $h^i_t = (a_{1:t-1}, w_{1:t}^i)$. In our model, due to independence of types and observations, player $i$’s private observations $w_{1:t}^i$ do not provide any information about $(x_{1:t}, w_{1:t}^i)$, as will be shown later. Then, player $i$’s relevant uncertainty, as required to compute its expected reward-to-go, can be sufficiently represented by its private belief $\xi^i_t$ and public belief on all players’ private beliefs, $\xi_t$ (both of which are defined later). Hence, instead of $\mu^*_t[h_t][h^i_t]$, every agent uses a private belief, $\xi^i_t$, and a public belief, $\mu^*_t[a_{1:t-1}](\xi_t)$, derived from the common history $h^i_t = a_{1:t-1}$, where $\mu^*_t[a_{1:t-1}](\xi_t)$ itself factorizes into a product of marginals $\prod_{j \in \mathcal{N}} \mu^*_{t,j}[a_{1:t-1}](\xi^j_t)$.

June 27, 2017

DRAFT
C. PBE of the game $\mathcal{D}$

In this section, we provide a methodology to find PBE of the game $\mathcal{D}$ that consists of strategies whose domain is time-invariant (while there may exist other equilibria that can not be found using this methodology). Specifically, we seek equilibrium strategies that are structured in the sense that they depend on players’ common and private information through belief states. In order to achieve this, at any time $t$, we summarize player $i$’s common information, $a_{1:t-1}$, in belief $\pi_t$, and its private information, $w_{1:t}^i$ (together with $a_{1:t-1}$), in the belief $\xi_t^i$, where $\xi_t^i$ and $\pi_t$ are defined as follows. For a strategy profile $g$, let $\xi_t^i(x_t^i) := P^g(x_t^i|a_{1:t-1}, w_{1:t}^i)$ be the belief of player $i$ on its current state conditioned on its information, where $\xi_t^i \in \Delta(X^i)$. Also, we define $\pi_t^i(\xi_t^i) := P^g(\xi_t^i|a_{1:t-1})$ as common belief on $\xi_t^i$ based on the common information of the players, $a_{1:t-1}$, where $\pi_t^i \in \Delta(\Delta(X^i))$. As it will be shown later, due to the independence of states and their evolution as independent controlled Markov processes, for any strategy profile of the players, joint beliefs on states can be factorized as product of their marginals i.e. $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(\xi_t^i)$. To accentuate this independence structure, we define $\bar{\pi}_t \in \times_{i \in N} \Delta(X^i)$ as vector of marginal beliefs where $\bar{\pi}_t := (\pi_t^i)_{i \in N}$.

Inspired by the common agent approach in decentralized team problems [22], we now generate players’ structured strategies as follows: player $i$ at time $t$ observes a common belief vector $\bar{\pi}_t$ and takes action $\gamma_t^i$, where $\gamma_t^i : \Delta(X^i) \rightarrow \Delta(A^i)$ is a partial (stochastic) function from its private belief $\xi_t^i$ to $a_t^i$ of the form $\gamma_t^i(a_t^i|\xi_t^i)$. These actions are generated through some policy $\theta^i = (\theta_t^i)_{t \in T}$, $\theta_t^i : \times_{i \in N} \Delta(\Delta(X^i)) \rightarrow \{\Delta(X^i) \rightarrow \Delta(A^i)\}$, that operates on the common belief vector $\bar{\pi}_t$, so that $\gamma_t^i(\cdot|\xi_t^i) = \theta_t^i(\cdot|\bar{\pi}_t)$. Then, the generated policy of the form $A_t^i \sim \theta_t^i(\cdot|\bar{\pi}_t)$ is also a policy of the form $A_t^i \sim g_t^i(\cdot|a_{1:t-1}, w_{1:t}^i)$ for an appropriately defined $g$. Although this is not relevant to our proofs, it can be shown (similar to [9] sec. III) that these structured policies form a sufficiently large, rich set of policies, which provides a good motivation for restricting attention to such equilibria. Specifically, it can be shown that policies $g$ are outcome equivalent to policies of state $\theta$, i.e., any expected total reward profile of the players that can be generated through a general policy profile $g$ can also be generated through some policy profile $\theta$. In the following lemma, we present the update functions of the private belief $\xi_t^i$ and the public belief $\pi_t^i$.

**Lemma 1**: There exist update functions $G^i$, independent of players’ strategies $g$, such that

$$\xi_{t+1}^i = G^i(\xi_t^i, w_{t+1}^i, a_t)$$  \hfill (5)

and update functions $F^i$, independent of $\theta$, such that

$$\pi_{t+1}^i = F^i(\pi_t^i, \gamma_t^i, a_t).$$  \hfill (6)

Thus $\bar{\pi}_{t+1} = F(\bar{\pi}_t, \gamma_t, a_t)$ where $F$ is appropriately defined through (6).
Proof: The proofs are straightforward using Bayes’ rule and the fact that players’ state and observation histories, $X_{1:t}^i, W_{1:t}^i$, are conditionally independent across players given the action history $a_{1:t}$. They are provided in Appendix A.

Based on (5), we define an update kernel of $\xi_t^i$ in (35) as $Q^i(\xi_{t+1}^i | \xi_t^i, a_t) := P(\xi_{t+1}^i | \xi_t^i, a_t)$. We now present the backward-forward algorithm to find PBE of the game $\mathcal{D}$, where strategies of the players are of state $\theta$. The algorithm resembles the one presented in [9] by the same authors for perfectly observable states.

1) Backward Recursion: In this section, we define an equilibrium generating function $\theta = (\theta_t^i)_{i \in \mathcal{N}, t \in T}$ and a sequence of functions $(V_t^i)_{i \in \mathcal{N}, t \in \{1,2,...,T+1\}}$, where $V_t^i : \times_{i \in \mathcal{N}} \Delta(\mathcal{X}^i) \times \Delta(\mathcal{X}^i) \rightarrow \mathbb{R}$, in a backward recursive way, as follows.

1. Initialize $\forall \pi_{T+1} \in \times_{i \in \mathcal{N}} \Delta(\mathcal{X}^i), \xi_{T+1}^i \in \Delta(\mathcal{X}^i)$,

\[
V_{T+1}^i(\pi_{T+1}, \xi_{T+1}^i) := 0. \tag{7}
\]

2. For $t = T, T-1, \ldots, 1$, $\forall \pi_t \in \times_{i \in \mathcal{N}} \Delta(\mathcal{X}^i)$, let $\hat{\gamma}_t = \hat{\gamma}_t(\pi_t)$ be generated as follows. Set $\hat{\gamma}_t = \hat{\gamma}_t(\pi_t)$, where $\hat{\gamma}_t$ is the solution, if it exists\footnote{Similar to the existence results shown in [23], it can be shown that in the special case where agent $i$’s instantaneous reward does not depend on its private state $x_t^i$, and for uncontrolled states and observations, the fixed point equation always has a state-independent, myopic solution $\hat{\gamma}_t^i(\cdot)$, since it degenerates to a Bayesian-Nash like best-response equation.}, of the following fixed point equation, of the following fixed point equation,

\[
\hat{\gamma}_t^i(\cdot | \xi_t^i) \in \arg \max_{\hat{\gamma}_t^i(\cdot | \xi_t^i)} \mathbb{E}_{\hat{\gamma}_t^i(\cdot | \xi_t^i)} [\hat{R}^i(X_t, A_t) + V_{t+1}^i(\pi_t, \hat{\gamma}_t, A_t, \xi_{t+1}^i)] | \xi_t^i \right\}, \tag{8}
\]

where expectation in (8) is with respect to random variables $(X_t, A_t, \xi_{t+1}^i)$ through the measure $\xi_t^i \pi_t^{-1}(\xi_t^i) \hat{\gamma}_t^i(a_t | \xi_t^i) \pi_t^{-1}(a_t | \xi_t^i) Q_t^i(\xi_{t+1}^i | \xi_t^i, a_t)$. $\hat{R}$ is defined in Lemma 1 and $Q^i$ is defined in (35). Furthermore, set

\[
V_t^i(\pi_t, \xi_t^i) := \mathbb{E}^{\hat{\gamma}_t^i(\cdot | \xi_t^i)} [\hat{R}^i(X_t, A_t) + V_{t+1}^i(\pi_t, \hat{\gamma}_t, A_t, \xi_{t+1}^i)] | \xi_t^i \right\}. \tag{9}
\]

It should be noted that (8) is a fixed point equation where the maximizer $\hat{\gamma}_t^i$ appears in both, the left-hand-side and the right-hand-side of the equation. However, it is not the outcome of the maximization operation as in a best response equation, similar to that of a Bayesian Nash equilibrium.

2) Forward Recursion: Based on $\theta$ defined above in (7–9), we now construct a set of strategies $\beta^*$ and beliefs $\mu^*$ for the game $\mathcal{D}$ in a forward recursive way, as follows. As before, we will use the notation $\mu^*_{1:t-1} := (\mu^*_t)_{i \in \mathcal{N}}$ and $\mu^*_{t[a_{1:t-1}]}$ can be constructed from $\mu^*_{1:t-1}$ as $\mu^*_t[a_{1:t-1}] := \prod_{i=1}^{N} \mu^*_t[a_{1:t-1}](\xi_t^i)$, where $\mu^*_t[a_{1:t-1}]$ is a belief on $\xi_t^i$.

1. Initialize at time $t = 0$,

\[
\mu^*_{0[a_{1:t-1}]}(\xi_0) := \delta_{Q_{0}^i}(\xi_0^i). \tag{10}
\]
2. For \( t = 1, 2 \ldots T, i \in \mathcal{N}, \forall a_{1:t}, w_{1:t} \)

\[
\beta^{*,i}_{t}(a^i_{1:t-1}, w_{1:t}) := \theta_{t}^{i}[\mu_{t}^{*,i}[a_{1:t-1}]](a_t^i) \quad (11a)
\]

\[
\mu_{t+1}^{*,i}[a_{1:t}] := E(\mu_{t}^{*,i}[a_{1:t-1}], \theta_{t}^{i}[\mu_{t}^{*,i}[a_{1:t-1}], a_t]) \quad (11b)
\]

where \( \bar{F} \) is defined in Lemma 1.

**Theorem 1:** A strategy and belief profile \((\beta^*, \mu^*)\), constructed through backward/forward recursive algorithm is a PBE of the game, i.e. \( \forall i \in \mathcal{N}, t \in T, (a_{1:t-1}, w_{1:t}^i), \beta^i, \)

\[
\mathbb{E}^{\beta^i_{t+1}\beta^i_{t-1}^{-i}, \mu^i_{t-1}} \left\{ \sum_{n=t}^{T} R^n(X_n, A_n) \Big| a_{1:t-1}, w_{1:t}^i \right\} \geq \mathbb{E}^{\beta^i_{t+1}\beta^i_{t-1}^{-i}, \mu^i_{t-1}} \left\{ \sum_{n=t}^{T} R^n(X_n, A_n) \Big| a_{1:t-1}, w_{1:t}^i \right\}. \quad (12)
\]

**Proof:** The proof relies crucially on the specific fixed point construction in (8) and the conditional independence structure of states and observations, and is provided in Appendix B.

**Remark 1:** When players observe their types perfectly, i.e. when \( \forall \beta^i_t = X^i_t \) and \( Q^n_{w_t}(x^i_t, a_{t-1}) = \delta_{x^i_t}(w^i_t), \forall i, w^i_t, x^i_t, a_{t-1}, \) then \( \xi^i_t(\cdot) = \delta_{x^i_t}(\cdot), \forall x^i_t \) and the results in this section reduce to the results in [9], as expected.

**IV. INFORMATIONAL CASCADES**

We now define informational cascades for a dynamic game with asymmetric information, and for a given PBE of that game, as those public histories of the game at which the future actions of the players are predictable\(^2\).

**Definition 1:** For a given strategy and belief profile \((\beta^*, \mu^*)\) that constitute a PBE of the game\(^3\), and for any time \( t \) and a sequence of action profile \( a_{t:T} \), informational cascades can be defined as set of public histories \( h^i_t \) of the game such that at \( h^i_t \) and under \((\beta^*, \mu^*)\), actions \( a_{t:T} \) are played almost surely, irrespective of players’ future private history realizations, i.e., for a PBE \((\beta^*, \mu^*)\) and time \( t \) and actions \( a_{t:T} \), cascades are defined by

\[
C^a_{t:T} := \{ h^i_t \in \mathcal{H}^i_t \mid \forall i, \forall n \geq t, \forall h^i_n, h^i_t \text{ that are consistent with } h^i_n \text{, and occur with non-zero probability, } \beta^{*,i}_{n}(a^i_{n}|h^i_{n}) = 1 \}. \quad (13)
\]

\(^2\)More generally, informational cascades can be thought as those histories of the game at which the system dynamics, from that point on, only depend on the common information. For our model, since the common information consists of past actions, it is equivalent to future actions being predictable.

\(^3\)A stronger notion of informational cascade could be defined for all PBEs of the game.
We also call an informational cascade a constant informational cascade if action profiles in the cascade are constant across time, i.e. for time $t$ and action profile $a$, constant cascades are defined by

$$\mathcal{C}^a_t := \{ h^i_t \in \mathcal{H}^i_t \mid \forall i, \forall n \geq t, \forall h^i_n \text{ that are consistent with } h^i_t, \text{ and occur with non-zero probability, } \beta_n^{a,i}(a^i|h^i_n) = 1 \}. \quad(14)$$

For the general games considered in this paper, which are dynamic games with asymmetric information and independent states, a more useful definition of cascades is the following.

**Definition 2:** For a given equilibrium generating function $\theta$, and for time $t$ and actions $a_{t:T}$, informational cascades are defined by the sets $\{ \tilde{\mathcal{C}}^a_t \}_{t=1}^{T+1}$, which are defined as follows. For $t = T, T-1, \ldots, 1$,

$$\tilde{\mathcal{C}}_{T+1} := \{ \text{All possible common beliefs } \pi_{T+1} \} \quad(15)$$

$$\tilde{\mathcal{C}}^{a_{t:T}}_t := \{ \pi_t \mid \forall i, \forall \xi^i_t \in \text{supp}(\pi^i_t), \theta^i_t[\pi_i](a^i_t | \xi^i_t) = 1 \text{ and } F(\pi_t, \theta_t[\pi_t], a_t) \in \tilde{\mathcal{C}}^{a_{t+1:T}}_{t+1} \}. \quad(16)$$

A constant informational cascade for time $t$ and actions profile $a$ is defined as,

$$\tilde{\mathcal{C}}_{T+1} := \{ \text{All possible common beliefs } \pi_{T+1} \} \quad(17)$$

$$\tilde{\mathcal{C}}^a_t := \{ \pi_t \mid \forall i, \forall \xi^i_t \in \text{supp}(\pi^i_t), \theta^i_t[\pi_i](a^i_t | \xi^i_t) = 1 \text{ and } F(\pi_t, \theta_t[\pi_t], a_t) \in \tilde{\mathcal{C}}^a_{t+1} \}. \quad(18)$$

In the following lemma, we show the connection between the two definitions.

**Lemma 2:** Let $(\beta^*, \mu^*)$ be an SPBE of a dynamic game with asymmetric information and independent states, generated by an equilibrium generating function $\theta$. Then $\forall t, a_{t:T}$,

$$(\mu_t^*)^{-1}(\tilde{\mathcal{C}}^{a_{t:T}}_t) = \mathcal{C}^{a_{t:T}}_t. \quad(19)$$

**Proof:** See Appendix [D].

**Corollary 1:** Let $(\beta^*, \mu^*)$ be an SPBE of a dynamic game with asymmetric information and independent states, generated by an equilibrium generating function $\theta$. Then $\forall t, a$,

$$(\mu_t^*)^{-1}(\tilde{\mathcal{C}}^a_t) = \mathcal{C}^a_t. \quad(20)$$

### A. Specific learning model

We now consider a specific model that captures the learning aspect in a dynamic setting with strategic agents and decentralized information. The model is inspired by the model considered in [3], [4] where now we consider a finite number of players who take action in every epoch and participate during the entire duration of the game. We assume that players’ states are uncontrollable and static.
i.e. $Q^i_w(x_{t+1}^i | x_t^i, a_t) = \delta_{x_t^i} (x_{t+1}^i)$, where $X^i = \{-1, 1\}$ and $P(X^i = -1) = P(X^i = 1) = 1/2$. Since the set of states, $X^i$ is has cardinality 2, the measure $\xi^i_t$ can be sufficiently described by $\xi^i_t(1)$. Henceforth, in this section and in Appendix F with slight abuse of notation, we also denote $\xi^i_t(1)$ by $\xi^i_t$, and reference is clear from context. In each epoch $t$, player $i$ makes independent observation $w_t^i$ about its state where $W^i = \{-1, 1\}$, through an observation kernel of the form $Q^i_w(w_t^i | x_t^i, a_{t-1}^i)$ which does not depend on $a_{t-1}^i$. These observations are made through a binary symmetric channel such that $Q^i_w(-1|1, a^i) = Q^i_w(1|-1, a^i) = p_{a^i}$, where $p_1 < p_0 < 1/2$. This model implies that taking action 1 can improve the quality of a player’s future private belief. Based on its information, it takes action $a_t^i$, where $A^i = \{0, 1\}$, and earns an instantaneous reward given by

$$R^i(x, a_t^i) = a_t^i \left( \lambda x^i + \lambda \frac{\sum_{j \neq i} x^j}{N-1} \right),$$

(21)

where $\lambda \in [0, 1]$, $\bar{\lambda} = 1 - \lambda$. This scenario can thought of as the case when players’ states represent their talent, capabilities or popularity, and a player makes a decision to either invest (action = 1) or not invest (action = 0) in these players, where its instantaneous reward depends on some combination of the capabilities of all the players. We note that the instantaneous reward does not depend on other players’ actions but on their states, and thus learning players’ states is an important aspect of the problem.

In this case, the update functions of $\xi^i_t$ and $\pi^i_t$ in (5), (6) reduce to

$$\xi_{t+1}^i = G^i(\xi_t^i, w_{t+1}^i, a_t^i)$$

(22a)

and (8) in the backward recursion reduces to

$$\tilde{\gamma}_t^i(\xi_t^i) \in \arg\max \sum_{a_t^i} a_t^i \tilde{\gamma}_t^i(a_t^i | \xi_t^i) \left( \lambda(2\xi_t^i - 1) + \bar{\lambda}(2\xi_t^i - 1) \right)$$

$$+ \mathbb{E} \tilde{\gamma}_t^i(\xi_t^i) \tilde{\gamma}_t^i, \pi_t \left\{ V_{t+1}^i \left( F(t, \tilde{\gamma}_t^i, A_t), \Xi_t^i \right) \right\}.$$

(23)

In the following theorem, we show that for the specific learning model considered in this section, the players learn their true state asymptotically.

Theorem 2:

$$\Xi_t \overset{a.s.}{\longrightarrow} \delta_{x_t^i}$$

(24)

Proof: This is a classical Bayesian learning problem and there are many proving techniques to prove the above result (e.g. see [24, pages 314-316]). We provide a proof here for convenience. We prove this for $x_t^i = 1$ and similar arguments follow for $x_t^i = 0$. For $x_t^i = 1$, we show in Lemma 8 in Appendix F that the process $\{\Xi_t^i\}$ is a strict sub-martingale for $p_{a_t^i} < 1/2$ and $\xi^i_t \notin \{0, 1\}$. Since it is also bounded, from Doob’s martingale convergence theorem [25], it converges almost surely to 1 since $\xi^i_0 = Q^i_x(1) \neq 0$. 

DRAFT June 27, 2017
We characterize constant informational cascades for this learning model through a time invariant set \( \hat{\mathcal{C}}^a \) of common beliefs \( \pi \), defined as follows. Let

\[
\hat{\mathcal{C}}^a := \left\{ \pi \mid \forall i, \frac{1}{2} - \frac{\bar{\lambda}}{\lambda}(\hat{\xi}^{-1} - \frac{1}{2}) \geq 1 \text{ if } a^i = 0, \right.
\]

\[
\left. \frac{1}{2} - \frac{\bar{\lambda}}{\lambda}(\hat{\xi}^{-1} - \frac{1}{2}) \leq 0 \text{ if } a^i = 1 \right\},
\]  

(25)

where

\[
\hat{\xi}^{-1} := \frac{1}{N-1} \sum_{j \neq i} E^{\pi_j}[\Xi].
\]  

(26)

In the following theorem we show that the set \( \hat{\mathcal{C}}^a \) defined in (25) characterizes a set of constant informational cascades for this problem. Specifically, we show that \( \hat{\mathcal{C}}^a \subset \check{\mathcal{C}}^a \).

**Theorem 3:** If for some time \( t_0 \) and action profile \( a, \pi_{t_0} \in \hat{\mathcal{C}}^a \), then \( \forall t \geq t_0, \pi_t \in \hat{\mathcal{C}}^a \) and solutions of (23) satisfy \( \check{\gamma}^i_t(a^i|\xi^i_t) = 1 \forall \xi^i_t \in [0, 1] \). Moreover, for \( t_0 \leq t \leq T \), \( V^i_t \) is given by, \( \forall \pi_t \in \hat{\mathcal{C}}^a \),

\[
V^i_t(\pi_t, \xi^i_t) = (T - t + 1)(\lambda(2\xi^i_t - 1) + \bar{\lambda}(2\hat{\xi}^{-1} - 1))a^i.
\]  

(27)

**Proof:** See Appendix [F].

It is useful to note that, in general, for any \( \pi_t \) in a cascading set \( \hat{\mathcal{C}}^a \), \( V^i_t(\pi_t, \cdot) \) represents the reward-to-go for player \( i \) for the open loop control policy \( a_{1:T} \), as was the case in (27) for the specific learning model considered.

**B. Discussion**

We characterize informational cascades by those histories of the game where learning stops for the players as a whole. Conceptually, they could be thought of as absorbing states of the system. It begets questions regarding the dynamics of the process that could lead to those states, for example hitting times of such sets and absorption probabilities. For the simplified problem considered in [3], cascades can be characterized as the fixed points of common belief update function, so that the common belief gets “stuck” once it reaches that state. It was shown that cascades eventually occur with probability 1 for that model. For the learning model considered in this section, common beliefs \( \pi_t \) still evolve in a cascade, although uninformatively, i.e., their evolution is directed by the primitives of the process and not on the new random variables being generated, namely, players’ private observations. Also, if players’ observations are informative, they asymptotically learn their true states, i.e., their private beliefs converge to dirac delta function on their true states. One trivial case when cascades could occur for this model is if the system was born in a cascade, i.e., the initial common belief, based on the prior distributions, is in cascades, \( \pi_1 \in \hat{\mathcal{C}}^a \). In general, a cascade could occur as in the following case. Suppose all players have low states (i.e. \( x^i = -1 \)), but they get atypical observations initially, which lead them into believing that their states are high (\( x^i = 1 \)). This information is conveyed through their
actions, which leads the public belief into a cascade. Now, even though the players eventually learn their true states, yet they remain in a (bad) cascade, each player believing that others have high states on average.

V. CONCLUSION

In this paper we study Bayesian learning dynamics of a specific class of dynamic games with asymmetric information. In the literature, a simplifying model is considered where herding behavior by selfish players is shown in a sequential buyers’ game where a countable number of strategic buyers buy a product exactly once in the game. In this paper, we consider a more general scenario where players could participate in the game throughout the duration of the game. Players’ states evolve as conditionally independent controlled Markov processes and players made noisy observations of their states. We first present a sequential decomposition methodology to find SPBE of the game. We then study a specific learning model and characterize information cascades using the general methodology described before. In general, the methodology presented serves as a framework for studying learning dynamics of decentralized systems with strategic agents. Some important research directions include characterization of cascades for specific classes of models, studying convergent learning behavior in such games including the probability and the rate of “falling” into a cascade, and incentive or mechanism design to avoid bad cascades.

APPENDIX A

Proof: We first prove the following lemma on conditional independence of \( x_{1:t}, w_{1:t} \) given \( a_{1:t-1} \).

Lemma 3: For any policy profile \( g \) and \( \forall t, \)

\[
P_g(x_{1:t}, w_{1:t}|a_{1:t-1}) = \prod_{i=1}^{N} P^{g_i}(x_{1:t}, w_{1:t}|a_{1:t-1})
\]  

(28)

Proof:

\[
P_g(x_{1:t}, w_{1:t}|a_{1:t-1})
\]

\[
= \frac{P^g(x_{1:t}, w_{1:t}, a_{1:t-1})}{\sum_{x_{1:t}, w_{1:t}} P^g(x_{1:t}, w_{1:t}, a_{1:t-1})} \tag{29a}
\]

\[
= \frac{\prod_{i=1}^{N} Q^g_i(x^i_{1:t}) Q^g_w(w^i_{1:t}|x^i_{1:t}) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_n(x^i_{n+1}|a_n, x^i_{1:t}, w^i_n|x^i_{n+1}, a_n)}{\sum_{x_{1:t}, w_{1:t}} \prod_{i=1}^{N} Q^i_w(w^i_{1:t}|x^i_{1:t}) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_n(x^i_{n+1}|a_n, x^i_{1:t}, w^i_n|x^i_{n+1}, a_n)} \tag{29b}
\]

\[
= \frac{\prod_{i=1}^{N} Q^g_i(x^i_{1:t}) Q^g_w(w^i_{1:t}|x^i_{1:t}) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_n(x^i_{n+1}|a_n, x^i_{1:t}, w^i_n|x^i_{n+1}, a_n)}{\prod_{i=1}^{N} \sum_{x_{1:t}, w_{1:t}} Q^i_w(w^i_{1:t}|x^i_{1:t}) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_n(x^i_{n+1}|a_n, x^i_{1:t}, w^i_n|x^i_{n+1}, a_n)} \tag{29c}
\]
and thus
\[ P^g(x_{1:t}, w_{1:t}|a_{1:t-1}) = \prod_{i=1}^N P^g(x^i_{1:t}, w^i_{1:t}|a_{1:t-1}) \] (29d)

Now for any \( g \) we have,
\[ \xi^i_{t+1}(x^i_{t+1}) \triangleq P^g(x^i_{t+1}|a_{1:t}, w^i_{1:t+1}) \]
\[ = \sum_{x^i_{t+1}} P^g(x^i_{t+1}, a_t, x^i_{t+1}, w^i_{1:t+1}|a_{1:t-1}, w^i_{1:t}) \]
\[ = \sum_{x^i_{t+1}} \xi^i_t(x^i_t) P^g(a_t^{-i}|a_{1:t-1}, w^i_{1:t}, x^i_t) Q^i_x(a^i_{t+1}|a_t, x^i_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t) \]
\[ = \sum_{x^i_{t+1}} \xi^i_t(x^i_t) P^g(a_t^{-i}|a_{1:t-1}, w^i_{1:t}, x^i_t) Q^i_x(a^i_{t+1}|a_t, x^i_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t), \] (30c)

where (30c) is true because \( a_t^i \) is a function of \( (a_{1:t-1}, w^i_{1:t}) \) and thus term involving can be cancelled in numerator and denominator. We now consider the quantity \( P^g(a_t^{-i}|a_{1:t-1}, w^i_{1:t}, x^i_t) \)
\[ P^g(a_t^{-i}|a_{1:t-1}, w^i_{1:t}, x^i_t) = \sum_{w^{-i}_{1:t}} P^g(a_t^{-i}, w^{-i}_{1:t}|a_{1:t-1}, w^i_{1:t}, x^i_t) \]
\[ = \sum_{w^{-i}_{1:t}} P^g(w^{-i}_{1:t}|a_{1:t-1}, w^i_{1:t}, x^i_t) \prod_{j \neq i} g^j_t(a^j_t|a_{1:t-1}, w^j_{1:t}) \]
\[ = \sum_{w^{-i}_{1:t}} P^g^{-i}(w^{-i}_{1:t}|a_{1:t-1}) \prod_{j \neq i} g^j_t(a^j_t|a_{1:t-1}, w^j_{1:t}) \]
\[ = P^g^{-i}(a_t^{-i}|a_{1:t-1}) \] (31d)

where (31d) follows from Lemma 3 in Appendix A, since \( w^{-i}_{1:t} \) is conditionally independent of \( (w^i_{1:t}, x^i_t) \) given \( a_{1:t-1} \) and is only a function of \( g^{-i} \). Since this term does not depend on \( x^i_t \), it gets cancelled in the final expression of \( \xi^i_{t+1} \)
\[ \xi^i_{t+1}(x^i_{t+1}) = \sum_{x^i_{t+1}} \xi^i_t(x^i_t) Q^i_x(x^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t) \]
\[ = \xi^i_t(x^i_t) Q^i_x(x^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t) \]
\[ = \xi^i_t(x^i_t) Q^i_x(x^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t), \] (32)

Thus the claim of the lemma follows. Based on this claim, we can conclude that
\[ \xi^i_t(x^i_t) = P^g(x^i_t|a_{1:t-1}, w^i_{1:t}) = P(x^i_t|a_{1:t-1}, w^i_{1:t}), \] (33)

Also, based on the update of \( \xi^i_t \) in (3), we define an update kernel
\[ Q^i(\xi^i_t|\xi^i_t, a_t) := P(\xi^i_{t+1}|\xi^i_t, a_t) \]
\[ = \sum_{x^i_{t+1}, w^i_{t+1}} \xi^i_t(x^i_t) Q^i_x(x^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t) I_{G^i}(\xi^i_t|w^i_{t+1}, a_t)(\xi^i_{t+1}) \] (35)
Lemma 4: There exists an update function \( \bar{F} \) of \( \pi_t \), independent of \( \psi \)

\[
\pi_t^{i+1} = \bar{F}(\pi_t, \gamma_t^i, a_t)
\]  

(36)

Proof:

\[
\pi_t+1(\xi_t+1)
\]

\[
= P^\psi(\xi_t+1|a_{1:t}, \gamma_{1:t+1})
\]  

(37a)

\[
= P^\psi(\xi_t+1|a_{1:t}, \gamma_{1:t})
\]  

(37b)

\[
= \sum_{\xi_t, x_t, \cdots} P^\psi(\xi_t, x_t, a_t, x_{t+1}, w_{t+1}, \xi_{t+1}|a_{1:t-1}, \gamma_{1:t})
\]

(37c)

\[
= \sum_{\xi_t, x_t, \cdots} \prod_{i=1}^N \pi_i^i(\xi_i^i) \xi_i^i(x_i^i) \gamma_i^i(a_i^i|\xi_i^i) Q_i^i(x_i^i+1|x_i^i, a_t) Q_w^i(w_i^i+1|x_i^i, a_t) I_{G_i}(\xi_i^i, w_i^i, a_i)(\xi_i^i+1)
\]

(37d)

\[
\prod_{i=1}^N \sum_{\xi_i^i, x_i^i, \cdots} \pi_i^i(\xi_i^i) \xi_i^i(x_i^i) \gamma_i^i(a_i^i|\xi_i^i) Q_i^i(x_i^i+1|x_i^i, a_t) Q_w^i(w_i^i+1|x_i^i, a_t) I_{G_i}(\xi_i^i, w_i^i, a_i)(\xi_i^i+1)
\]

(37e)

\[
\prod_{i=1}^N \sum_{\xi_i^i, x_i^i, \cdots} \pi_i^i(\xi_i^i) \xi_i^i(x_i^i) \gamma_i^i(a_i^i|\xi_i^i) Q_i^i(x_i^i+1|x_i^i, a_t) Q_w^i(w_i^i+1|x_i^i, a_t) I_{G_i}(\xi_i^i, w_i^i, a_i)(\xi_i^i+1)
\]

(37f)

Thus we have,

\[
\pi_{t+1} = \prod_{i=1}^N F_i(\pi_t, \gamma_t^i, a_t)
\]  

(37g)

APPENDIX B

(Proof of Theorem 1)

Proof: We prove (12) using induction and from results in Lemma 5, 6 and 7 proved in Appendix C. For base case at \( t = T \), \( \forall i \in N, (a_{1:T-1}, w_{1:T}) \in H^i_T, \beta^i \)

\[
\mathbb{E} \beta^i_T, \beta^i_{T-1}, \mu_T[a_{1:T-1}] \left\{ R^i(X_T, A_T) \right\}|a_{1:T-1}, w_{1:T}
\]

(38a)

\[
= V^i_T[a_{1:T-1}, \xi_T]
\]

(38b)

where (38a) follows from Lemma 7 and (38b) follows from Lemma 5 in Appendix C.
Let the induction hypothesis be that for \( t + 1 \), \( \forall i \in \mathcal{N}, (a_{1:t}, w_{1:t+1}^i) \in \mathcal{H}_{t+1}^i, \beta^i \),

\[
\mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t, w_{1:t+1}^i} \geq \mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t, w_{1:t+1}^i}.
\]

Then \( \forall i \in \mathcal{N}, (a_{1:t-1}, w_{1:t}^i) \in \mathcal{H}_{t}^i, \beta^i \), we have

\[
\mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t}^T R^i(X_n, A_n) \right\} a_{1:t-1, w_{1:t}^i} = V_t^i(\mu^*_t[a_{1:t-1}], \xi_t^i) \geq \mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ R^i(X_t, A_t) + V_t^i(\mu^*_t[a_{1:t-1}A_t], \Xi_{t+1}) \right\} a_{1:t-1, w_{1:t}^i}.
\]

\[
\mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i} = \mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i}.
\]

\[
\mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i} = \mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i}.
\]

\[
\mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i} = \mathbb{E}^{\beta^i_{t+1}, \theta^i_{t+1}, \mu^*_t[a_{1:t-1}] | a_{1:t-1}, A_t, w_{1:t}^i} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \right\} a_{1:t-1, A_t, w_{1:t}^i, W_{t+1}^i}.
\]

where (40a) follows from Lemma 7, (40b) follows from Lemma 5, (40c) follows from Lemma 7, (40d) follows from induction hypothesis in (39a) and (40e) follows from Lemma 6. Moreover, construction of \( \theta \) in (8), and consequently definition of \( \beta^* \) in (11a) are pivotal for (40e) to follow from (40d).

We note that \( \mu^* \) satisfies the consistency condition of [8, p. 331] from the fact that (a) for all \( t \) and for every common history \( a_{1:t-1} \), all players use the same belief \( \mu^*_t[a_{1:t-1}] \) on \( x_t \) and (b) the belief \( \mu^*_t \) can be factorized as \( \mu^*_t[a_{1:t-1}] = \prod_{i=1}^N \mu^*_t[a_{1:t-1}] \forall a_{1:t-1} \in \mathcal{H}_t^i \) where \( \mu^*_t \) is updated through Bayes’ rule (F) as in Lemma 1 in Appendix A.
APPENDIX C

Lemma 5: \( \forall t \in T, i \in \mathcal{N}, (a_{1:t-1}, w_{1:t}) \in \mathcal{H}_i, \beta_i^t \)

\[
V_i^t(\mu_i^t[a_{1:t-1}], \xi_i^t) \geq \mathbb{E}^{\beta_i^t, \beta_i^t \cdot i, \mu_i^t[a_{1:t-1}]} \{ R_i^t(X_t, A_t) + V_{i+1}^i(\mathcal{E}^{\mu_i^t[a_{1:t-1}], \beta_i^t(\cdot|a_{1:t-1}, \cdot), A_t), \mathcal{E} \mid a_{1:t-1}, w_{1:t}) \} .
\]

(41)

\[
V_{i+1}^i(\mathcal{E}^{\mu_i^t[a_{1:t-1}], \beta_i^t(\cdot|a_{1:t-1}, \cdot), A_t), \mathcal{E} \mid a_{1:t-1}, w_{1:t}) .
\]

(42)

Proof: We prove this lemma by contradiction.

Suppose the claim is not true for \( t \). This implies \( \exists i, \hat{\beta}_i^t, \hat{a}_{1:t-1}, \hat{w}_{1:t} \) such that

\[
\mathbb{E}^{\hat{\beta}_i^t, \hat{\beta}_i^t \cdot i, \hat{\mu}_i^t[a_{1:t-1}]} \{ R_i^t(X_t, A_t) + V_{i+1}^i(\mathcal{E}^{\hat{\mu}_i^t[a_{1:t-1}], \hat{\beta}_i^t(\cdot|a_{1:t-1}, \cdot), A_t), \mathcal{E} \mid a_{1:t-1}, \hat{w}_{1:t}) \} > V_i^t(\hat{\mu}_i^t[a_{1:t-1}], \hat{\xi}_i^t) .
\]

(43)

We will show that this contradicts the definition of \( V_i^t \) in (9).

Construct \( \hat{\gamma}_i^t(a_i^t|\hat{\xi}_i^t) = \begin{cases} \hat{\beta}_i^t(a_i^t|\hat{a}_{1:t-1}, \hat{w}_{1:t}) \quad \hat{\xi}_i^t = \hat{\xi}_i^t \\ \text{arbitrary otherwise.} \end{cases} \)

Then for \( \hat{a}_{1:t-1}, \hat{w}_{1:t} \), we have

\[
V_i^t(\hat{\mu}_i^t[a_{1:t-1}], \hat{\xi}_i^t) = \max_{\gamma_i^t(a_i^t|\hat{\xi}_i^t)} \mathbb{E}^{\hat{\gamma}_i^t(a_i^t|\hat{\xi}_i^t)} \{ R_i^t(X_t, A_t) + V_{i+1}^i(\mathcal{E}^{\hat{\mu}_i^t[a_{1:t-1}], \hat{\beta}_i^t(\cdot|a_{1:t-1}, \cdot), A_t), \mathcal{E} \mid a_{1:t-1}, \hat{w}_{1:t}) \} \times \hat{\xi}_i^t(\hat{x}_i^t) \hat{x}_i^t(x_i^t)^{\hat{\mu}_i^t \cdot i}[\hat{a}_{1:t-1}](\hat{\xi}_i^t)^{\hat{\beta}_i^t \cdot i}(\hat{a}_i^t|\hat{a}_{1:t-1}, \hat{w}_{1:t})^Q(\hat{\xi}_i^t, \hat{\xi}_i^t, \hat{a}_i^t) .
\]

(44a)

(44b)

(44c)

\[
(44d)
\]

(44e)

\[
(44f)
\]

where (44a) follows from the definition of \( V_i^t \) in (9), (44d) follows from definition of \( \hat{\gamma}_i^t \) and (44f) follows from (43). However this leads to a contradiction. ■
Lemma 6: \( \forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{i:t}, w_{i:t+1}) \in \mathcal{H}_{i+1}^{t} \) and \( \beta_{i}^{t} \)

\[
\mathbb{E}_{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} \left\{ \sum_{n=t+1}^{T} R(X_n, A_n) \big| a_{i:t}, w_{i:t+1} \right\} = \\
\mathbb{E}_{\beta_{i}^{t+1} > T, \mu_{i}^{t+1} > T} \left\{ \sum_{n=t+1}^{T} R(X_n, A_n) \big| a_{i:t}, w_{i:t+1} \right\}.
\]

(45)

Thus the above quantities do not depend on \( \beta_{i}^{t} \).

Proof: Essentially this claim stands on the fact that \( \mu_{t+1}^{s-1} \) can be updated from \( \mu_{t}^{s-1} | a_{1:t-1}, \beta_{t}^{s-1} \) and \( a_{t} \), as \( \mu_{t+1}^{s-1} | a_{1:t-1} = \prod_{i \neq t} F^{-i} (\mu_{t}^{s-1} | a_{1:t-1}, \beta_{t}^{s-1}, a_{t}) \) as in Lemma [1]. Since the above expectations involve random variables \( X_{t+1:T}, A_{t+1:T} \), we consider \( P^{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} (a_{t+1:T}, a_{1:t}, w_{i:t+1}) \).

\[
P^{\beta_{i}^{t} > T, \mu_{i}^{t} > T} (a_{1:t}, w_{i:t+1}) = \\
P^{\beta_{i}^{t} > T, \mu_{i}^{t} > T} (a_{t+1:T}, a_{1:t}, w_{i:t+1}) \frac{P^{\beta_{i}^{t} > T, \mu_{i}^{t} > T} (a_{t+1:T}, a_{1:t}, w_{i:t+1})}{P^{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} (a_{t+1:T}, a_{1:t}, w_{i:t+1})}.
\]

(46a)

We consider the numerator and the denominator separately. The numerator in (46a) is given by

\[
N_{t} = \sum_{x_{t}, \xi_{t}^{i}} P^{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} (x_{t}, \xi_{t}^{i} | a_{1:t-1}, w_{i:t+1}) \beta_{t}^{i}(a_{t} | a_{1:t-1}, w_{i:t+1}) \beta_{t}^{s-1}(a_{t} | a_{1:t-1}, \xi_{t}^{i}) Q_{x}(x_{t+1} | x_{t}, a_{t})
\]

(46b)

\[
Q_{w}^{i}(w_{i+1} | x_{t+1}, a_{t}) P^{\beta_{i}^{t} > T, \mu_{i}^{t} > T} (a_{1:t}, w_{i:t+1}, x_{t+1})
\]

where (46c) follows from the fact that probability on \( (a_{1:t+1}, x_{t+1}, x_{t+2:T}) \) given \( a_{1:t}, w_{i:t+1}, x_{t+1} \) depends on \( a_{1:t}, w_{i:t+1}, x_{t+1}, \mu_{i}^{t} | a_{1:t} \) through \( \beta_{i}^{t+1:T} \). Similarly, the denominator in (46a) is given by

\[
D_{r} = \sum_{\tilde{x}_{t}, \tilde{\xi}_{t}^{i}, \tilde{x}_{i+1}} P^{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} (\tilde{x}_{t}, \tilde{\xi}_{t}^{i} | a_{1:t-1}, w_{i:t+1}) \beta_{t}^{i}(a_{t} | a_{1:t-1}, w_{i:t+1}) \beta_{t}^{s-1}(a_{t} | a_{1:t-1}, \tilde{\xi}_{t}^{i}) Q_{x}^{i}(\tilde{x}_{i+1} | \tilde{x}_{i+1}, a_{t})
\]

(46d)

\[
Q_{w}^{i}(w_{i+1} | \tilde{x}_{i+1}, a_{t})
\]

(46e)

By canceling the terms \( \beta_{i}(\cdot) \) in the numerator and the denominator, (46a) is given by

\[
\frac{N_{r}}{D_{r}} P^{\beta_{i}^{t-1} > T, \mu_{i}^{t-1} > T} (a_{t+1:T}, x_{t+2:T} | a_{1:t}, w_{i:t+1}, x_{t+1})
\]

(46f)
where

\[ N_r = \sum_{x_t, \xi_t^{-1}} \xi_t(x_t) \mu_t^{a_{1:t-1}}[a_{1:t-1} \mid (\xi_t^{-1}), \beta_t^{a_{1:t-1}}(a_{1:t-1}, \xi_t^{-1})]R_x(x_{t+1} \mid x_t, a_t)Q_x(w_{t+1}^i \mid x_{t+1}^i, a_t) \quad (46g) \]

\[ = \sum_{x_t^i} \xi_t^i(x_t^i)Q_x^i(x_{t+1}^i \mid x_t^i, a_t)Q_w^i(w_{t+1}^i \mid x_{t+1}^i, a_t) \]

\[ = \sum_{x_t^i, \xi_t^{-1}} \xi_t^{-1}(x_t^{-i}) \mu_t^{a_{1:t-1}}[a_{1:t-1} \mid (\xi_t^{-1}), \beta_t^{a_{1:t-1}}(a_{1:t-1}, \xi_t^{-1})]Q_x^{-i}(x_{t+1}^{-i} \mid x_t^{-i}, a_t) \quad (46h) \]

and

\[ D_r = \sum_{\tilde{x}_t, \tilde{\xi}_t^{-1}, \tilde{x}_{t+1}} \xi_t^i(\tilde{x}_t^i) \tilde{\xi}_t^{-i}(\tilde{x}_t^{-i}) \mu_t^{a_{1:t-1}}[a_{1:t-1} \mid (\tilde{\xi}_t^{-1}), \beta_t^{a_{1:t-1}}(a_{1:t-1}, \tilde{\xi}_t^{-1})]Q_x^i(\tilde{x}_{t+1}^i \mid \tilde{x}_{t+1}, a_t)Q_w(\tilde{x}_{t+1}^i \mid \tilde{x}_{t+1}, a_t) \]

\[ = \sum_{\tilde{x}_t^i, \tilde{\xi}_t^{-1}, \tilde{x}_{t+1}} \xi_t^i(\tilde{x}_t^i) \tilde{\xi}_t^{-i}(\tilde{x}_t^{-i}) \mu_t^{a_{1:t-1}}[a_{1:t-1} \mid (\tilde{\xi}_t^{-1}), \beta_t^{a_{1:t-1}}(a_{1:t-1}, \tilde{\xi}_t^{-1})]Q_x^i(\tilde{x}_{t+1}^i \mid \tilde{x}_{t+1}, a_t)Q_w(\tilde{x}_{t+1}^i \mid \tilde{x}_{t+1}, a_t) \quad (46i) \]

Thus (46a) is given by

\[ = \xi_{t+1}(x_{t+1}) \mu_{t+1}^{a_{1:t}}[a_{1:t} \mid (\xi_{t+1}^{-1})]P_{t+1}^{a_{1:t}}(\xi_{t+1}^{-1}, a_{1:t+1} \mid a_{1:t}) \quad (46k) \]

\[ = P_{t+1}^{a_{1:t}}(\xi_{t+1}^{-1}, a_{1:t+1} \mid a_{1:t}) \quad (46l) \]

\[ \text{Lemma 7: } \forall i \in \mathcal{N}, t \in \mathcal{T}, a_{1:t-1} \in \mathcal{H}_t^i, w_{1:t}^i \in (\mathcal{W}^i)^t \]

\[ V_t^i(\mu_t^{a_{1:t-1}}, \xi_t^i) = \mathbb{E}^{\beta_t^{a_{1:T}}[a_{1:T-1}], \mu_t^{a_{1:T-1}}} \left\{ \sum_{n=t}^T R^i(X_n, A_n) \middle| a_{1:t-1}, w_{1:t}^i \right\} \quad (47) \]

\[ \text{Proof:} \]

We prove the Lemma by induction. For \( t = T \),

\[ \mathbb{E}^{\beta_T^{a_{1:T}}, \mu_T^{a_{1:T-1}}} \left\{ R^i(X_T, A_T) \middle| a_{1:T-1}, w_{1:T}^i \right\} \]

\[ = \sum_{x_T^i, a_T} R^i(x_T, a_T) \xi_T(x_T) \mu_T[a_{1:T-1} \mid (\xi_T^{-1}), \beta_T^{a_{1:T-1}}(a_T \mid a_{1:T-1}, \xi_T^i)] \beta_T^{a_{1:T-1}}(a_T \mid a_{1:T-1}, \xi_T^i) \quad (48a) \]

\[ = V_T^i(\mu_T^{a_{1:T-1}}, \xi_T^i) \]

(48b)

where (48b) follows from the definition of \( V_t^i \) in (9) and the definition of \( \beta_T^{a_{1:T}} \) in the forward recursion in (11a).

Suppose the claim is true for \( t + 1 \), i.e., \( \forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, w_{1:t+1}^i) \in \mathcal{H}_t^{i+1} \)

\[ V_{t+1}^i(\mu_{t+1}^{a_{1:t}}, \xi_{t+1}^i) = \mathbb{E}^{\beta_{t+1}^{a_{1:T}}[a_{1:T}], \mu_{t+1}^{a_{1:T}}} \left\{ \sum_{n=t+1}^T R^i(X_n, A_n) \middle| a_{1:t}, w_{1:t+1}^i \right\} \quad (49) \]
Then \( \forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, w_{1:t}) \in \mathcal{H}_{t} \), we have

\[
\begin{align*}
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ \sum_{n=t+1}^{T} R^{i}(X_{n}, A_{n}) \big| a_{1:t-1}, w_{1:t} \right\} \\
= \mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ R^{i}(X_{t}, A_{t}) + \right. \\
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ \sum_{n=t+1}^{T} R^{i}(X_{n}, A_{n}) \big| a_{1:t-1}, A_{t}, w_{1:t}, W_{t+1}^{i} \right\} \bigg| a_{1:t-1}, w_{1:t} \right) \\
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ R^{i}(X_{t}, A_{t}) + \right. \\
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ \sum_{n=t+1}^{T} R^{i}(X_{n}, A_{n}) \big| a_{1:t-1}, A_{t}, w_{1:t}, W_{t+1}^{i} \right\} \bigg| a_{1:t-1}, w_{1:t} \right) \\
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ R^{i}(X_{t}, A_{t}) + V^{i}_{t+1}(\mu_{t+1}^{a_{1:t-1}A_{t+1}}[a_{1:t-1}A_{t}], \Xi_{t+1}^{i}) \bigg| a_{1:t-1}, w_{1:t} \right) \\
\mathbb{E}_{\hat{\beta}_{t}^{*,i}, \hat{\beta}_{t-1}^{*,i}, \mu_{t+1}^{a_{1:t-1}}} & \left\{ R^{i}(X_{t}, A_{t}) + V^{i}_{t+1}(\mu_{t+1}^{a_{1:t-1}A_{t+1}}[a_{1:t-1}A_{t}], \Xi_{t+1}^{i}) \bigg| a_{1:t-1}, w_{1:t} \right) \\
= V^{i}_{t}(\mu_{t}^{a_{1:t-1}A_{t}}, \xi_{t}^{i}),
\end{align*}
\]

where (50b) follows from Lemma 6 in Appendix C, (50c) follows from the induction hypothesis in (49), (50d) follows because the random variables involved in expectation, \( X^{i}_{t-1}, A_{t}, X^{i}_{t+1} \) do not depend on \( \hat{\beta}_{t+1}^{*,i}, \hat{\beta}_{t+1}^{*,i} \) and (50e) follows from the definition of \( \beta^{*} \) in the forward recursion in (11a), the definition of \( \mu_{t+1}^{a_{1:t-1}} \) in (11b) and the definition of \( V^{i}_{t} \) in (9).

**APPENDIX D**

**Proof:** We will prove the result by induction on \( t \). The result is vacuously true for \( T + 1 \).

Suppose it is also true for \( t + 1 \), i.e.

\[
(\mu_{t+1}^{a_{1:t-1}})^{-1}(\hat{C}_{t+1}^{*}) = \hat{C}_{t+1}^{a_{1:t-1}}.
\]

We show that the result holds true for \( t \). In the following two cases, we show that if there exists an element in one set, it also belongs to the other. From the contrapositive of the statement, if one is empty, so is the other.

**Case 1.** We prove \( (\mu_{t}^{a_{1:t-1}})^{-1}(\hat{C}_{t}^{*}) \subset \hat{C}_{t}^{a_{1:t-1}} \).

Let \( h_{t}^{i} \in (\mu_{t}^{a_{1:t-1}})^{-1}(\hat{C}_{t}^{*}) \). We will show that \( h_{t}^{i} \in \hat{C}_{t}^{a_{1:t-1}} \).

Since \( h_{t}^{i} \in (\mu_{t}^{a_{1:t-1}})^{-1}(\hat{C}_{t}^{a_{1:t-1}}) \), this implies \( \mu_{t}^{a_{1:t-1}}[h_{t}^{i}] \in \hat{C}_{t}^{a_{1:t-1}} \). Then by the definition of \( \hat{C}_{t}^{a_{1:t-1}} \), \( \forall i \in \text{supp}(\mu_{t}^{a_{1:t-1}}[h_{t}^{i}]) \), \( \theta_{i}[\mu_{t}^{a_{1:t-1}}[h_{t}^{i}]](a_{t}^{i}[\xi_{t}^{i}]) = 1 \). Since \( \xi_{t}(x_{t}^{i}) = P(x_{t}^{i} | h_{t}^{i}) \forall x_{t}^{i}, \mu_{t}^{a_{1:t-1}}[h_{t}^{i}][\xi_{t}^{i}] = P^{0}(\xi_{t}^{i} | h_{t}^{i}) \forall \xi_{t}^{i} \) and \( \beta_{t}^{*,i}(a_{t}^{i} | h_{t}^{i}) = \theta_{i}[\mu_{t}^{a_{1:t-1}}[h_{t}^{i}]](a_{t}^{i}[\xi_{t}^{i}]) \) by the definition of \( \beta^{*} \), this implies \( \forall i, \beta_{t}^{*,i}(a_{t}^{i} | h_{t}^{i}) = 1, \forall h_{t}^{i} \) that are consistent with \( h_{t}^{i} \) and occur with non-zero probability.

Also since \( \mu_{t}^{a_{1:t-1}}[h_{t}^{i}] \in \hat{C}_{t}^{a_{1:t-1}} \), this implies \( F(\mu_{t}^{a_{1:t-1}}[h_{t}^{i}], \theta_{i}[\mu_{t}^{a_{1:t-1}}[h_{t}^{i}]], a_{t}) \in \hat{C}_{t+1}^{a_{1:t-1}} \) by definition of \( \hat{C}_{t}^{a_{1:t-1}} \).

Thus \( \mu_{t+1}^{a_{1:t-1}}[h_{t}^{i}, a_{t}] \in \hat{C}_{t+1}^{a_{1:t-1}} \), since \( \mu_{t+1}^{a_{1:t-1}}[h_{t}^{i}, a_{t}] = F(\mu_{t}^{a_{1:t-1}}[h_{t}^{i}], \theta_{i}[\mu_{t}^{a_{1:t-1}}[h_{t}^{i}]], a_{t}) \) by definition. Using the
induction hypothesis, \((h^i_t, a_t) \in C_{t+1}^{a_i+T}\), which implies \(\forall i, \beta_{n,i}^+(a_i^n|h^n_i) = 1, \forall n \geq t, \forall h^n_i\) that are consistent with \((h^i_t, a_t)\) and occur with non-zero probability.

The above two facts conclude that \(\forall i, \beta_{n,i}^+(a_i^n|h^n_i) = 1, \forall n \geq t, \forall h^n_i\) that are consistent with \(h^i_t\) and occur with non-zero probability, which implies \(h^i_t \in C_{t+T}^+\) by the definition of \(C_{t+T}^+\).

Case 2. We prove \((\mu^+_t)^{-1}(\tilde{C}_{t+T}^+)) \supset C_{t+T}^+\).

Let \(h^i_t \in C_{t+T}^+\). We will show that \(\mu^+_t[h^i_t] \in \tilde{C}_{t+T}^+\).

Since \(h^i_t \in C_{t+T}^+\), this implies \(\forall i, \beta_{t,i}^+(a_i^t|h^t_i) = 1, \forall h^t_i\) that are consistent with \(h^i_t\) and occur with non-zero probability. Since \(\beta_{t,i}^+(a_i^t|h^t_i) = \theta_i^+[\mu^+_t[h^i_t]](a_i^t|\xi_i^t)\), by the definition of \(\beta^+\), where \(\xi_i^t(\xi_i^t) = P(x^t_i|h^t_i) \forall x^t_i\), this implies \(\forall i, \theta_i^+[\mu^+_t[h^i_t]](a_i^t|\xi_i^t) = 1, \forall \xi_i^t \in \text{supp}(\mu^+_t[h^i_t]), \forall \epsilon_i^t \in \text{supp}(\mu^+_t[h^i_t])\).

Also, since \(h^i_t \in C_{t+T}^+\), it is implied by the definition of \(C_{t+T}^+\) that \((h^i_t, a_t) \in \tilde{C}_{t+1}^{a_i+T}\). This implies \(\mu^+_{t+1}[h^i_t, a_t] \in \tilde{C}_{t+1}^{a_i+T}\) by the induction hypothesis. Since, by definition, \(\mu^+_{t+1}[h^i_t, a_t] = F(\mu^+_t[h^i_t], \theta_i[\mu^+_t[h^i_t]], a_t)\), this implies \(F(\mu^+_t[h^i_t], \theta_i[\mu^+_t[h^i_t]], a_t) \in \tilde{C}_{t+1}^{a_i+T}\).

Since we have shown that \(\forall i, \theta_i^+[\mu^+_t[h^i_t]](a_i^t|\xi_i^t) = 1, \forall \xi_i^t \in \text{supp}(\mu^+_t[h^i_t])\) and \(F(\mu^+_t[h^i_t], \theta_i[\mu^+_t[h^i_t]], a_t) \in \tilde{C}_{t+1}^{a_i+T}\), this implies \(\mu^+_t[h^i_t] \in \tilde{C}_{t+T}^+\) by the definition of \(\tilde{C}_{t+T}^+\).

The above two cases complete the induction step.

### APPENDIX E

**Lemma 8:** Conditioned on \(x^t = 1\), \(\{\Xi^t\}_t\) is a sub-martingale.

**Proof:**

\[
\Xi^t_{t+1} = \begin{cases} 
G^t(\xi_i^t, w^t_{i+1} = 0, a_i^t) = \xi_i^t p_{a_i^t} & \text{with probability } p_{a_i^t} \\
G^t(\xi_i^t, w^t_{i+1} = 1, a_i^t) = \xi_i^t (1 - p_{a_i^t}) & \text{with probability } 1 - p_{a_i^t}
\end{cases}
\]

Thus,

\[
\mathbb{E}[\Xi^t_{t+1} | \Xi^t_i, a_i^t] - \Xi^t_i = \frac{\xi_i^t (p_{a_i^t})^2}{\xi_i^t p_{a_i^t} + (1 - \xi_i^t)(1 - p_{a_i^t})} + \frac{\xi_i^t (1 - p_{a_i^t})^2}{\xi_i^t (1 - p_{a_i^t}) + (1 - \xi_i^t) p_{a_i^t}} - \xi_i^t
\]

\[
= \frac{\xi_i^t (1 - \xi_i^t)^2 (1 - 2 p_{a_i}^t)}{(\xi_i^t p_{a_i}^t + (1 - \xi_i^t)(1 - p_{a_i}^t))(\xi_i^t (1 - p_{a_i}^t) + (1 - \xi_i^t) p_{a_i}^t)}
\]

\[
\geq 0
\]

with the inequality being strict for \(p_{a_i}^t < \frac{1}{2}\) and \(\xi_i^t \notin \{0, 1\} \).

### APPENDIX F

**Proof:** We prove this by induction on \(t_0\). For \(t_0 = T\), \((23)\) reduces to

\[
\hat{\gamma}_T^T(\cdot | \xi_T^T) = \arg \max_{\gamma^{\cdot}_T(\cdot | \xi_T^T)} \sum_{a_T} a_T^T \gamma^T_T(a_T^T|\xi_T^T) (\lambda(2 \xi_T^T - 1) + \hat{\lambda}(2 \xi_T^T - 1)),
\]

\[\text{DRAFT}\]

June 27, 2017
and since $\pi_T \in \hat{C}^a$, it is easy to verify that $\tilde{\gamma}_t^i(a^i|\xi_T^i) = 1, \forall \xi_T^i \in [0,1]$ and thus $V_t^i(\pi_T, \xi_T^i) = (\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1)) a^i$. This establishes the base case.

Now, suppose the claim is true for $t_0 = \tau + 1$ i.e. if $\pi_{\tau + 1} \in \hat{C}^a$, then $\forall t \geq \tau + 1, \pi_t \in \hat{C}^a$ and $\tilde{\gamma}_t^i(a^i|\xi_T^i) = 1 \forall \xi_T^i \in [0,1]$. Moreover, for $\tau + 1 \leq t \leq T$, $V_t^i$ is given by, $\forall \pi_t \in \hat{C}^a$,

$$V_t^i(\pi_t, \xi_T^i) = (T - t + 1)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1)) a^i. \quad (57)$$

Then if $\pi_T \in \hat{C}^a$, then $\tilde{\gamma}_t^i(a^i|\xi_T^i) = 1 \forall \xi_T^i \in [0,1]$ satisfies (23) since,

$$\tilde{\gamma}_t^i(\cdot | \xi_T^i) \in \arg \max_{\gamma_t^i(\cdot | \xi_T^i)} \sum a^i \gamma_t^i(a^i|\xi_T^i)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1))$$

$$+ E(\gamma_t^i(\cdot | \xi_T^i)\tilde{\gamma}_t^i, \pi_t \{ V_{t+1}^i(F(\pi_T, \tilde{\gamma}_T, A_T, \tilde{\Xi}_{t+1}^i) | \xi_T^i) \} \} \quad (58a)$$

$$= \arg \max_{\gamma_t^i(\cdot | \xi_T^i)} \sum a^i \gamma_t^i(a^i|\xi_T^i)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1))$$

$$+ E(\gamma_t^i(\cdot | \xi_T^i)\tilde{\gamma}_t^i, \pi_t \{ (T - \tau)(\lambda(2\Xi_T^i_{t+1} - 1) + \bar{\lambda}(2\hat{T}_T^i - 1))a^i|\xi_T^i \} \} \quad (59a)$$

$$= \arg \max_{\gamma_t^i(\cdot | \xi_T^i)} \sum a^i \gamma_t^i(a^i|\xi_T^i)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1))$$

$$+ (T - \tau)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1))a^i \quad (59b)$$

$$= \arg \max_{\gamma_t^i(\cdot | \xi_T^i)} \sum a^i \gamma_t^i(a^i|\xi_T^i)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1)), \quad (59c)$$

where (59a) follows from the fact that $F(\pi_T, \tilde{\gamma}_T, a_T) \in C^a, \forall a_T$, as shown in Lemma 9 and induction hypothesis, (59b) follows from Lemma 9 and Lemma 10 and (59c) follows from the fact that the second term does not depend on $\gamma_t^i(\cdot | \xi_T^i)$. This also shows that, $\forall \pi_t \in \hat{C}^a$,

$$V_t^i(\pi_T, \xi_T^i) = (T - \tau + 1)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{T}_T^i - 1)) a^i, \quad (60)$$

which completes the induction step.

\textbf{Lemma 9:} Expectation of $\pi_{t+1}$ under non-informative $\tilde{\gamma}_t^i$ of the form $\tilde{\gamma}_t^i(a^i|\xi_T^i) = 1 \forall \xi_T^i \in [0,1]$, remains the same as mean of $\pi_t^i$, i.e.,

$$E\{\Xi_{t+1}^i(1)|\pi_t^i, \tilde{\gamma}_t^i, a^i\} = E\{\Xi_1^i(1)|\pi_1^i\} \quad (61)$$
Lemma 10:

\[ E_\{\Xi_{t+1}(1)|\pi_t, \hat{\gamma}_t, a^t\} \]

\[ = \sum_{\xi_{t+1}^t} \xi_{t+1}^t(1)F^t(\pi_t, \hat{\gamma}_t, a^t)(\xi_{t+1}^t) \]

(62a)

\[ = \sum_{\xi_t^t, \xi_{t+1}^t} \xi_t^t(1)\pi_t^t(\xi_t^t)\xi_{t+1}^t(1)E_{\{\Xi_{t+1}(1)|\xi_{t+1}^t\}} \]

(62b)

\[ = \sum_{\xi_t^t, \xi_{t+1}^t} \xi_t^t(1)\pi_t^t(\xi_t^t)\xi_{t+1}^t(1)E_{\{\Xi_{t+1}(1)|\xi_{t+1}^t\}} \]

(62c)

\[ = \sum_{\xi_t^t, \xi_{t+1}^t} \xi_t^t(1)\pi_t^t(\xi_t^t)\xi_{t+1}^t(1)E_{\{\Xi_{t+1}(1)|\xi_{t+1}^t\}} \]

(62d)

\[ = \sum_{\xi_t^t, \xi_{t+1}^t} \xi_t^t(1)\pi_t^t(\xi_t^t)\xi_{t+1}^t(1)E_{\{\Xi_{t+1}(1)|\xi_{t+1}^t\}} \]

(62e)

\[ = \sum_{\xi_t^t, \xi_{t+1}^t} \xi_t^t(1)\pi_t^t(\xi_t^t)\xi_{t+1}^t(1)E_{\{\Xi_{t+1}(1)|\xi_{t+1}^t\}} \]

(62f)

\[ = \sum_{\xi_t^t} \xi_t^t(1)\pi_t^t(\xi_t^t) \]

(62g)

Proof:

\[ E\{\Xi_{t+1}(1)|\pi_t, \hat{\gamma}_t, a^t\} \]

(63)

\[ = \xi_t^t(1) \]

(64a)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(64b)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(64c)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(64d)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(64e)

Proof:

\[ E\{\Xi_{t+1}(1)|\xi_t^t, \gamma_t^t\} \]

(65)

\[ = \xi_t^t(1) \]

(66a)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(66b)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(66c)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(66d)

\[ = \xi_t^t(1)I_{F, \{\xi_t^t, \xi_{t+1}^t, a^t\}}(\xi_{t+1}^t) \]

(66e)
REFERENCES


