On the Difficulty of Achieving Equilibrium in Interactive POMDPs

Prashant Doshi PDOSHI@CS.UGA.EDU
Department of Computer Science
University of Georgia
Athens, GA 30606

Piotr J. Gmytrasiewicz PIOTR@CS.UIC.EDU
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607

Abstract

We analyze the asymptotic behavior of agents engaged in an infinite horizon partially observable stochastic game as formalized by the interactive POMDP framework. We show that when agents' initial beliefs satisfy a truth compatibility condition, their behavior converges to a subjective ϵ-equilibrium in a finite time, and subjective equilibrium in the limit. This result is a generalization of a similar result in repeated games, to partially observable stochastic games. However, it turns out that the equilibrating process is difficult to demonstrate computationally because of the difficulty in coming up with initial beliefs that are both natural and satisfy the truth compatibility condition. Our results, therefore, shed some negative light on using equilibria as a solution concept for decision making in partially observable stochastic games.

1. Introduction

We analyze the interactions taking place between agents participating in an infinite horizon partially observable stochastic game formalized within the framework of interactive POMDPs (I-POMDPs) [5, 6]. I-POMDPs represent and solve a partially observable stochastic game (POSG) from the perspective of an agent playing the game. This approach, also called the decision-theoretic approach to game theory [10], differs from the objective representation of POSGs as outlined in [8]. We consider the setting in which an agent may be unaware of the other agents' behavioral strategies, it is uncertain about their observations, and it may be unable to perfectly observe the other agents' actions. In accordance with Bayesian decision theory, the agent maintains and updates its belief about the physical state as well as the strategies of the other agents, and its decisions are best responses to its beliefs.

Under the assumption of compatibility of agents' prior beliefs about future observations with the true distribution induced by the actual strategies of all agents, we show that for agents modeled within the I-POMDP framework, the following properties hold: (i) the agents' beliefs about the future observation paths of the game coincide in the limit with the true distribution over the future, and (ii) the agents' beliefs about the opponents' strategies do not change in the limit. Strategies with these properties are said to be in *subjective equilibrium*, which is stable with respect to learning and optimization. One way (and possibly the only way in the absence of any information about the other agent's true strategy) to satisfy the truth compatibility condition on prior beliefs is to consider beliefs that assign a non-zero probability to each possible strategy of other agents. In other words, the beliefs must have a *grain of truth*. However, for computable strategies, we show by borrowing a result from Nachbar and Zame [14] that it is impossible for all the agents' beliefs that assign some non-zero probability to each possible strategy of the others, to simultaneously satisfy the grain of truth assumption. While this negative result does not question the existence of the equilibrium nor does it preclude reaching the equilibrium, it does point out the difficulty in guaranteeing it within the I-POMDP framework. Specifically, prior beliefs that satisfy the truth compatibility condition must be unnatural – the agents will have to start the game convinced that others will not act according to some strategies.

In prior work, Kalai and Lehrer [11, 12] have shown that the strategies of agents engaged in infinitely repeated games with discounted payoffs, who are unaware of others' strategies, and under the assumptions of perfect observability of others' actions (perfect monitoring) and truth compatibility of prior beliefs will converge to a subjective equilibrium. We complement this result by showing the asymptotic existence of subjective equilibrium in a more general and realistic multiagent setting, one in which the assumptions of perfect observability of state and others' actions have been relaxed. Additionally, we address the research problem posed in [11] regarding the existence of subjective equilibrium in POSGs. Hahn [7] introduced the concept of a *conjectural equilibrium* in economies where the signals generated by the economy do not cause changes in the agents' theories, nor do they induce changes in the agents' policies. Fudenberg and Levine [4] consider a general model of finitely repeated extensive form games wherein strategies of opponents may be correlated (unlike [11] where strategies are assumed independent), and show that behavior of agents that maintain beliefs and optimize according to their beliefs, converges to a *self-confirming equilibrium*. There is a strong link between the subjective equilibrium and its objective counterpart – the Nash equilibrium. Specifically, under the assumption of perfect monitoring, both [11] and [4] show that the strategy profile in subjective and self-confirming equilibrium induce a distribution over the future action paths that coincides with the distribution induced by a set of strategies in Nash equilibrium. Of course, this does not imply that strategies in subjective equilibrium are also in Nash equilibrium; however, the converse is always true. While proving a similar link between subjective and Nash equilibrium for POSGs is beyond the scope of this paper, we conjecture its existence. Work of a similar vein is reported in [9]. It assumes agents have a common prior over the possible types of agents engaged in a repeated game, and shows that the sequence of Bayesian-Nash equilibrium beliefs of agents converges to a Nash equilibrium.

2. Overview of Interactive POMDPs

Interactive POMDPs [5, 6] generalize POMDPs to account for presence of other agents in the environment. They do this by including models of other agents in the state space. Models of other agents, analogous to *types* in game theory, encompass all private information influencing their behavior.

For simplicity of presentation let us consider an agent, i, that is interacting with one other agent, j. The formalism easily generalizes to a larger number of agents.

I-POMDP An *interactive POMDP* of agent i, *I-POMDP*$_i$, is:

$$\textit{I-POMDP}_i = \langle IS_i, A, T_i, \Omega_i, O_i, R_i \rangle$$

where:

- IS_i is a set of **interactive** states defined as $IS_i = S \times \langle \mathcal{O}_j \times M_j \rangle$, where S is the set of states of the physical environment, and $\langle \mathcal{O}_j \times M_j \rangle$ is the set of pairs consisting of a possible observation function and a model of agent j. Each model, $m_j \in M_j$, is a pair $m_j = \langle h_j, \pi_j \rangle$, where $\pi_j : H_j \rightarrow \Delta(A_j)$ is j's policy tree (strategy), assumed computable [1], which maps possible histories of j's observations to distributions over its actions. [2] h_j is an element of H_j.[3] $O_j \in \mathcal{O}_j$, also computable, specifies the way in which the environment is supplying the agent with its input.
- $A = A_i \times A_j$ is the set of joint moves of all agents
- T_i is a transition function, $T_i : S \times A \times S \rightarrow [0, 1]$ which describes results of agents' actions on the physical state
- Ω_i is the set of agent i's observations
- O_i is an observation function $O_i : S \times A \times \Omega_i \rightarrow [0, 1]$
- R_i is defined as $R_i : S \times A \rightarrow$ **R**. We allow the agent to have preferences over the physical states and actions of all agents.

The task of computing a solution for an I-POMDP, similar to that of a POMDP, can be decomposed into two steps: (1)**Belief update** during which the agent updates its belief to reflect newly available information. (2)**Policy computation** during which the agent computes the optimal action(s) to perform from each belief state.

2.1 Bayesian Belief Update

There are two differences that complicate state estimation in multiagent settings, when compared to single agent ones. First, since the state of the physical environment depends on the actions performed by both agents, the prediction of how the physical state changes has to be made based on the predicted actions of the other agent. The probabilities of other's actions are obtained based on their models. Thus, as opposed to the literature on learning in repeated games, we do not assume that actions are fully observable by other agents. Rather, agents can attempt to infer what actions other agents have performed by sensing their results on the environment. Second, changes in the models of other agents have to be included in the update. Specifically, update of the other agent's models due to its new observation has to be included. In other words, the agent has to update its beliefs based on what it anticipates that the other agent observes and how it updates. Consequently, an agent's beliefs record what it thinks about how the other agent will behave as it learns. For simplicity we decompose the I-POMDP belief update into two steps:

[1] We assume computability in the Turing machine sense, i.e. strategies are (total) recursive functions.

[2] Note that if $|A_j| \geq 2$, then the space of policy trees is uncountable; however, by assuming π_j to be computable, we restrict the space to be countable.

[3] In [5, 6], we replace $\langle \mathcal{O}_j \times M_j \rangle$ with a special class of models called *intentional models*. These models ascribe beliefs, preferences, and rationality to other agents. We do not introduce those models here for the purpose of generality.

• *Prediction* When an agent, say i, with a previous belief, b_i^{t-1}, performs a control action a_i^{t-1} and if the other agent performs its action a_j^{t-1}, the predicted belief state is,

$$\begin{aligned} Pr(is^t|a_i^{t-1},a_j^{t-1},b_i^{t-1}) = & \sum_{IS^{t-1}:(f_j,O_j)^{t-1}=(f_j,O_j)^t} b_i^{t-1}(is)Pr(a_j^{t-1}|m_j)T(s^{t-1},a_i^{t-1},a_j^{t-1},s^t) \\ & \times \sum_{\omega_j^t} O_j(s^t,a_i^{t-1},a_j^{t-1},\omega_j^t)\delta(\text{APPEND}(h_j^{t-1},\omega_j^t)-h_j^t) \end{aligned}$$

where δ is the Kronecker delta function, and APPEND$(\cdot,\cdot)$ returns a string in which the second argument is appended to the first.

•*Correction* When agent i perceives an observation, ω_i^t, the intermediate belief state, $Pr(\cdot|a_i^{t-1},a_j^{t-1},b_i^{t-1})$, is corrected according to,

$$Pr(is^t|\omega_i^t,a_i^{t-1},b_i^{t-1}) = \alpha\sum_{a_j^{t-1}} O_i(s^t,a_i^{t-1},a_j^{t-1},\omega_i^t)Pr(is^t|a_i^{t-1},a_j^{t-1},b_i^{t-1})$$

where α is the normalizing constant.

The update extends to more than two agents in a straightforward way. We represent possible correlations between actions of other agents as dependencies between their models, which are expressed in i's beliefs.

2.2 Policy Computation

Each belief state in I-POMDP has an associated value reflecting the maximum payoff the agent can expect in this belief state:

$$V(b_i) = \max_{a_i \in A_i}\left\{\sum_{is} ER_i(is,a_i)b_i(is) + \gamma\sum_{\omega_i\in\Omega_i} Pr(\omega_i|a_i,b_i)V(SE(b_i,a_i,\omega_i))\right\} \tag{1}$$

where, $ER_i(is,a_i) = \sum_{a_j} R_i(is,a_i,a_j)Pr(a_j|m_j)$ (since $is = (s,m_j)$), and $SE(\cdot)$ represents the belief update defined previously in Section 2.1. Eq. 1 is a basis for value iteration in I-POMDPs. As shown in [5], the value iteration converges in the limit.

Agent i's optimal action, a_i^*, for the case of infinite horizon criterion with discounting, is an element of the set of optimal actions for the belief state, $OPT(b_i)$, defined as:

$$OPT(b_i) = \underset{a_i \in A_i}{argmax}\left\{\sum_{is} ER_i(is,a_i)b_i(is) + \gamma\sum_{\omega_i\in\Omega_i} Pr(\omega_i|a_i,b_i)V(SE(b_i,a_i,\omega_i))\right\} \tag{2}$$

Equation 2 enables the computation of a policy tree, π_i, for each belief b_i. The policy, π_i, gives i's best response long term strategy for the belief.

For additional details on the I-POMDP framework, and how they compare with other multiagent planning frameworks, see [5].

2.3 Background: Stochastic Processes, Martingales, and Bayesian Learning

A stochastic process is a sequence of random variables, $\{X_t\}, t = 0,1,\ldots$, whose values are realized one at a time. Well-known examples of stochastic processes are Markov chains, as well as sequences of beliefs updated using the Bayesian update. Bayesian learning turns out to exhibit an additional property that classifies it as a special type of stochastic process, called a Martingale.

A Martingale is a stochastic process that, for any observation history up to time t, h^t, exhibits the property that for all $l \geq t$:

$$E[X_l|h^t] = X_t.$$

Consequently, for all future time points $l \geq t$ the expected change, $E[X_l - X_t|h^t] = 0$. A sequence of an agent's beliefs updated using Bayesian learning is known to be a Martingale. Intuitively, this means that the agent's current estimate of the state is equal to what the agent expects its future estimates of the state will be, based on its current observation history. Because the Martingale property of Bayesian learning is central to our results, we sketch a formal proof below.

Let an agent's initial belief over some state space Θ be $X_0 = Pr(\theta)$. The agent receives some observation, ω, in the future according to a distribution ϕ that depends on θ. Let the future revised belief be $X_1 = Pr(\theta|\omega)$. By Bayes theorem, $Pr(\theta|\omega) = \phi(\omega|\theta)Pr(\theta)/Pr(\omega)$. We will show that $E[Pr(\theta|\omega)] = Pr(\theta)$:

$$\begin{aligned} E[Pr(\theta|\omega)] &= \sum_\omega Pr(\theta|\omega)Pr(\omega) = \sum_\omega \frac{\phi(\omega|\theta)Pr(\theta)}{Pr(\omega)} Pr(\omega) \\ &= \sum_\omega \phi(\omega|\theta)Pr(\theta) = Pr(\theta)\sum_\omega \phi(\omega|\theta) \\ &= Pr(\theta) = X_0 \end{aligned}$$

The above result extends immediately to observation histories of any length t. Formally, $E[X_{t+1}|h^t] = X_t$, therefore the beliefs satisfy the Martingale property.

All Martingales share the following convergence property:

Theorem 1 (Martingale Convergence Theorem (§4 of Chapter 7 in [3]). *If $\{X_t\}, t = 0, 1, \ldots$ is a Martingale with $E[X_t^2] < U < \infty$ for some U and all t, then the sequence of random variables, $\{X_t\}$, converges with probability 1 to some X_∞ in mean-square.*

3. Subjective Equilibrium in I-POMDPs

In Section 2, we reviewed a framework for two-agent POSG in which each agent computes the discounted infinite horizon strategy which is the subjective best response of the agent to its belief. During each step of game play, the agent starting with a prior belief revises it in light of the new information using the Bayesian belief update process outlined in Section 2.1, and computes the optimal strategy given its beliefs. The latter step is equivalent to using its observation history to index into its policy tree (computed offline using the process given in Section 2.2) [4]– to compute the best response future strategy.

3.1 Truth Compatible Beliefs

We investigate the asymptotic behavior of agents playing an infinite horizon POSG, in which each agent learns and optimizes. Sequential behavior of agents in a POSG may be represented using their observation histories. For an agent, say i, let ω_i^t be its observation at time step t. Let $\omega^t = [\omega_i^t, \omega_j^t]$. An observation history of the game is a sequence, $h = \{\omega^t\}, t = 1, 2, \ldots$. The set of all histories is, $H = \cup_{t=1}^{\infty}\Omega^t$ where $\Omega^t = \Pi_1^t(\Omega_i \times \Omega_j)$. The set of observation histories upto time t is, $H^t = \Pi_1^t(\Omega_i \times \Omega_j)$, and the set of future observation paths from time t onwards is, $H_t = \Pi_t^\infty(\Omega_i \times \Omega_j)$.

Example: We use the multiagent tiger problem described in [5] as a running example throughout this paper. Briefly, the problem consists of two doors, behind one is a tiger and behind the other is some gold, and two agents i and j. The agents are unaware of where the tiger is (TL or TR), and each can either open any one of two doors, or listen (OL,OR, or L). On opening any door, the tiger appears randomly behind a door, the next time. A tiger emits a growl periodically, which reveals its position behind a door (GL or GR) but only with some certainty. Additionally, each agent can also hear a creak with some certainty, if the other agent opens a door (CL,CR, or S). We will assume that neither agent can perceive other's observations nor actions. The problem is non-cooperative since either i or j may open a door, thereby reseting the location of the tiger, and rendering any information collected by the other agent about the tiger's location useless to it. Example histories in the multiagent tiger problem are shown in Fig. 1.

In the I-POMDP framework, each agent's belief over the physical state and others' candidate models, together with the agent's perfect information regarding its own model, induces a predictive probability distribution over the future observation paths. These distributions play a critical role in our analysis; we represent them mathematically using a collection of probability measures, $\{\mu_k\}, k = 0, i, j$ defined over the space $M \times H$, where $M = M_i \times M_j$ and H is as defined previously, such that,

1. μ_0 is the objective true distribution over models of each agent and the histories,
2. $proj_{M_k}\ \mu_k = proj_{M_k}\ \mu_0 = \delta_{m_k} \quad k = i, j$

Here, condition 2 states that each agent knows its own model (δ_{m_k} is the Kronecker delta function). Additionally, $proj_H\ \mu_0$ gives the true distribution over the histories as induced by the initial strategy profile, and $proj_H\ \mu_k(\cdot|b_k^0)$ for $k = i, j$ gives the predictive probability distribution for each agent over the histories at the start of the game. [5]

[4] In the infinite horizon case, convergence of value iteration allows us to conveniently represent the policy tree as a finite state machine.

[5] Following [9, 15] the unconditional measure μ_k may be seen as a prior before an agent knows its own model, and μ_k along with the conditions as an *interim* prior once an agent knows its own model.

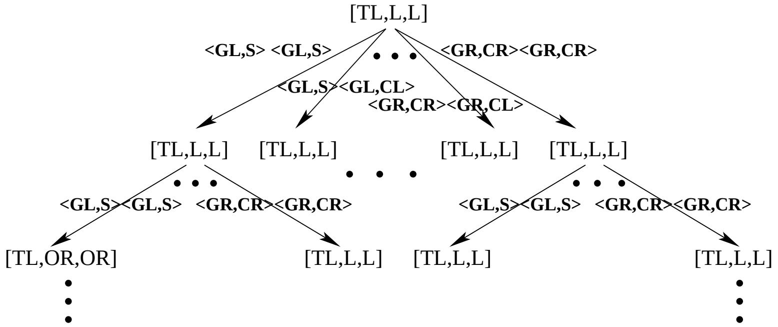

Figure 1: Observation histories in the infinite horizon multiagent tiger problem. The nodes represent the state of the game and play of agents, while the edges are labeled with the possible observations. This example starts with the tiger on the left and each agent listening. Each agent may receive one of six observations (labels on the arrows), and performs an action that optimizes its resulting belief.

If the actual sequence of observations in the game does not proceed along a history that is assigned some positive predictive probability by an agent, then the agent's observations would contradict its beliefs and the Bayesian update would not be possible. Clearly, it is desirable for each agent's initial belief to assign nonzero probability to each possible observation history; this is called the truth compatibility condition. To formalize this condition we need a notion of absolute continuity of two probability measures.

Definition 1 (Absolute Continuity). *A probability measure p_1 is absolutely continuous with p_2, denoted as $p_1 \ll p_2$, if $p_2(E) = 0$ implies $p_1(E) = 0$, for any measurable set E.*

Condition 1 (Absolute Continuity Condition (ACC)). *ACC holds for any agent $k = i, j$ if $proj_H\ \mu_0 \ll proj_H\ \mu_k(\cdot|b_k^0)$.*

Condition 1 states that the probability distribution induced by an agent's initial belief on observation histories should not rule out positive probability events according to the real probability distribution on the histories. A sure way to satisfy ACC is for each agent's initial belief to have a "*grain of truth*" – assign a non-zero probability to the true model of the other agent. Since an agent has no way of knowing the true model of its opponent from beforehand, it must assign a non-zero probability to each candidate model of the other agent. We emphasize that the grain of truth assumption on the agents' prior beliefs is a stronger assumption than ACC. In other words, as we show later, it is possible to satisfy the ACC while violating the grain of truth assumption.

3.2 Subjective Equilibrium

Truth compatible beliefs of an agent that performs Bayesian learning tend to converge in the limit to the opponent model(s) that most likely generates the observations of the agent. In the context of the I-POMDP framework, an agent's belief over the other's models updated using the process outlined in Section 2.1, will converge in the limit. Formally,

Proposition 1 (Bayesian Learning in I-POMDPs). *For an agent in the I-POMDP framework, if its initial belief over the other's models satisfies the ACC, its posterior beliefs will converge with probability 1.*

Proof. As we mentioned before, Bayesian learning is a Martingale. In Section 2.3, set $\Theta = M_j$, and $\phi = O_i$. Noting that the I-POMDP belief update is Bayesian, its Martingale property follows from applying the proof appropriately. In order to apply Theorem 1 to the I-POMDP belief update, set $X_t = Pr(m_j|h_i^t)$ where h_i^t is agent i's observation history up to time t. We must first show that $E[X_t^2]$ is bounded.

$$\begin{aligned} E[|b_i^t|^2] &= \textstyle\sum_{k=1}^{(|A_i||\Omega_i|)^t} |b_i^t = \widehat{b}_i^k|^2 Pr(\widehat{b}_i^k) = \sum_{k=1}^{(|A_i||\Omega_i|)^t} \sum_{M_j} \widehat{b}_i^k(m_j)^2 Pr(b_i^k) && (L_2 \text{ norm}) \\ &\leq \textstyle\sum_{k=1}^{(|A_i||\Omega_i|)^t} 1 \cdot Pr(\widehat{b}_i^k) && (\textstyle\sum p^2 \leq 1) \\ &= 1 \end{aligned}$$

Proposition 1 now follows from a straightforward application of Theorem 1. □

The above result does not imply that an agent's belief always converges to the true model of the other agent. This is due to the possible presence of *observationally equivalent* models of the other agent. For example, for agent i, all models of j

that induce identical distributions over all possible future observation paths are said to be observationally equivalent. When a particular observation history obtains, agent i is unable to distinguish between the observationally equivalent models of j. In other words, observationally equivalent models generate distinct behaviors for histories which are never observed. Models that are observationally equivalent to the true model are also the reason why it is possible for prior beliefs to violate the grain of truth assumption, and yet satisfy ACC.

Example: For an example of observationally equivalent models, consider a version of the multiagent tiger game in which the tiger persists behind its original door once any door has been opened. Additionally, i has superior observation capabilities compared to j, and each agent is able to perfectly observe other's actions but observes the growls imperfectly. Let i's utility dictate that it will not open any doors until it's 100% certain that the tiger is behind the opposite door. The corresponding strategy for i is to listen for an infinite number of time steps. Suppose that as a best response to its belief, j were to adopt a strategy in which it would listen for an infinite number of steps, but if at any time i opened a door, it would also do so at the next time step and then continue opening the same door. The true distribution assigns a probability 1 to the histories $\{[\langle GL|GR,S\rangle,\langle GL|GR,S\rangle]\}_1^\infty$. Instead of the above mentioned strategy if j were to adopt a follow-the-leader strategy, i.e. j performs the action which i did in the previous time step, then the true distribution would again assign probability 1 to the previously mentioned histories. The two different strategies of j turn out to be observationally equivalent for i.

An immediate consequence of the convergence of Bayesian learning is that the predictive distribution over the future observation paths induced by each agent's belief after a finite sequence of observations h_k^t, $proj_{H_t}\ \mu_k(\cdot|b_k^0,h_k^t),\ k=i,j$ becomes arbitrary close to the true distribution, $proj_{H_t}\ \mu_0(\cdot|h^t)$, for a finite t, and converges uniformly in the limit. This is an important result, because it establishes that no matter what the initial beliefs of the agents about the future are, provided that these beliefs are truth compatible, the agents' opinions (about the future) will merge and correctly predict the true future in the limit. This result was first noted in [2]; we present it below and refer the reader to the paper for its proof.

Lemma 1 (Blackwell and Dubins [2]). *Suppose that P is a predictive probability on X, and Q is absolutely continuous w.r.t. P. Then for each conditional distribution $P^t(x_1,\ldots,x_t)$ of the future given the past w.r.t. P, there exists a conditional distribution $Q^t(x_1,\ldots,x_t)$ of the future given the past w.r.t. Q such that, $||P^t(x_1,\ldots,x_t)-Q^t(x_1,\ldots,x_t)|| \underset{t\to\infty}{\to} 0$ with Q-probability 1.*

We use Lemma 1 to establish predictive convergence in the context of the I-POMDP framework.

Proposition 2 (ϵ-Predictive Convergence in I-POMDPs). *For all agents in the I-POMDP framework, if their initial beliefs satisfy the ACC, then for every $\epsilon>0$, there exists a finite T which is a function of ϵ, such that for all $t\geq T$ and with μ_0-probability 1,*

$$||proj_{H_t}\ \mu_0(\cdot|h^t)-proj_{H_t}\ \mu_k(\cdot|b_k^0,h_k^t)||\leq\epsilon \quad \textit{for} \quad k=i,j$$

Proof. Referring to Lemma 1, let $X=H$. We observe that $proj_H\ \mu_0$ and $proj_H\ \mu_k(\cdot|b_k^0)$ for $k=i,j$ are predictive as defined in [2]. Set $Q=proj_H\ \mu_0$, and $P=proj_H\ \mu_k(\cdot|b_k^0)$. Subsequently, $Q^t=proj_{H_t}\ \mu_0(\cdot|h^t)$, and $P^t=proj_{H_t}\ \mu_k(\cdot|b_k^0,h_k^t)$. Proposition 2 then follows immediately from a straightforward application of Lemma 1. □

We have shown that for a POSG modeled using the I-POMDP formalism, the players' beliefs over opponent's models converge in the limit if they satisfy the ACC property. However, the limit beliefs may be incorrect, due to the inability of agents to distinguish between observationally equivalent models of the opponent on the basis of the observation history. Nevertheless, their beliefs over the future paths come arbitrary close, and remain close, to the true distribution over the future, after a finite amount of time. Further observations will only confirm their beliefs about the truth, and will not alter their beliefs. We capture this notion using the concept of a subjective equilibrium [11], defined as follows:

Definition 2 (Subjective ϵ-Equilibrium). *Let $b_k^t,\ \ k=i,j$ be the agents' beliefs at some time t. A pair of policy trees, $\pi^*=[\pi_i^*,\pi_j^*]$ is a subjective ϵ-equilibrium if,*

1. $\pi_i^*\in OPT(b_i^t),\pi_j^*\in OPT(b_j^t)$
2. $||proj_{H_t}\ \mu_0(\cdot|h^t)-proj_{H_t}\ \mu_k(\cdot|b_k^0,h_k^t)||\leq\epsilon,\ \ k=i,j$ *with a μ_0-probability 1.*

For ϵ = 0, subjective equilibrium obtains. Condition 1 of subjective ϵ-equilibrium states that agents are subjectively rational, i.e. their strategies are best responses to their beliefs. As we mentioned before, these strategies are the policy trees computed using Equations 1 and 2. The second condition states that the agents' beliefs have attained ϵ-predictive convergence. In other words, a strategy profile is in subjective ϵ-equilibrium when the strategies are best responses to agents' beliefs that have attained ϵ-predictive convergence.

We now establish a key result of this paper, which is that behavior strategies of agents playing a POSG modeled using the I-POMDP framework, attain subjective ϵ-equilibrium in finite time and subjective equilibrium in the limit, provided that their initial beliefs satisfy the ACC.

Proposition 3 (Convergence to Subjective Equilibrium in I-POMDPs). *Let $\pi = [\pi_i, \pi_j]$ be the strategies of agents i, and j respectively, playing a POSG modeled using the I-POMDP formalism. Let b_i^0, and b_j^0 be their initial beliefs. If the following conditions are met,*

1. $\pi_i \in OPT(b_i^0), \pi_j \in OPT(b_j^0)$
2. $proj_H\ \mu_0 \ll proj_H\ \mu_k(\cdot|b_k^0),\ \ k = i, j$ *(ACC)*

then for any $\epsilon > 0$, and for all μ_0-positive probability histories, there exists some finite time step T which is a function of ϵ, such that for all $t \geq T$, the strategy profile, $\pi^* = [\pi_i^*, \pi_j^*]$ is a subjective ϵ-equilibrium where,

- *b_i^t and b_j^t are the agents' beliefs at time t*
- $\pi_i^* \in OPT(b_i^t), \pi_j^* \in OPT(b_j^t)$

Proof. Proposition 3 follows in part from Proposition 2, and in part from noting that agents in the I-POMDP framework compute strategies that are best responses to their posterior beliefs at each time step, and that the beliefs are updated using their observation history. □

Strategy profiles in subjective ϵ-equilibrium for arbitrarily small $\epsilon \geq 0$ are stable. Specifically, further play will bring agents' beliefs over the future closer to the truth statistically, and the corresponding strategy profiles will remain in the subjective ϵ-equilibrium. Note that ACC is a sufficient condition, but not a necessary one. An example setting in which even though ACC is violated, yet subjective ϵ-equilibrium still results is given in [11].

4. Limitations

Recall that in order to guarantee subjective equilibrium, the agents' prior beliefs must satisfy the ACC condition. We now investigate how to satisfy this condition. As we mentioned previously, since an agent has no way of knowing the true model of the other a'priori, it must assign some probability to all possible models of the other agent. As a result, agents' beliefs will exhibit a grain of truth. However, under the assumption of computability of the agents' strategies, we observe that it is impossible for all the agents' beliefs to simultaneously satisfy the grain of truth. In order to show this, we borrow a result from [14].

Lemma 2 (Nachbar and Zame [14]). *There exists a subgame perfect equilibrium strategy profile, [π_1, π_2], for which π_2 is computable, and no exact best response to π_2 is computable.*

Proposition 4 (Impossibility Result). *Within the I-POMDP framework, agents' prior beliefs that assign a non-zero probability to all possible models of the other, cannot simultaneously satisfy the grain of truth assumption.*

Proof. We first note that within the I-POMDP framework, the model spaces, M_i and M_j, are restricted to the set of computable models. Let b_i^0 be agent i's initial belief that assigns a non-zero probability to all models in M_j. Lemma 2 implies that the support of b_i^0 must contain a strategy of j for which no computable best response exists. This implies that $OPT(b_i^0)$ – i's best response strategy to its belief – is not computable, since in computing OPT, we must enumerate all possible models of j. Therefore, j's belief that has a full support in M_i, fails to account for the true strategy of i. □

Consequences of this negative result point toward a subtle tension between learning (to predict the true distribution over the observation histories) and optimization. Indeed, as Binmore indicates in [1], this conflict is at the heart of why perfect rationality is an unattainable ideal. Binmore proves that a Turing machine cannot always predict truthfully the behavior of an opponent Turing machine (given its complete description) and optimize simultaneously. Similar questions about the plausibility of realizing the subjective equilibrium in repeated games have also been raised, for example see [13]. We emphasize that

Proposition 4 applies only to *exact* best responses. One may always find a computable ϵ-optimal response by optimizing over a long enough finite horizon[6].

While Proposition 4 does not question the existence of equilibrium, it bears relevance to whether the equilibrium can be realized. The implication of Proposition 4 is that it is difficult to satisfy the truth compatibility condition in practice, and therefore ensure convergence to equilibrium. Consequently, our results cast a negative shadow on using equilibrium as a solution concept for decision making in POSGs.

5. Discussion

In this paper we theoretically analyzed the play of agents engaged in a partially observable stochastic game formalized using the interactive POMDP framework. In particular, we have considered subjectively rational agents which may not know others' strategies. Therefore, they maintain beliefs over the physical state and models of other agents and optimize with respect to their beliefs. Within this framework, we first proved that if agents' beliefs satisfy a truth compatibility condition, then strategies of agents that learn and optimize converge to the subjective equilibrium in the limit, and subjective ϵ-equilibrium for arbitrarily small $\epsilon > 0$ in finite time. This result is an extension of a similar result in repeated games, to POSGs as formalized by the I-POMDP framework. Secondly, we argued about the implausibility of satisfying the truth compatibility condition and therefore reaching the equilibrium in practice. Our results therefore bear some relevance to the notion of using equilibrium for solving POSGs. As part of future work, we are investigating the relation between subjective and Nash equilibrium. Since a link between the two has already been established for repeated games, its existence for POSGs can be speculated.

6. Acknowledgements

This research was supported by the NSF CAREER award IRI-9702132 and NSF award IRI-0119270. Prashant would also like to acknowledge the support of a UIC University Fellowship.

References

[1] Ken Binmore. *Essays on Foundations of Game Theory*. Pittman, 1982.

[2] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33(3):882–886, 1962.

[3] J L Doob. *Stochastic Processes*. John Wiley and Sons, 1953.

[4] Drew Fudenberg and David Levine. Self confirming equilibrium. *Econometrica*, 61:523–545, 1993.

[5] Piotr Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research (JAIR)*, 24:49–79, 2005.

[6] Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. In *AMAI*, Ft. Lauderdale, Florida, 2004.

[7] Frank Hahn. *On the Notion of Equilibrium in Economics: An Inaugural Lecture*. Cambridge University Press, 1973.

[8] Eric Hansen, Daniel Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *AAAI*, 2004.

[9] J S Jordan. Bayesian learning in repeated games. *Games and Economic Behavior*, 9:8–20, 1995.

[10] Joseph Kadane and Patrick Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, 1982.

[11] Ehud Kalai and Ehud Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.

[12] Ehud Kalai and Ehud Lehrer. Subjective equilibrium in repeated games. *Econometrica*, 61(5):1231–1240, 1993.

[6] In fact, infinite horizon strategies computed in practice for I-POMDPs (and POMDPs) are always ϵ-optimal.

[13] John H. Nachbar. Prediction, optimization, and rational learning in repeated games. *Econometrica*, 65:275–309, 1997.

[14] John H. Nachbar and William Zame. Non-computable strategies and discounted repeated games. *Economic Theory*, 8:103–122, 1996.

[15] Yaw Nyarko. Convergence in economic models with bayesian hierarchies of beliefs. *Journal of Economic Theory*, 74:266–296, 1997.