Subjective Equilibria in Interactive POMDPs: Theory and Computational Limitations

Prashant Doshi and Piotr Gmytrasiewicz
Dept. of Computer Science,
University of Illinois at Chicago, IL 60607
{pdoshi,piotr}@cs.uic.edu

Abstract

We analyze the asymptotic behavior of agents engaged in a infinite horizon partially observable stochastic game 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. Imposing an additional assumption of mutual singularity on agents' initial beliefs makes their behavior converge to Nash equilibrium. While theoretically sound, the equilibrating process is difficult to demonstrate computationally because of the difficulty in coming up with initial beliefs that satisfy the truth compatibility condition.

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) [Gmytrasiewicz and Doshi, 2005, Gmytrasiewicz and Doshi, 2004]. 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 [Kadane and Larkey, 1982], differs from the objective representation of POSGs as outlined in [Hansen et al., 2004]. 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 to Bayesian decision theory, the agent maintains and updates its belief about the physical state as well as the strategies of the other agents, and it 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, and hence their own strategies (which are best responses to their beliefs), 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.

In prior work, [Kalai and Lehrer, 1993a, Kalai and Lehrer, 1993b] 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.

Hahn [Hahn, 1973] 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 [Fudenberg and Levine, 1993] consider a general model of finitely repeated extensive form games wherein strategies of opponents may be correlated (unlike [Kalai and Lehrer, 1993a] 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 [Kalai and Lehrer, 1993a] and [Fudenberg and Levine, 1993] 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. Work of a similar vein is reported in [Jordan, 1995]. 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.

The results contained in this paper complement prior results. Specifically, we show 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. Further, we address the research problem posed in [Kalai and Lehrer, 1993a] regarding the existence of subjective equilibrium in POSGs. We also draw a parallel with works in multiagent learning [Hu and Wellman, 1998, Bowling and Veloso, 2002] that show convergence of play to equilibrium. However, our results differ in that we assume that the state and others' actions are partially observable, and the plan is computed offline using a given model of the environment. Finally, we comment on the difficulties in achieving subjective equilibria when a computational perspective is adopted.

The rest of this paper is structured in the following manner. In the next section, we briefly review the Interactive POMDP framework. We focus on the key steps of belief update, and policy computation. In Section 3, we introduce the concept of a subjective equilibrium and theoretically prove that the strategy profile of agents playing a POSG modeled using an I-POMDP, in the limit, is in subjective equilibrium. In Section 4, we remark on the computational infeasibility of arriving at this equilibrium. Finally, we conclude this paper with a discussion of our results and open research issues in Section 5.

2 Overview of Interactive POMDPs

Interactive POMDPs [Gmytrasiewicz and Doshi, 2005, Gmytrasiewicz and Doshi, 2004] 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 $R_i : S \times A \rightarrow \mathbf{R}$. We allow the agent to have preferences over 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:

• *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,

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

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,

[1] We assume computability in the Turing machine sense, i.e. strategies are partial 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 [Gmytrasiewicz and Doshi, 2005, Gmytrasiewicz and Doshi, 2004], 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.

$Pr(\cdot|a_i^{t-1}, a_j^{t-1}, b_i^{t-1})$, 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:

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

where, $ER_i(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|m_j)$ (since $is = (s, m_j)$). Eq. 1 is a basis for value iteration in I-POMDPs. As shown in [Gmytrasiewicz and Doshi, 2005], 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:

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

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.

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 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.

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

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

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

The above result extends immediately to observation histories of any length t. Formally, $E[X_{t+1}|h^t] = X_t$, and from the law of conditional expectations, $E[X_l|h^t] = X_t, \;\; l \geq t$. Therefore the beliefs satisfy the Martingale property.

All Martingales share the following convergence property:

Theorem 1 (Martingale Convergence Theorem (§4 of Chapter 7 in [Doob, 1953]). *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.*

3.2 Subjective Equilibrium

We investigate the asymptotic behavior of agents playing an infinite horizon POSG, in which each agent learns and optimizes. Specifically, each agent starts with a prior belief which is revised on performing an action and receipt of sensory information, followed by computing the strategy which optimizes its beliefs. In the context of I-POMDPs, each agent uses its prior beliefs to index into its policy (computed offline using Equations 1 and 2) resulting in the policy tree that will form its behavior strategy.

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 [Gmytrasiewicz and Doshi, 2005] as a running example throughout this paper. Briefly, the game 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). 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 game 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.

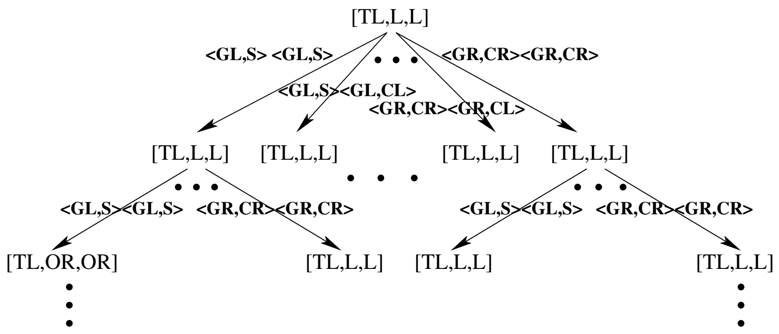

Figure 1: Joint observation histories in the infinite horizon multiagent tiger problem. The nodes represent the state of the game and play of agents, while the edges are labelled with the possible observations. This example starts with the tiger on the left and each agent listening. Each agent may receive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

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]

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 $proj_H\ \mu_0 \ll proj_H\ \mu_k(\cdot|b_k^0)$.*

[5] Following [Nyarko, 1997, Jordan, 1995] 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.

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.

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 updated using the process outlined in Section 2.1, will converge in the limit. Formally,

Theorem 2 (Bayesian Learning in I-POMDPs). *For an agent i in the I-POMDP framework, if its initial belief satisfies the ACC, its posterior beliefs will converge with probability 1.*

Proof. As we mentioned before, Bayesian learning is a Martingale. In Section 3.1, set $\Xi = IS_i$, 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 $X_t = Pr(is^t|h_i^t)$ where h_i^t is agent i's observation history upto time t. We must first show that $E[|X_t|^2]$ is bounded.

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

Theorem 2 now follows from a straightforward application of Theorem 1. □

The above result does not imply that an agent's belief 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.

Example: For an example of observationally equivalent models, consider a version of the multi-agent 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, and then open the door. 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 compatible, the agents' opinions (about the future) will merge and correctly predict the true future in the limit. This result was first noted in [Blackwell and Dubins, 1962]; we present the theorem below and refer the reader to the paper for its proof.

Theorem 3 ([Blackwell and Dubins, 1962]). *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 Theorem 3 to establish predictive convergence in the context of the I-POMDP framework.

Theorem 4 (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,*

$$||proj_{H_t}\ \mu_0(\cdot|h^t) - proj_{H_t}\ \mu_k(\cdot|b_k^0, h_k^t)|| \leq \epsilon$$

for $k = i, j$.

Proof. Referring to Theorem 3, 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 [Blackwell and Dubins, 1962]. 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 $P^t = proj_{H_t}\ \mu_k(\cdot|b_k^0, h_k^t)$. Theorem 4 then follows immediately from a straightforward application of Theorem 3. □

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 [Kalai and Lehrer, 1993a], 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 the main 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. The following corollary gives our result.

Corollary 1 (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. Corollary 1 follows in part from Theorem 4, 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 [Kalai and Lehrer, 1993a].

4 Computational Limitations of Our Results

Recall that in Section 2, we made the assumption that agent strategies are computable. This restricts the space of possible strategies to be countable. However, as observed in [Nachbar and Zame, 1996], there may exist strategies which are exact best responses to computable strategies but are themselves not computable, and even when computable best responses do exist, the decision procedure of computing these best responses may not be computable. Consequences of these negative results lead to a subtle tension between learning and optimization. Specifically, if agents' best response strategies are not computable, then their beliefs fail to account for such strategies of others, thereby possibly violating ACC and preventing predictive convergence. On the other hand, if we posit that best responses be computable, then the corresponding beliefs may be unnatural – for example, they may not assign non-zero probability to all possible strategies of others. Nachbar [Nachbar, 1997] makes an argument along similar lines in the context of repeated games using the notion of a conventional set of strategies (analogous to computable set in our setting) attributed to each agent. We believe that these implausibility issues are a direct implication of Binmore's claim in [Binmore, 1982] that 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. His claim rests on a particular construction of a two-agent game in which a supposedly rational Turing machine when required to compute the best response is unable to predict truthfully, and when required to predict truthfully is unable to terminate its computations and optimize.

Though the above mentioned negative results are existential, they serve to show that it may be problematic to fulfill the assumptions laid out in our analysis in practice. Nevertheless, there may be ways to overcome these limitations. One interesting direction is to replace exact optimization with approximate optimization. Specifically, rather than computing the exact best response to its subjective belief, an agent may compute an ϵ-best response [6] that is guaranteed to be always computable. However, the effect of ϵ-optimality on predictive convergence remains an open question.

5 Discussion

In this paper we analyzed 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. We have also shown how such agents update their beliefs on performing actions and receiving observations, and compute best responses to their beliefs. Within this framework, we 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. For completeness, we also gave conditions on agents prior beliefs that will allow play to converge to Nash equilibrium.

Though the concept of subjective equilibrium is not novel, we believe that our results complement the existing literature. Specifically, using the I-POMDP framework for learning and optimizing, we have shown the existence of equilibria in a POSG. The POSGs provide a more realistic setting than repeated games, in which existence of equilibria was known. We pointed out that attempts to apply these theoretical results could run into obstacles. One problem is the inherent difficulty in perfect optimization and simultaneous prediction. One may be forced to resort to ϵ-optimality. Whether any form of equilibrium obtains when the players are boundedly rational is a topic of future work.

References

[Binmore, 1982] Binmore, K. (1982). *Essays on Foundations of Game Theory*. Pittman.

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

[Bowling and Veloso, 2002] Bowling, M. and Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence Journal*, 136:215–250.

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

[Fudenberg and Levine, 1993] Fudenberg, D. and Levine, D. (1993). Self confirming equilibrium. *Econometrica*, 61:523–545.

[Gmytrasiewicz and Doshi, 2004] Gmytrasiewicz, P. and Doshi, P. (2004). Interactive pomdps: Properties and preliminary results. In *AAMAS*, pages 1374–1375, NYC, NY.

[Gmytrasiewicz and Doshi, 2005] Gmytrasiewicz, P. and Doshi, P. (2005). A framework for sequential planning in multiagent settings. *Journal of AI Research*, 23.

[6] One way to compute an ϵ-best response is to consider finite horizons for maximization, rather than infinity.

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

[Hansen et al., 2004] Hansen, E., Bernstein, D., and Zilberstein, S. (2004). Dynamic programming for partially observable stochastic games. In *AAAI*.

[Hu and Wellman, 1998] Hu, J. and Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *15th Intl Conference on Machine Learning*, pages 242–250.

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

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

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

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

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

[Nachbar and Zame, 1996] Nachbar, J. H. and Zame, W. (1996). Non-computable strategies and discounted repeated games. *Economic Theory*, 8:103–122.

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