Empirical-evidence Equilibria in Stochastic Games

Nicolas Dudebout

October 2012

Contents

Nomenclature v

1 Introduction 1

2 Origin and History of the Problem 3
2.1 Decentralized Control through Learning in Stochastic Games 3
2.2 Decision Making 4
2.2.1 Formulation 4
2.2.2 Dynamic Problems 4
2.2.3 Learning a Strategy 6
2.2.4 Imperfect-information Problems 6
2.3 Game Theory 6
2.3.1 Games and Equilibria 6
2.3.2 A Game Example 7
2.3.3 Are Nash Equilibria the Solution? 8
2.3.4 Correlated Equilibria 9
2.4 Learning in Games 10
2.4.1 Correlated Equilibria 11
2.4.2 Weakly Acyclic Games 11
2.4.3 Stochastically Stable States 11
2.5 Stochastic Games 11
2.5.1 Multiagent Reinforcement Learning 12
2.5.2 Subjective and Self-confirming Equilibria 12
2.5.3 Weakly Belief-free Equilibria 13
2.5.4 Analogy-based Expectation Equilibria 13
2.6 Bounded Rationality 14
2.6.1 Linear Modeling 15
2.6.2 Mean-field Games 15
2.6.3 Incomplete Theories 15
2.6.4 Egocentric Modeling 15

3 Preliminary Research 17
3.1 Empirical-evidence Equilibria 17
3.1.1 Single-agent Setup 17

3.1.2 Depth-k Consistency . . . 18
3.1.3 Empirical-evidence Optimality . . . 19
3.1.4 Multiagent Setup . . . 21
3.2 Existence of Empirical-evidence Equilibria . . . 22
3.3 Learning Empirical-evidence Equilibria . . . 23
3.3.1 A Learning Rule . . . 23
3.3.2 Simulation Results . . . 24

4 Proposed Research **27**
4.1 Open Questions . . . 27
4.1.1 Implications of Using Consistency . . . 27
4.1.2 Large Number of Agents . . . 28
4.1.3 Learning . . . 28
4.1.4 Price of Anarchy . . . 28
4.1.5 Payoff Folk Theorem . . . 29
4.2 Proposed Work . . . 29

Bibliography **31**

Nomenclature

Symbols

t denotes a discrete time step. b^t is the value of variable b at time t. When unambiguous, b and b^+ are short notations for b^t and b^{t+1} respectively.

$\Delta(\mathcal{B})$ is the set of distributions over finite set $\mathcal{B}$. $b \sim \beta$ denotes that b is drawn according to distribution β. $\mathbb{P}_\beta[B]$ is the probability of event B under distribution β. $\beta[e]$ denotes the quantity $\mathbb{P}_\beta[b = e]$, for $b \sim \beta$. $\mathbb{E}_\beta[b]$ is the expected value of b under distribution β.

$\mathcal{I}$ is a set of agents and i denotes one agent. $-i$ represents the set of all agents excluding agent i, i.e., $\mathcal{I} \setminus \{i\}$. If $\mathcal{B}_i$ is a set associated with agent i, $\mathcal{B}$ denotes the Cartesian product $\prod_{i \in \mathcal{I}} \mathcal{B}_i$. If b_i is a variable associated with agent i, b denotes the tuple $\big(b_1, b_2, \ldots, b_{|\mathcal{I}|}\big)$.

Acronyms

ABEE analogy-based expectation equilibrium 13, 14

EEE empirical-evidence equilibrium 1, 21–25, 27–29

EEO empirical-evidence optimum 19, 21

MDP Markov decision process 1, 3–6, 11, 15, 18–20, 22, 23, 27

MFE mean-field equilibrium 15, 28

POMDP partially observable Markov decision process 1, 3, 4, 6, 12, 17–19, 27

1 Introduction

The objective of the proposed research is to develop the framework of empirical evidence equilibria (EEEs) in stochastic games and to design decentralized controllers using learning in that framework. The goal is to enable a set of agents to control a dynamical system in a decentralized fashion. To do so, the agents play a stochastic game crafted such that its equilibria are decentralized controllers for the dynamical system. Unfortunately, there exists no algorithm to compute equilibria in stochastic games. One explanation for this lack of results is the full-rationality requirement of game theory. In the case of stochastic games, full rationality imposes that two requirements be met at equilibrium. First, each agent has a perfect model of the game and of its opponents' strategies. Second, each agent plays an optimal strategy for the partially observable Markov decision process (POMDP) induced by its opponents' strategies. Both requirements are unrealistic. An agent cannot know the strategies of its opponents; it can only observe the combined effect of its own strategy interacting with its opponents'. Furthermore, POMDPs are intractable; an agent cannot compute an optimal strategy in a reasonable time. In addition to these two requirements, engineered agents cannot carry perfect analytical reasoning and have limited memory; they naturally exhibit bounded rationality. In the proposed research, bounded rationality is not seen as a limitation and is instead used to relax the two requirements. In the EEE framework, agents formulate low-order empirical models of observed quantities called mockups. Mockups have unmodeled states and dynamic effects, but they are statistically consistent; the empirical evidence observed by an agent does not contradict its mockup. Each agent uses its mockup to derive an optimal strategy. Since agents are interconnected through the system, these mockups are sensitive to the specific strategies employed by other agents. In an EEE, the two requirements are weakened. First, each agent has a consistent mockup of the game and the strategies of its opponents. Second, each agent plays an optimal strategy for the Markov decision process (MDP) induced by its mockups. The EEE framework provides a new solution concept for stochastic games. This solution concept has been contrasted to other classical notions. General conditions for the existence of EEEs have been derived. The properties of EEEs will be investigated. The EEE framework will be used to solve decentralized control problems through the design of an algorithm converging to EEEs.

2 Origin and History of the Problem

2.1 Decentralized Control through Learning in Stochastic Games

A complex system is a set of agents connected through a network. The subsystems of a car, a robotic plant, and the power grid are examples of complex systems at different scales. The advances in information technology made these complex systems ubiquitous, and tools to control them are needed. These systems can be controlled in a centralized fashion. However, a centralized controller represents a single point of failure, does not scale to large networks, and incurs high communication costs. Adaptive decentralized controllers address these problems. A controller is decentralized if each agent in the system makes some decisions. Decentralization renders the system more robust by not having a single point of failure. A controller is adaptive if each agent is doing simple computations using local information. Adaptivity mitigates the scalability and communication issues.

In optimal control, centralized controllers are the optima of a function. Unfortunately, in the multiagent setting, the notion of optimality is ill defined. Game theory, the study of interacting decision makers, addresses this issue by replacing optima with equilibria. An equilibrium is a joint decision satisfying all the agents at once; at equilibrium, no agent has an incentive to unilaterally deviate. In a game-theoretic approach, decentralized controllers are equilibria of a game.

Equilibria can be computed by a centralized algorithm. However, this centralized approach brings back the issue of scalability and prevents the addition of new agents without designing a new controller. Game-theoretic learning enables the decentralized computation of equilibria. Each agent modifies its strategy according to a learning rule using local information. The learning rules used by the agents are chosen to guarantee convergence to an equilibrium. Game-theoretic learning is an adaptive decentralized approach to designing adaptive decentralized controllers.

Engineering problems often involve dynamical systems with a state, such as MDPs. When the decision maker cannot observe the state directly it is facing a POMDP. Solving an MDP is tractable for reasonable sizes of the state space, whereas solving a POMDP is intractable. Stochastic games extend these processes to the multiagent case. In a complex system, agents only observe local information. Therefore, the games used to control these systems are stochastic games of imperfect information. These games are, like POMDPs, intractable. To this day, there exists no centralized algorithm nor learning rule for computing equilibria in stochastic games.

The full-rationality requirement of game theory is in part to blame for this lack of results. Full rationality requires agents to have perfect understanding of the game being played. This requirement is not realistic for engineered agents which have, by nature, bounded rationality. The proposed research uses bounded rationality to make each agent face an MDP instead of a POMDP.

The rest of this chapter introduces the formal setting of learning in stochastic games along with relevant previous work.

2.2 Decision Making

2.2.1 Formulation

Decision making is the rational process of finding the best action given the information available. An agent is given a set of actions $\mathcal{A}$ and preferences over these actions. Preferences are expressed by a utility function $u\colon \mathcal{A} \to \mathbb{R}$, such that for two actions a and a' in $\mathcal{A}$, the following two properties hold:

- The agent prefers a over a' if and only if $u(a) > u(a')$.
- The agent is indifferent between a and a' if and only if $u(a) = u(a')$.

The utility of an action can be interpreted as a payoff that the agent wants to maximize. The agent can make nondeterministic decisions. Instead of committing to a specific action, it can choose a mixed action. A mixed action α is a distribution over the action set. A mixed action's payoff is the expected value of the payoffs of the actions in its support. For example, choosing a with probability $\frac{1}{3}$ and a' with probability $\frac{2}{3}$ yields a payoff $\frac{1}{3}u(a) + \frac{2}{3}u(a')$. The domain of the utility function is usually extended from the action set $\mathcal{A}$ to distributions over the action set $\Delta(\mathcal{A})$. For an element α in $\Delta(\mathcal{A})$, $u(\alpha) = \mathbb{E}_{a\sim\alpha}[u(a)]$. Solving a decision-making problem is equivalent to solving a stochastic optimization problem.

Note 1 (Von Neumann–Morgenstern utility theorem)**.** *The representation of preferences by utility functions was characterized by von Neumann and Morgenstern [1]. They proved that rational preferences can always be represented by a utility function to be maximized in expectation and that the utility function is unique up to a positive affine transformation. Preferences are rational if they satisfy four axioms: completeness, transitivity, continuity, and independence. Human decision makers might not verify these axioms, but engineered agents can always be designed to verify them. Insuring the validity of these axioms is therefore not a concern for this research.*

2.2.2 Dynamic Problems

Problems in controls are dynamic. The simplest dynamic problems are MDPs. In an MDP, a state evolves in discrete time controlled by an action. The state at time $t+1$ is a random

variable depending only on the state of the system at time t and the action played at time t. This dynamic is captured by the short notation

$$x^+ \sim f(x, a), \tag{1}$$

where x and x^+ are states in a finite state space $\mathcal{X}$ and a is an action in a finite action set $\mathcal{A}$. At each time step, the agent observes the state and chooses an action. A history is a sequence of states and actions. The history up to time t is $h^t = (x^0, x^1, \ldots, x^t, a^0, a^1, \ldots, a^t)$. In state x, choosing action a yields a payoff $u(x, a)$. The agent is interested in maximizing its expected sum of discounted payoffs. For a given infinite history h^t, the agent receives a sum of discounted payoffs $\sum_{t=0}^{\infty} \delta^t u(x^t, a^t)$, where $\delta \in [0, 1)$ is the discount factor. Note that δ^t denotes δ to the power t whereas x^t and a^t denote the state and the action at time t.

In a static decision-making problem, the agent has to choose one action. In an MDP, the agent has to choose a strategy which is a plan of action for all the possible outcomes of the process. A strategy σ determines an action a^t depending on (h^{t-1}, x^t), the information available at time t. The domain of a strategy is infinite; therefore, the set of strategies Σ is infinite. An agent using strategy σ with initial state x receives an expected sum of discounted payoffs

$$U_\sigma(x) = \mathbb{E}_\sigma \left[\sum_{t=0}^{\infty} \delta^t u\big(x^t, a^t\big) \,\middle|\, x^0 = x \right]. \tag{2}$$

A solution to the MDP is an element of $\cap_{x \in \mathcal{X}} \arg\max_{\sigma \in \Sigma} U_\sigma(x)$. It is not obvious that the maximum is attainable or that the intersection is not empty. Furthermore, since Σ is infinite, looking for a solution with an exhaustive-search method is in vain. However, by using the Markovian structure of the problem these issues can be bypassed. The main theorem in the MDP literature states that for any finite MDP, there exists a stationary deterministic optimal strategy [2, 3]. A strategy is stationary if the next action is computed using only the current state; the history leading to the current state and the time are not used. A strategy is deterministic if the actions selected are not mixed.

This result reduces the set of strategies to be considered to a finite number. However, solving (2) for each of the $|\mathcal{A}|^{|\mathcal{X}|}$ strategies is prohibitively expensive. The structure of the problem can be used to explore the solution space more efficiently.

A central concept in the analysis of MDPs is the value function $U^* \colon \mathcal{X} \to \mathbb{R}$ of the problem. The utility received by using an optimal strategy from the initial state x is $U^*(x)$. The value function is characterized by the Bellman equation

$$U^*(x) = \max_{a \in \mathcal{A}} \big\{ u(x, a) + \delta \mathbb{E}\big[U^*\big(x^+\big) \,\big|\, x, a \big] \big\}. \tag{3}$$

Dynamic-programming algorithms search the solution space by using the recursive structure of the Bellman equation. These algorithms are more efficient than exhaustive-search algorithms but are under the curse of dimensionality; the amount of computations required grows exponentially with the size of the state space. The two main dynamic-programming algorithms are value iteration and policy iteration.

2.2.3 Learning a Strategy

When the dynamic of the system is not known but can be easily simulated, reinforcement-learning algorithms can be used [4, 5]. A reinforcement-learning algorithm learns the transition probabilities while using its current optimal strategy. As the algorithm accumulates information, it computes better strategies. Reinforcement-learning algorithms work by balancing exploration and exploitation. Exploration refers to learning the transition probabilities, and exploitation refers to using a strategy maximizing the expected sum of discounted payoffs. Dynamic programming is an offline approach, whereas reinforcement learning is an online approach.

Dynamic-programming algorithms compute the value function $U^*\colon \mathcal{X} \to \mathbb{R}$. Reinforcement-learning algorithms compute the action value function $Q\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined by

$$Q(x,a) = u(x,a) + \delta \mathbb{E}\left[U^*\left(x^+\right) \,\middle|\, x, a\right]. \tag{4}$$

SARSA and Q-learning are reinforcement-learning versions of policy iteration and value iteration respectively.

2.2.4 Imperfect-information Problems

POMDPs model situations where the agent is uncertain about the state of the dynamical system. In a POMDP, the state evolves according to (1). However, at each time step, the agent cannot observe the state and can only observe a signal

$$y \sim g(x). \tag{5}$$

This small change in the information available to the agent has big consequences: POMDPs are intractable.

In an MDP, the state is the only necessary information needed to compute the next action of an optimal strategy. In a POMDP, the agent does not know the state and needs to use beliefs to implement an optimal strategy. Beliefs are probability distributions over sequences of states computed using the signals observed and Bayes' inference. An optimal solution for a POMDP is a function from the belief space $\bigcup_{t=0}^{\infty} \Delta(\mathcal{X}^t)$ to the action set. The fact that the belief space is continuous is what makes the problem intractable.

2.3 Game Theory

2.3.1 Games and Equilibria

In a game setting, a set of agents $\mathcal{I}$ faces decision-making problems. Each agent i in $\mathcal{I}$ has an action set $\mathcal{A}_i$ and a utility function $u_i\colon \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} = \prod_{i \in \mathcal{I}} \mathcal{A}_i$ is called the joint action set. Note that this utility function depends on the actions of all the agents. As mentioned earlier, decision making is the rational process of finding an optimal action

given the information available. There is no obvious way to extend that definition to the multiagent setting. Preferences of different agents cannot be aggregated; therefore, the notion of optimality for the set of agents is ill defined.

Optimality for an individual agent is still well defined. Denote the opponents of agent i by $-i = \mathcal{I} \setminus \{i\}$. For fixed actions of its opponents, agent i faces a decision-making problem. The actions in $\mathcal{A}_i$ that are optimal for the fixed actions of $-i$ are called best responses. For $a_{-i} = \big(a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{|\mathcal{I}|}\big)$, the best-response set of agent i is $\mathrm{BR}_i(a_{-i}) = \arg\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i})$. Note that $\mathrm{BR}_i \colon \mathcal{A}_{-i} \rightrightarrows \mathcal{A}_i$ is a correspondence and not a function.

A joint action that is simultaneously a best response for all the agents is a good candidate to replace optimality in the multiagent setting. This concept, at the core of game theory, is called a Nash equilibrium. A joint action a^* in $\mathcal{A}$ is a Nash equilibrium if

$$\forall i \in \mathcal{I},\ a_i^* \in \mathrm{BR}_i\big(a_{-i}^*\big). \tag{6}$$

The definitions of best response and Nash equilibrium extend readily to mixed actions by replacing as by αs. Note that a deterministic action is sometimes called a pure action. Nash proved that any game with a finite number of players having a finite number of actions has at least one, potentially mixed, Nash equilibrium [6].

2.3.2 A Game Example

Game-theoretic concepts are illustrated on the the following game known as battle of the sexes. A couple, composed of a man m and a woman w, is planning a date. Each one chooses between two actions: going to a football match F or going to an opera performance O. The joint action of the couple is represented by an ordered pair $(a_\mathrm{m}, a_\mathrm{w})$, where a_m is the action chosen by the man and a_w by the woman. For example, (F, O) denotes that he chooses football and she chooses opera.

The man prefers to be with the woman rather than separated from her. If they are together, he prefers football $(\mathbf{F}, \mathrm{F})$ to opera $(\mathbf{O}, \mathrm{O})$. If they are not together, he is indifferent between football $(\mathbf{F}, \mathrm{O})$ and opera $(\mathbf{O}, \mathrm{F})$. The woman prefers to be with the man rather than separated from him. If they are together, she prefers opera $(\mathrm{O}, \mathbf{O})$ to football $(\mathrm{F}, \mathbf{F})$. If they are not together, she still prefers opera $(\mathrm{F}, \mathbf{O})$ to football $(\mathrm{O}, \mathbf{F})$. Their preferences can be implemented by utility functions u_m for the man and u_w for the woman with the following values:

$$\begin{aligned} &u_\mathrm{m}(\mathrm{F},\mathrm{F}) = 2, \quad u_\mathrm{m}(\mathrm{O},\mathrm{O}) = 1, \quad u_\mathrm{m}(\mathrm{F},\mathrm{O}) = 0, \quad u_\mathrm{m}(\mathrm{O},\mathrm{F}) = 0, \\ &u_\mathrm{w}(\mathrm{O},\mathrm{O}) = 3, \quad u_\mathrm{w}(\mathrm{F},\mathrm{F}) = 2, \quad u_\mathrm{w}(\mathrm{F},\mathrm{O}) = 1, \quad u_\mathrm{w}(\mathrm{O},\mathrm{F}) = 0. \end{aligned} \tag{7}$$

The action sets and the utility functions of battle of the sexes are represented in a compact form as follows:

	F	O
F	2, 2	0, 1
O	0, 0	1, 3

. (8)

The man's action determines the row and the woman's determines the column. Numbers in the cell are the utilities received: the first by the man and the second by the woman. This is called the normal-form representation of the game.

To compute the best response of the man, fix the action α_{w} of the woman. With α_{w}, she chooses F with probability p_{w} and O with probability $(1 - p_{\mathrm{w}})$. Note that she chooses a pure action for $p_{\mathrm{w}} = 0$ or $p_{\mathrm{w}} = 1$. The utility received by the man, if he plays F, is $u_{\mathrm{m}}(\mathrm{F}, \alpha_{\mathrm{w}}) = p_{\mathrm{w}} u_{\mathrm{m}}(\mathrm{F}, \mathrm{F}) + (1 - p_{\mathrm{w}}) u_{\mathrm{m}}(\mathrm{F}, \mathrm{O}) = 2p_{\mathrm{w}}$. If he plays O, his utility is $u_{\mathrm{m}}(\mathrm{O}, \alpha_{\mathrm{w}}) = p_{\mathrm{w}} u_{\mathrm{m}}(\mathrm{O}, \mathrm{F}) + (1 - p_{\mathrm{w}}) u_{\mathrm{m}}(\mathrm{O}, \mathrm{O}) = 1 - p_{\mathrm{w}}$. His optimal action depends on p_{w} with a critical value of $\frac{1}{3}$. If $p_{\mathrm{w}} > \frac{1}{3}$, he prefers F to O. If $p_{\mathrm{w}} < \frac{1}{3}$, he prefers O to F. If $p_{\mathrm{w}} = \frac{1}{3}$, he is indifferent between F and O; any combination of F and O is a best response.

The best response of the woman is computed in a similar fashion. The critical value of p_{m} making her indifferent is $\frac{3}{4}$.

The best responses are plotted in Figure 1. The intersections of the graphs correspond to the Nash equilibria of the game. Battle of the sexes has three Nash equilibria: two pure ones and one mixed. The pure Nash equilibria arise from the man and the woman choosing the same event. The mixed one corresponds to the man and the woman independently randomizing their choices with probabilities $p_{\mathrm{m}} = \frac{3}{4}$ and $p_{\mathrm{w}} = \frac{1}{3}$.

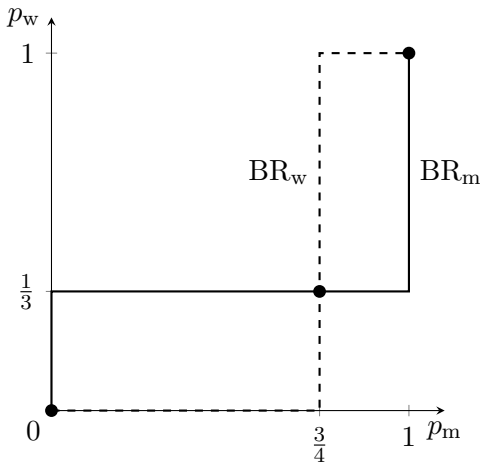

Figure 1. Best responses and Nash equilibria for battle of the sexes. The man plays F with probability p_{m}. The woman plays F with probability p_{w}. The solid line is the man's best response. The dashed line is the woman's best response. The filled circles indicate the Nash equilibria.

2.3.3 Are Nash Equilibria the Solution?

Nash equilibria form a solution concept for multiagent decision problems. Nash equilibria are self-enforcing agreements; if you give actions forming a Nash equilibrium to some agents, none of them has an incentive to unilaterally deviate. This self-enforcing property makes

Nash equilibria relevant for decentralized control. Nash equilibria cannot, however, be called **the** solution to multiagent decision problems. They have weaknesses made clear by the following questions:

- When a game has multiple equilibria, how do agents coordinate to choose one equilibrium?
- Different utility functions can represent the same preferences. However, changing the utility functions changes the probabilities in a mixed Nash equilibrium. Therefore, what is the intrinsic meaning of a mixed Nash equilibrium?
- The social welfare is the sum of the utilities of the agents. Should the social welfare be maximized instead of finding an equilibrium?
- A Nash equilibrium guarantees no profitable unilateral deviation. What happens when two agents deviate simultaneously?

Authors have been debating for decades regarding which equilibrium is the right solution concept. This debate is not relevant for the proposed research. Any solution concept can be used as long as its limitations are understood.

2.3.4 Correlated Equilibria

The mixed Nash equilibrium in battle of the sexes seems fairer than the two pure ones: each agent has a chance to go on his or her preferred date. However, the two agents end up in different locations with positive probability: some utility is wasted. In the mixed Nash equilibrium, the expected utility for the man is $\frac{2}{3}$ and for the woman $\frac{3}{2}$.

When facing these kinds of incompatible decisions, humans sometimes have recourse to a coin toss. The agents agree that on heads they go to the football match and on tails they go to the opera performance. This is not a Nash equilibrium. However, once the outcome of the coin toss is known, no agent has an incentive to unilaterally deviate. This is another form of equilibrium, introduced by Aumann under the name correlated equilibrium [7]. The correlated equilibrium maintains the fairness of the mixed Nash equilibrium but does not waste utility. In the correlated equilibrium, the expected utility for the man is $\frac{3}{2}$ and for the woman $\frac{5}{2}$. Note that every Nash equilibrium is a correlated equilibrium but the converse is not true.

Two equivalent definitions of a correlated equilibrium exist. The first one describes the equilibrium in terms of the common random source and the actions taken given each possible outcome; the second one, in terms of the resulting distribution over joint actions. The first definition is more intuitive whereas the second one is easier to work with. As a result, we give the second definition. A probability distribution over joint actions γ is a correlated equilibrium if

$$\forall i \in \mathcal{I},\, a_i, a_i' \in \mathcal{A}_i, \quad \sum_{a_{-i} \in \mathcal{A}_{-i}} \gamma[a](u_i(a) - u_i(a_i', a_{-i})) \geq 0. \tag{9}$$

2.4 Learning in Games

As previously mentioned, Nash equilibria are self-enforcing agreements. Learning studies the question of how agents reach such an agreement. Learning in the economics literature tries to explain behaviors observed in experiments; the authors look for simple rules human decision makers likely use. The ensuing debate concerning the validity of learning algorithms for human decision makers is irrelevant for the proposed research.

In the learning framework, a game is played repeatedly at discrete time steps. Agents use strategies to choose their actions. At a given time step, an agent plays an action and receives a signal. This signal is most of the time the joint action. The agent then updates its strategy depending on the received signal. The update rule is called a learning algorithm. The goal is to define learning algorithms making the joint action converge to a Nash equilibrium [8].

A learning algorithm is composed of the three following components:

- Information accumulation
- Optimization of a function constructed from that information
- Randomization to avoid being trapped in local optima

Randomization commonly takes the form of smoothing; instead of playing a best response, an agent plays a mixed action favoring the best response and putting a small probability on other actions. A learning algorithm is called adaptive if the information is accumulated locally and the optimization is an easy computational task. In economics, the easiness of a computational task is defined with human decision makers in mind. In the proposed research, the easiness is defined for an engineered decision maker; for example, computing the eigenvectors of a medium size matrix is considered an easy computational task. Adaptivity is an important characteristic of learning algorithms for scaling.

Fictitious play is an example of an adaptive learning algorithm. In fictitious play, agents keep track of the empirical frequencies of the actions played by their opponents. At each time step, an agent plays a best response to the mixed action induced by these empirical frequencies of play. Information is accumulated through the empirical frequencies. Optimization takes the form of playing a best response. Smooth fictitious play is a variant incorporating the randomization component.

Unfortunately, fictitious play does not always converge to a Nash equilibrium [9]. In fact, no adaptive learning rule converges to Nash equilibria for all games [10]. This result is in part due to the fact that computing a Nash equilibrium is PPAD complete [11]; the complexity of finding a Nash equilibrium is exponential in the number of actions.

Three approaches to designing simple convergent algorithms are presented below. One considers correlated equilibria with a weaker notion of convergence, another focuses on the class of weakly-acyclic games, and the last one uses the less constraining solution concept of stochastically stable states.

2.4.1 Correlated Equilibria

Hart and Mas-Colell proved that a family of adaptive learning rules converge to the set of correlated equilibria [12–14]. These algorithms rely on the notion of regret. A regret measures the payoff difference between two actions. Formally, the regret for playing a instead of a' is the average increase in payoff the agent would have received, had it replaced every play of a by a'. The optimization step seeks to minimize the regrets. As a result, the family of algorithms is called no regret. The guaranteed convergence of these algorithms comes not only from the simpler equilibrium concept but also from the use of a looser notion of convergence on a different quantity. Indeed, no-regret algorithms guarantee the convergence of the empirical distribution of play to the set of correlated equilibria. Note that the empirical distribution of play is different from the joint action and that convergence to a set is less constraining than convergence to a point.

2.4.2 Weakly Acyclic Games

A game is weakly acyclic if from any joint action there exists a better-reply path ending at some pure Nash equilibrium. This structure on the utility functions, introduced by Young [15], insures that better-reply learning algorithms converge to a Nash equilibrium in weakly acyclic games [16]. Weakly acyclic games are an extension of potential games, a class of games used to model congestion problems and to systematically design decentralized controllers [17].

2.4.3 Stochastically Stable States

Young introduced the notion of stochastically stable states to characterize the long-run behavior of a system subject to a diminishing random noise [18]. A state is stochastically stable if it is visited infinitely often as the noise fades. Learning in this context is different from learning an equilibrium. Agents should, as a whole, make the noise fade in a way guaranteeing that the stochastically stable states of the system are the desirable ones. This notion of stability was used to control wind farms [19], to characterize the yield of self-assembly mechanisms [20], and to study language evolution [21].

2.5 Stochastic Games

Stochastic games are the extension of MDPs to the multiagent setting. The utility functions of the agents depend on a state whose dynamic is impacted by the joint actions. In other terms, for each state, the agents play a different game. Their actions impact the payoffs and the transition probabilities between states. In a stochastic game, the agents want to maximize the expected sum of their discounted payoffs. Repeated games are a subset of stochastic games with no state; the same stage game is played at each time step [22]. Note that repeated play of a game, for example as used in learning, differs from a repeated game;

in the repeated play of a game, the quantity to maximize is not the sum of discounted payoffs. Requiring adaptive agents implies that they use local, thus imperfect information. For fixed strategies of its opponents, an agent is facing a POMDP. Therefore, there are no results to compute equilibrium strategies nor learning algorithms for the stochastic games of imperfect information.

In a stochastic game, a state evolves in discrete time controlled by the joint action of a set of agents $\mathcal{I}$. The state at time $t+1$ is a random variable depending only on the state of the system at time t and the joint action played at time t. This dynamic is captured by the short notation

$$x^+ \sim f(x, a), \tag{10}$$

where x and x^+ are states in a finite state space $\mathcal{X}$ and $a = \left(a_1, \ldots, a_{|\mathcal{I}|}\right)$ is a joint action in the finite joint action set $\mathcal{A} = \prod_{i \in \mathcal{I}} \mathcal{A}_i$. When the game is of imperfect information, at each time step, agent i observes a signal $y_i \sim g_i(x)$. The private history of agent i up to time t is $h_i^t = ({y_i}^0, {y_i}^1, \ldots, y_i^t, {a_i}^0, {a_i}^1, \ldots, a_i^t)$. A strategy for agent i is a mapping from its private histories to the distribution over its actions, $\sigma_i \colon \mathcal{H}_i \to \Delta(\mathcal{A}_i)$. In state x, joint action a yields a payoff $u_i(x, a)$ to agent i. Each agent is interested in maximizing its expected sum of discounted payoffs.

Extensions of reinforcement learning to the multiagent setting and three equilibrium concepts lowering the requirements on the beliefs are presented below.

2.5.1 Multiagent Reinforcement Learning

Hu and Wellman attempted to apply results from reinforcement learning to the multiagent setting with the Nash–Q-learning algorithm [23]. In stochastic games, it is unfortunately not enough to balance exploration and exploitation. The Nash–Q-learning algorithm requires the agents to keep track of action-value functions for their opponents and to play Nash-equilibrium strategies. This approach is computationally expensive and only yields results for agents with identical or opposite utility functions.

2.5.2 Subjective and Self-confirming Equilibria

Subjective equilibria, introduced by Kalai and Lehrer, lower the requirements on the beliefs [24]. They only require the beliefs to be correct on the path of play. Self-confirming equilibria, introduced by Fudenberg and Levine, are a closely related concept [25]. In a self-confirming equilibrium an agent can hold the false belief that its opponents correlate their actions off the path of play. Agents playing a subjective or self-confirming equilibrium never see plays contradicting their beliefs.

Subjective and self-confirming equilibria are formally defined in terms of belief strategies. Belief strategy $\tilde{\sigma}_j^i \colon \mathcal{H}_j \to \Delta(\mathcal{A}_j)$ is the strategy agent i believes agent j is playing. Agent i's belief is composed of one belief strategy for each agent $\tilde{\sigma}^i = \left(\tilde{\sigma}_1^i, \ldots, \tilde{\sigma}_{|\mathcal{I}|}^i\right)$. In particular, its belief strategy for itself is its actual strategy, $\tilde{\sigma}_i^i = \sigma_i$.

A set of $|\mathcal{I}|$ strategies, one per agent, induces a distribution over the possible histories. The histories having a positive probability of being visited are called the path of play. This set of strategies can be the actual strategies or the beliefs of one agent. Note that a distribution over beliefs also induces a distribution over the possible histories.

Strategies σ and beliefs $\tilde{\sigma}$ form a subjective equilibrium when the following two conditions hold for each agent i:

- Strategy σ_i is a best response to the belief strategies $\tilde{\sigma}^i_{-i}$.
- Strategies σ and strategies $\tilde{\sigma}^i$ induce the same distribution over the path of play.

Strategies σ and distributions over beliefs $\tilde{\nu}$ form a self-confirming equilibrium when the following two conditions hold for each agent i:

- Strategy σ_i is a best response to the distribution over belief strategies $\tilde{\nu}^i_{-i}$.
- Strategies σ and the distribution over belief $\tilde{\nu}^i$ induce the same distribution over the path of play.

These two equilibrium concepts loosens the requirements of full rationality. Agents can be mistaken about events that will never happen. However, these concepts require each agent to be aware of the existence of every other agent. An agent needs to understand what its opponents actions and signals are to build belief strategies. It also needs to know the exact impact of its opponents actions to verify the optimality of its own strategy. Therefore, these two equilibrium concepts are only a first step towards the goal of the proposed research.

2.5.3 Weakly Belief-free Equilibria

In repeated games, the notion of Nash equilibrium is not satisfactory. Some Nash equilibria exhibit noncredible threats; some strategies are best response to each other on the path of play but not off of it. Strategies using only credible threats are called sequentially rational or subgame perfect. Payoff folk theorems establish that the set of payoffs achievable at equilibrium with sequentially rational strategies is the set of feasible individually strictly rational payoffs. These theorems rely on the recursive structure of optimal solutions. Sugaya recently derived the proof of this result for games of private imperfect information [26]. The previous best result for that class of games was established by Kandori [27]. He characterized weakly belief-free equilibria, a subset of the sequentially rational equilibria exhibiting a recursive structure. In a weakly belief-free equilibrium, agents only need to have correct beliefs about the last action played by their opponents.

2.5.4 Analogy-based Expectation Equilibria

Jehiel introduced the concept of analogy-based expectation equilibria (ABEEs) for games of perfect information to keep the belief space size constant [28]. ABEEs can be expressed

in terms of belief strategies Each agent partitions the history set in a finite number of analogy classes. An analogy class for agent i is denoted by κ_i and the set of analogy classes by $\mathcal{K}_i$. Each agent i believes that its opponents' actions are fully determined by the analogy class; for two histories h and h' in the same analogy class κ_i and for all agent j, $\tilde{\sigma}^i_j(h) = \tilde{\sigma}^i_j(h') = \alpha^{i,\kappa_i}_j$. The, potentially mixed, action α^{i,κ_i}_j is called an analogy-based belief.

Strategies σ, analogy classes $\mathcal{K}$, and analogy-based beliefs α form an ABEE when the following two conditions hold for each agent i:

- Strategy σ_i is a sequentially rational best response to the analogy-based beliefs α^i_{-i}.
- For all agent j, the analogy-based belief α^i_j is consistent with σ_j, i.e., for all κ_i in $\mathcal{K}_i$ and a_j in $\mathcal{A}_j$, $\alpha^{i,\kappa_i}_j[a_j] = \mathbb{P}[\sigma_j(h) = a_j \,|\, h \in \kappa_i]$.

The ABEE concept is a substantial step in the direction of the proposed research. The perfect understanding required by full rationality is replaced by the notion of consistency. Beliefs are consistent if they are accurate on average even though they might be inexact upon closer inspection. This relaxation simplifies the problem that each agent is facing. However, each agent is still required to have a good understanding about the game being played and the role of its opponents. The proposed research goes beyond this limitation by using consistency in a setup where agents do not need to know they are playing a game. The following section exposes other approaches using consistency.

2.6 Bounded Rationality

In classical game theory, agents are assumed to be fully rational. Bounded rationality studies scenarii where agents have limited computation power or make mistakes [29]. In the economics literature, bounded rationality is used to take into account human nature and to explain discrepancies with experiments. Fully rational agents can perfectly use any knowledge they have about the problem they face. For example, in a stochastic game of imperfect information, fully rational agents propagate beliefs accurately. Propagating beliefs means doing Bayesian inference on a belief space whose size increases with time. Engineered agents have limited computation power, limited memory, and bounded precision. Furthermore, adaptivity requires the use of local and therefore incomplete information. As a result, there is no hope to build fully rational adaptive agents in a dynamic world. In the proposed research, the bounded rationality of engineered agents is used as an advantage. Instead of relying on propagation of beliefs regarding the imperfect information, simple consistent models are used. A model is consistent if the agent does not observe evidence contradicting it.

Four approaches using bounded rationality to lower the complexity of the problem are presented below. The first one uses Kalman filtering to update a model while the others use the notion of consistency. All the consistency approaches use exogenous models, whereas the proposed research lifts that restriction. Other differences with the proposed research are highlighted.

2.6.1 Linear Modeling

Chang, Ho and Kaelbling used modeling to simplify multiagent learning [30]. Each agent assumes that the signal received is generated from a linear system and uses Kalman filtering to get the best estimate of the current state.

2.6.2 Mean-field Games

Lasry and Lions studied a setting where a very large number of agents faces identical copies of an MDP [31]. The MDPs are coupled through a common signal received by the agents. This signal is the proportion of agents in each state; it is a stochastic process impacted by the strategies of all the agents. Agents compute their optimal strategies by considering a consistent, exogenous and stationary model of the signal. Agents are in a mean-field equilibrium (MFE) if their optimal strategies induce precisely this stationary signal. The goal of MFEs is to simplify the analysis of games with a very large number of agents. The main result in the MFE literature is that when the number of agents goes to infinity, MFEs coincide with Nash equilibria. The fact that the signal is not truly stationary nor exogenous washes away when the number of agents is large. MFEs aim at simplifying the analysis of Nash equilibria for a specific game with a large number of players. The proposed research aims for a different equilibrium concept in general games with any number of players. Furthermore, MFEs focuses on stationary models, whereas the proposed research considers more elaborate models. Weintraub, Benkard, and van Roy applied the mean-field methodology to approximate subgame-perfect equilibria in a problem of dynamic imperfect competition [32]. They named their equilibrium concept oblivious equilibrium.

2.6.3 Incomplete Theories

Eyster and Piccione analyzed a scenario in which traders have exogenous nonstationary consistent models of prices on the stock market; these models are called incomplete theories [33]. The traders use their theories to acquire assets. The key result is that traders with more complete theories do not necessarily perform better. The main difference with the proposed research is that, the actions of the traders do not influence prices. Therefore, prices are truly exogenous; traders do not need to update their models.

2.6.4 Egocentric Modeling

Seah and Shamma analyzed a specific game where two agents share a one-dimensional signal [34]. The signal is stochastic and influenced by the strategies of the agents. However, the agents model it with a consistent stationary exogenous model. Similarly, in the proposed research, an inaccurate simplified model is used to lower the complexity of computations.

3 Preliminary Research

The following presentation of the preliminary research was first developed in [35].

3.1 Empirical-evidence Equilibria

3.1.1 Single-agent Setup

Consider a discrete-time dynamical system governed by

$$x^+ \sim f(x, a, s), \tag{11}$$

where x is a state, a is an action, and s is a signal. Variables x, a, and s take values in finite sets $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{S}$, respectively. The agent picks the action a. Nature determines the signal s according to

$$w^+ \sim n(w, x, a), \tag{12a}$$

$$s \sim \nu(w), \tag{12b}$$

where w is a state of Nature evolving in the finite state space $\mathcal{W}$. The agent observes s but not w. Denote by $\mathbf{N}$ the dynamical system described by (11) and (12).

Define the agent's observation by $o = (x, a, s)$ and the actual realization of the system by $r = (w, x, a, s)$. At time t, the agent's private history is $p^t = (o^0, o^1, \ldots, o^t)$ and the true history is $h^t = (r^0, r^1, \ldots, r^t)$. Denote by $\mathcal{P}$ the set of finite private histories. A strategy $\sigma : \mathcal{P} \to \Delta(\mathcal{A})$ is a mapping from private histories to a distribution over the actions.

At each time step, the agent receives a payoff according to the utility function $u : \mathcal{X} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. For a given infinite private history the agent receives the sum of discounted payoffs $\sum_{t=0}^{\infty} \delta^t u(x^t, a^t, s^t)$, where $\delta \in (0, 1)$ is a discount factor. The agent wants to find a strategy maximizing its expected sum of discounted payoffs $U_{\mathbf{N},\sigma}(x^0) = \mathbb{E}_{\mathbf{N},\sigma}[\sum_{t=0}^{\infty} \delta^t u(x^t, a^t, s^t)]$.

When the agent knows (11) and (12), it is facing a POMDP. A natural solution concept for this type of problems is an optimal policy for the POMDP. The agent computes an optimal policy making use of beliefs, which are probability distributions over the true history. Beliefs are obtained from the private histories p^t, the signaling structure (12), and the application of Bayes's rule. Belief computation is intractable because every observation increases the size of the belief space.

When the agent knows (11) but does not know (12) it can still implement an optimal policy for the POMDP. However it cannot compute such an optimal policy anymore. In such a setting, a less constraining solution concept is required. Empirical-evidence optimality is one such solution concept that relies on the notion of statistical consistency.

The following section presents the simplest notion of statistical consistency, depth-k consistency.

3.1.2 Depth-k Consistency

Consider c, an $\mathcal{S}$-valued ergodic process. For k in $\mathbb{N}$, its depth-k characteristic χ^k is the long-run distribution of the strings of length $k+1$. For d in $\mathcal{S}^{k+1}$

$$\chi^k[d] = \lim_{t\to\infty} \mathbb{P}\big[\big(c^{t-k}, \dots, c^{t-1}, c^t\big) = d\big]. \tag{13}$$

Two processes with the same depth-k characteristic are called depth-k consistent.

The signal observed by the agent is one such $\mathcal{S}$-valued process. Consider another $\mathcal{S}$-valued process described by

$$z^+ = m^k(z, s), \tag{14a}$$
$$s \sim \mu(z), \tag{14b}$$

where z is a state in $\mathcal{S}^k$ and m^k is the length-k–memory function defined by

$$m^k\big(\big(s^{t-k}, \dots, s^{t-2}, s^{t-1}\big), s^t\big) = \big(s^{t-k+1}, \dots, s^{t-1}, s^t\big).$$

Under some technical assumptions, described in Section 3.1.3, the observed signal and the Markov chain described by (14) are ergodic processes. Furthermore, the Markov chain is depth-k consistent with the true signal when the following equality holds:

$$\mu(z)[s] = \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s^t = s \,\big|\, \big(s^{t-k}, \dots, s^{t-2}, s^{t-1}\big) = z\big].$$

Denote by $\mathbf{M}^k$ the dynamical system described by (11) and (14). The system $\mathbf{M}^k$ induces an MDP with state (x, z), action a, strategy $\hat{\sigma}\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{A}$, and the objective function $U_{\mathbf{M}^k,\hat{\sigma}}(x^0, z^0) = \mathbb{E}_{\mathbf{M}^k,\hat{\sigma}}[\sum_{t=0}^{\infty} \delta^t u(x^t, a^t, s^t)]$. A strategy $\hat{\sigma}$ for the MDP can be implemented in the real system by building z with (14a). From now on, no distinction will be made between a strategy for the MDP $\hat{\sigma}$ and its associated strategy built with (14a). Both strategies will be denoted σ.

Consider the following iterative process. The agent implements an initial strategy σ^0. It formulates a depth-k consistent model μ^0 of Nature's dynamic. Then, it computes an optimal strategy σ^1 for the MDP induced by this model μ^0. Upon implementation of this new strategy, the model μ^0 may lose the requisite statistical consistency. Therefore, the agent formulates a revised depth-k consistent model μ^1 and the process repeats. A fixed point of this iterative process is one way to define a solution to this problem. A strategy is

a solution if it is optimal with respect to the model it induces. Note that such a strategy is not a solution to the POMDP.

Using that model to design a strategy is equivalent to the agent making an assumption about the system. For example, when the agent uses a depth-k consistent model, it assumes the signal is generated exogenously, i.e., not impacted by x or a. This assumption might seem restrictive. However, note that the repeated-modeling and optimization phases create a feedback loop. Therefore, a model satisfying the consistency condition is exogenous but captures characteristics of Nature's dynamic.

The following section extends beyond the notion of depth-k consistency.

3.1.3 Empirical-evidence Optimality

The agent assumes that a Markov chain, with state z from a finite set $\mathcal{Z}$, generates the signal s and that it can construct z from its observations as follows:

$$z^+ \sim m(z, x, a, s), \tag{15a}$$

$$s \sim \mu(z). \tag{15b}$$

The model m represents the assumption the agent makes about the system. The predictor μ is the set of parameters the agent adjusts to be consistent with its observations. The pair (m, μ) is called a mockup.

In this setup, depth-k consistency is replaced with the following definition.

Definition 1. *Let σ be a strategy and (m, μ) be a mockup. Predictor μ is (σ, m) consistent with $\mathbf{N}$ if*

$$\mu(z)[s] = \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\left[s^{t+1} = s \,\middle|\, z^t = z\right].$$

The notion of optimality used is the following.

Definition 2. *Let σ be a strategy, (m, μ) be a mockup, and ε be a positive number. Strategy σ is (μ, m) optimal if it is optimal for the MDP induced by $\mathbf{M}$. Strategy σ is (ε, μ, m) optimal if it is ε optimal for the MDP induced by $\mathbf{M}$.*

Having defined consistency and optimality the definition of an empirical-evidence optimum (EEO) follows.

Definition 3. *Let σ be a strategy, (m, μ) be a mockup, and ε be a positive number. The pair (σ, μ) is an m EEO if the following two conditions hold:*

1. *Strategy σ is (μ, m) optimal.*
2. *Predictor μ is (σ, m) consistent with $\mathbf{N}$.*

The pair (σ, μ) is an (ε, m) EEO if the following two conditions hold:

1. *Strategy σ is (ε, μ, m) optimal.*

2. *Predictor μ is (σ, m) consistent with $\mathbf{N}$.*

A little care must be taken to make μ in Definition 1 well defined. Insuring the following assumption is verified guarantees it.

Assumption 1. *Let σ be a strategy, and T_σ be the Markov chain with state $X = (w, x, z)$ induced by $\mathbf{N}$ and σ, $X^+ \sim T_\sigma X$. The Markov chain T_σ is ergodic.*

Assumption 1 insures that T_σ has a unique stationary distribution π_σ such that

$$\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s \,\big|\, z^t = z\big] = \mathbb{P}_{\pi_\sigma}[s \,|\, z].$$

Furthermore, Assumption 1 guarantees that π_σ has full support, meaning that for all w in $\mathcal{W}$, x in $\mathcal{X}$, and z in $\mathcal{Z}$, $\pi_\sigma[w, x, z]$ is positive. This guarantees that μ in Definition 1 is well defined for all z and s as follows:

$$\begin{aligned}
\mu(z)[s] &= \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s \,\big|\, z^t = z\big] \\
&= \mathbb{P}_{\pi_\sigma}[s \,|\, z] \\
&= \sum_{w\in\mathcal{W}} \mathbb{P}_{\pi_\sigma}[s \,|\, z, w] \cdot \mathbb{P}_{\pi_\sigma}[w \,|\, z] \\
&= \sum_{w\in\mathcal{W}} \mathbb{P}_{\pi_\sigma}[s \,|\, w] \cdot \frac{\mathbb{P}_{\pi_\sigma}[w, z]}{\mathbb{P}_{\pi_\sigma}[z]} \\
&= \sum_{w\in\mathcal{W}} \nu(w)[s] \cdot \frac{\sum_{x\in\mathcal{X}} \pi_\sigma[w, x, z]}{\sum_{w'\in\mathcal{W}} \sum_{x\in\mathcal{X}} \pi_\sigma[w', x, z]}
\end{aligned}$$

Consistency yields a mapping associating to a strategy σ a unique predictor (σ, m) consistent with $\mathbf{N}$. Note that μ is a continuous function of π_σ.

Similarly, a mapping associating to a predictor μ a unique (ε, m)-optimal strategy can be defined. Denote by $\mathbf{M}$ the dynamical system described by (11) and (15). Consider the MDP induced by $\mathbf{M}$. Let $U^* : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be the value function for that MDP. Define $Q : \mathcal{X} \times \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ by

$$Q(x, z, a) = (1 - \delta)u(x, a) + \delta \mathbb{E}_{\mathbf{M}}\big[U^*\big(x^+, z^+\big) \,\big|\, x, z, a\big],$$

and σ by

$$\sigma(x, z)[a] = \frac{\mathrm{e}^{\frac{1}{\tau}Q(x,z,a)}}{\sum_{a'\in\mathcal{A}} \mathrm{e}^{\frac{1}{\tau}Q(x,z,a')}}.$$

As τ goes to 0, σ converges to a (μ, m)-optimal strategy. When τ is small enough, σ is (ε, μ, m) optimal. To guarantee uniqueness, define τ to be the largest value such that σ

is (ε, μ, m) optimal. Note that σ defined that way is a continuous function of the value function U^*.

One way to insure that Assumption 1 is verified is to have a small noise affect all the transitions. Formally, this means that for all $w \in \mathcal{W}$, $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$, $f(x, a, s)$, $n(w, x, a)$, $\nu(w)$, and $\sigma(x, z)$ have full support. From now on, Assumption 1 is always verified.

The following subsection extends the notion of EEOs to the multiagent case and defines EEEs.

3.1.4 Multiagent Setup

Consider a collection of agents $\mathcal{I}$. Each agent i has a state x_i, an action a_i, and a signal s_i. Let x be the tuple $\big(x_1, x_2, \ldots, x_{|\mathcal{I}|}\big)$. Define a and s similarly. Agent i is controlling the system described by

$$x_i^+ \sim f_i(x_i, a_i, s_i). \tag{16}$$

Agents $-i$ are controlling systems described as a whole by

$$x_{-i}^+ \sim f_{-i}(x_{-i}, a_{-i}, s_{-i}). \tag{17}$$

All these systems are coupled through Nature which determines the signals s according to

$$w^+ \sim n(w, x, a), \tag{18a}$$

$$s \sim \nu(w). \tag{18b}$$

Denote by $\mathbf{N}_i$ the system from agent i's perspective. In the single-agent setup, $\mathbf{N}$ was composed of a known part (11) and an unknown part (12). Similarly, $\mathbf{N}_i$ has a known part (16) and an unknown part (17) and (18).

The other definitions from previous sections can readily be extended to the multiagent case. Agent i has a utility function u_i, a discount factor δ_i, a strategy $\sigma_i : \mathcal{P}_i \to \Delta(\mathcal{A}_i)$, and a mockup of Nature and its opponents described by a state z_i, a model m_i, and a predictor μ_i.

From agent i's perspective, everything is identical to the single-agent setup. The notions of (μ, m) optimality, (ε, μ, m) optimality, and (σ, m) consistency can be replaced by (μ_i, m_i) optimality, $(\varepsilon_i, \mu_i, m_i)$ optimality, and (σ, m_i) consistency respectively. Therefore, the definition of EEO readily extends to the multiagent setting.

Definition 4. *Let σ, (m, μ), and ε such that for all i in $\mathcal{I}$, σ_i is a strategy, (m_i, μ_i) is a mockup, and ε_i is a positive number. The pair (σ, μ) is an m EEE if the following two conditions hold for all i in $\mathcal{I}$:*

1. *Strategy σ_i is (μ_i, m_i) optimal.*
2. *Predictor μ_i is (σ, m_i) consistent with $\mathbf{N}$.*

The pair (σ, μ) is an (ε, m) EEE if the following two conditions hold for all i in $\mathcal{I}$:

1. *Strategy σ_i is $(\varepsilon_i, \mu_i, m_i)$ optimal.*
2. *Predictor μ is (σ, m_i) consistent with $\mathbf{N}$.*

For a given m and ε such that for all i in $\mathcal{I}$, ε_i is a positive number, denote by $F^{\mathrm{O},m,\varepsilon}$ the optimization mapping from predictors to strategies and by $F^{\mathrm{M},m}$ the modeling mapping from strategies to predictors. These mappings are defined by direct extension of their single agent counterparts. Define $F^{m,\varepsilon}$, a mapping from the space of predictors to itself, by $F^{m,\varepsilon} = F^{\mathrm{M},m} \circ F^{\mathrm{O},m,\varepsilon}$.

3.2 Existence of Empirical-evidence Equilibria

Fix models m and ε such that for all i in $\mathcal{I}$, ε_i is a positive number.

Theorem 1. *There exists an (ε, m) EEE.*

Proof. First, show that $F^{m,\varepsilon}$ has a fixed point. The set of predictors is representable by a product of simplices. Therefore, $F^{m,\varepsilon}$ is a mapping from a convex and compact set to itself. By Propositions 2 and 3, $F^{\mathrm{O},m,\varepsilon}$ and $F^{\mathrm{M},m}$ are continuous. As the composition of two continuous functions, $F^{m,\varepsilon}$ is continuous. By application of Brouwer's fixed-point theorem, $F^{m,\varepsilon}$ has a fixed point.

Proposition 1 therefore implies that an (ε, m) EEE exists. □

Proposition 1. *Let μ^* be a fixed point of $F^{m,\varepsilon}$. Define σ^* by $\sigma^* = F^{\mathrm{O},m,\varepsilon}(\mu^*)$. The pair (μ^*, σ^*) is an (ε, m) EEE.*

Proof. By definition, strategy σ^* is $(\varepsilon_i, \mu_i^*, m_i)$ optimal. Note that

$$F^{\mathrm{M},m}(\sigma^*) = F^{\mathrm{M},m} \circ F^{\mathrm{O},m,\varepsilon}(\mu^*) = F^{m,\varepsilon}(\mu^*) = \mu^*.$$

This implies that predictor μ^* is (σ^*, m_i) consistent with $\mathbf{N}_i$. Therefore, (μ^*, σ^*) is an (ε, m) EEE. □

Proposition 2. *The optimization mapping $F^{\mathrm{O},m,\varepsilon}$ is continuous.*

Proof. Agent i's predictor only affects agent i's strategy. Therefore, proving that $F^{\mathrm{O},m,\varepsilon}$ is continuous, only requires showing that $F_i^{\mathrm{O},m,\varepsilon} : \mu_i \mapsto \sigma_i$ is continuous for all $i \in \mathcal{I}$. Decomposing this function as follows:

$$F_i^{\mathrm{O},m,\varepsilon} : \mu_i \overset{\text{(a)}}{\longmapsto} U_i^* \overset{\text{(b)}}{\longmapsto} \sigma_i,$$

it is sufficient to prove that (a) and (b) are continuous.

Lemma 1 shows that the value function of a finite MDP is a continuous function of the parameters of the problem. Since μ_i is one of the parameters of the MDP whose value function is U_i^*, (a) is continuous. It was noted in Section 3.1.3 that (b) is continuous. □

Proposition 3. *The modeling mapping $F^{\mathrm{M},m}$ is continuous.*

Proof. Agent i's strategy impacts all the agents' predictors. Proving the continuity of $F^{\mathrm{M},m}$, requires showing that $F^{\mathrm{M},m}_{i,j} : \sigma_i \mapsto \mu_j$ is continuous for all $i, j \in \mathcal{I}$. Decomposing this function as follows:

$$F^{\mathrm{M},m}_{i,j} : \sigma_i \overset{\text{(c)}}{\longmapsto} T_\sigma \overset{\text{(d)}}{\longmapsto} \pi_\sigma \overset{\text{(e)}}{\longmapsto} \mu_j,$$

it is sufficient to prove that (c), (d), and (e) are continuous.

Since (c) is linear, it is continuous. [36, Theorem 4.1] shows that the stationary distribution of a finite ergodic Markov chain is a continuous function of the elements of its transition matrix, which proves that (d) is continuous. It was noted in Section 3.1.3 that (e) is continuous. □

Lemma 1. *Consider a finite MDP described by a dynamic $x^+ \sim f(x, a)$, a utility function $u(x, a)$, and a discount factor δ. Denote by θ the finite vector of all the entries in f and u. Let B_θ be the Bellman operator associated with the problem. By definition, the value function of the problem U^*_θ is the fixed point of B_θ, $U^*_\theta = B_\theta U^*_\theta$.*

The function $\theta \mapsto U^*_\theta$ is continuous.

Proof. Let θ and θ' be two vectors of parameters. The value function U^*_θ is a fixed point of B_θ. The Bellman operator B_θ is a contraction mapping with Lipschitz constant δ. As a result,

$$\begin{aligned}
\|U^*_\theta - U^*_{\theta'}\| &= \|B_\theta U^*_\theta - U^*_{\theta'}\| \\
&\leq \|B_\theta U^*_\theta - B_\theta U^*_{\theta'}\| + \|B_\theta U^*_{\theta'} - U^*_{\theta'}\| \\
&\leq \delta \|U^*_\theta - U^*_{\theta'}\| + \|B_\theta U^*_{\theta'} - U^*_{\theta'}\| \\
&\leq \frac{1}{(1-\delta)} \|B_\theta U^*_{\theta'} - U^*_{\theta'}\|.
\end{aligned}$$

The continuity of $\theta \mapsto B_\theta U^*_{\theta'}$ can now be established. By definition, $(B_\theta U^*_{\theta'})(x) = \max_{a \in \mathcal{A}} v(x, a, \theta)$, where $v(x, a, \theta) = (1-\delta)u(x, a) + \delta f(x, a)^{\mathrm{T}} U^*_{\theta'}$. For fixed x and a, $\theta \mapsto v(x, a, \theta)$ is linear and therefore continuous. For a fixed x, $\theta \mapsto B_\theta U^*_{\theta'}(x)$ is the maximum of a finite number of continuous functions and as such is continuous. The function $\theta \mapsto B_\theta U^*_{\theta'}$ is continuous because each of its finitely many components is continuous.

Continuity of $\theta \mapsto B_\theta U^*_{\theta'}$ implies that $\|U^*_\theta - U^*_{\theta'}\|$ goes to zero as θ goes to θ'. This last statement concludes the proof. □

3.3 Learning Empirical-evidence Equilibria

3.3.1 A Learning Rule

The fixed points of $F^{m,\varepsilon}$ are (ε, m) EEEs. A natural approach to try and learn an (ε, m) EEE is to use an adaptive rule that converges to fixed points. Consider the following

adaptive rule:

$$\mu^{t+1} = \mu^t + \alpha^t\big(F^{m,\varepsilon}(\mu^t) - \mu^t\big), \tag{19}$$

where α^t is a step size. The long-run behavior of (19) is related to properties of the following differential equation:

$$\dot{\mu} = F^{m,\varepsilon}(\mu) - \mu.$$

In particular, Benaïm showed that the limit set of (19) is a connected set internally chain-recurrent for the flow induced by $F^{m,\varepsilon} - \mathrm{Id}$, where Id is the identity function [37]. The fixed points of $F^{m,\varepsilon}$ are connected sets internally chain-recurrent for the flow induced by $F^{m,\varepsilon} - \mathrm{Id}$ but they might not be the only ones. Therefore, if (19) converges it might yield an (ε, m) EEE.

3.3.2 Simulation Results

This learning rule was successfully used on a simplified market example. Two agents can hold a quantity of a single asset between 0 and 4, $\mathcal{X} = \{0, 1, 2, 3, 4\}$. At each time step, each agent can sell one asset, buy one asset, or hold its position, $\mathcal{A} = \{\text{Sell}, \text{Hold}, \text{Buy}\}$. The assets can be traded at a low price or at a high price, $\mathcal{S} = \{\text{Low}, \text{High}\}$. Nature exogenously determines the market trend as a bull market or a bear market, $\mathcal{W} = \mathcal{S} \times \{\text{Bear}, \text{Bull}\}$. The price is impacted by the past price, the market trend, and the orders placed by the two agents. A high price in the past, buying orders, or a bull market increase the chances of seeing a high price in the future. The agents receive the price at each time step but are not aware of the price dynamic. In this model, they are not even aware of the existence of the market trend. The two agents use a discount factor $\delta = 0.95$.

Agent 1 starts with the idea that the price will be high with probability 1. Agent 2 starts with the idea that the price will be low with probability 1. Each agent is trying to learn a depth-0 model of the price. Two versions of (19) were simulated. The first one used (19) directly with a fixed step size of 0.1. The stationary distribution π_σ was computed at each time step to obtain the true value of $F^{m,\varepsilon}(\mu^t)$, The results of the simulations using the theoretical predictor are presented in Figure 2. Since the price is a public signal, after a transient phase due to the step size, the predictions of both agents agree. The prediction converges to probability of seeing a high price of 0.431. The two agents use the same strategy that is the optimal response for that prediction of the price. When the price is high sell. When the price is low, sell when having four units, hold when having three units, and buy otherwise. The learning rule has indeed converged to an EEE.

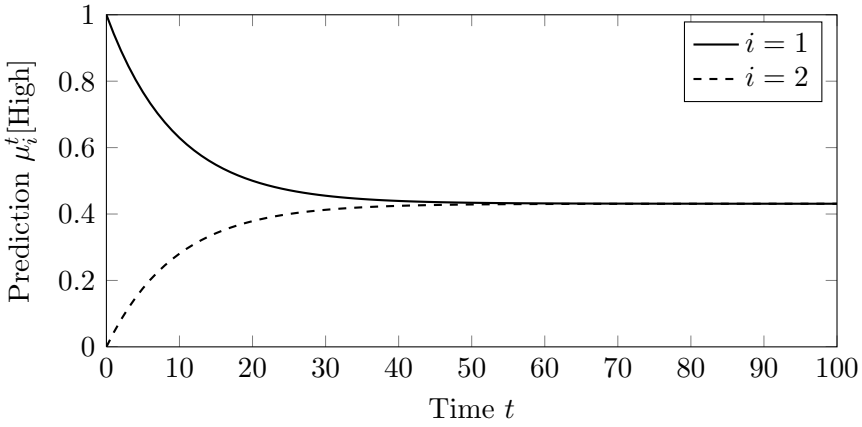

Figure 2. Simulation results of two agents learning a depth-0 model of the price for the market example with the theoretical predictor.

In the second simulated version of (19), the stationary distribution was only estimated by playing 100 rounds of the game at each time step. Because of the variance induced by this sampling process, the step size was taken to be diminishing, $\alpha^t = \left(\frac{1}{t}\right)^{\frac{3}{4}}$. The estimated predictors obtained in that case are denoted by $\hat{\mu}_i^t$. The results of the simulation using the empirical predictors are presented in Figure 3. Estimating, instead of using the true probability, induces some variations. The learning rule does not converge, but oscillates around the EEE reached by the theoretical predictor.

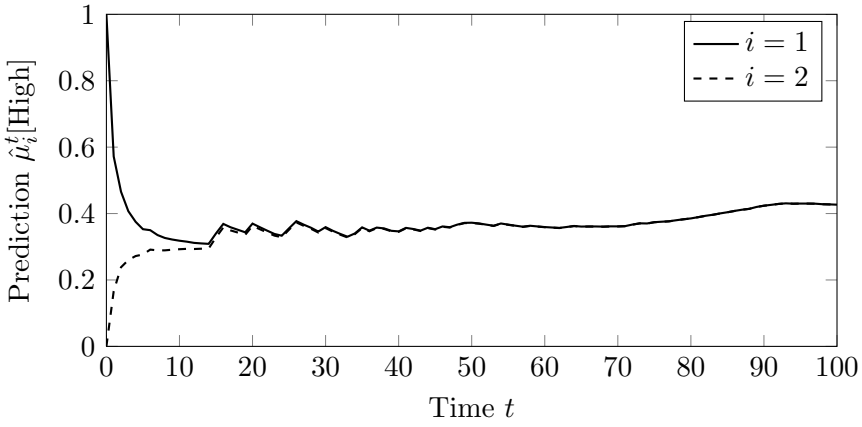

Figure 3. Simulation results of two agents learning a depth-0 model of the price for the market example with an empirical predictor.

4 Proposed Research

The objective of the proposed research is to develop the framework of EEEs in stochastic games and to design decentralized controllers using learning in that framework. This research started by trying to apply game-theoretic results to decentralized control. Using game theory to design a controller entails computing equilibrium strategies for a specific game. For decentralized controllers, computing the strategies in a decentralized fashion through learning is an undeniable advantage. Stochastic games are of particular interest for controls since they extend MDPs. However, the computation of equilibrium strategies in stochastic games is an open problem. The main reason for this lack of result is that computing equilibrium strategies in a general stochastic game requires each agent to solve a POMDP. As exposed in Chapter 2, this issue stems from the full rationality requirement imposed by classical game theory. With this consideration in mind, this research was steered towards bounded rationality. In stochastic games, bounded rationality commonly appears in the form of consistency. Agents using consistency are not required to have perfect understanding of their environment but only a statistically consistent understanding. In the preliminary research, the foundations of a general consistency framework have been laid down and EEEs have emerged as a solution concept for that framework. Understanding the properties of EEEs has become the primary goal of this research.

The next section list all the open questions surrounding EEEs for completeness sake. The following one defines the scope of the proposed work.

4.1 Open Questions

4.1.1 Implications of Using Consistency

The fact that agents use consistent models in EEEs diminishes the amount of computation they require to obtain optimal strategies. However, it also imposes constraints on the attainable equilibria and associated strategies. The first step to understand those constraints is to analyze the simplest notion of consistency, which is depth-k consistency.

What impact does varying k have? Eyster and Piccione gave an answer in a specific setting where the strategies of the agents did not impact the environment [33]. They proved that agents with a larger k, synonymous with better understanding of the environment, did not always receive a larger payoff. This question has to be addressed for a more general setting.

As k increases, the agent get a more accurate prediction of the strings of signals. This

raises the question to know what happens in the limit. Do depth-k EEEs converge to Nash equilibria as k goes to infinity?

4.1.2 Large Number of Agents

In a mean-field game, agents face identical problems and impact their opponents through the empirical distribution of states of all the agents. An MFE is an equilibrium in which these agents use depth-0 consistency. These MFEs are studied when the number of agents is large. Restricting the attention to this specific setting with a large number of agents allows for the derivation of strong results. The main result states that as the number of agents grows to infinity, an MFE converges to a Nash equilibrium. In other words, the approximation made by the agents regarding the empirical distribution of states does not change the behavior of the system. This result is a consequence of the central limit theorem and it would be interesting to generalize it to a broader setting.

In the MFE setting, the agents are homogeneous and impact their opponents only through their state. The EEE framework lifts these two restrictions. In particular the impact agents have on their opponents is embedded in the signal definition. Can results from the MFE literature be extended to the broader framework of EEEs? In particular, the EEEs framework offers the opportunity to explore the consequences of the central limit theorem for a broader class of consistency than the sole depth-0.

4.1.3 Learning

EEEs were informally defined as fixed points of a simple iterative process. The existence of fixed points was established in the preliminary research. However, the convergence of the iterative process to such a fixed point is not guaranteed. Building a learning rule converging to EEEs can be done in two steps. First, a theoretical learning rule converging to EEEs is designed. Then, a practical online version of the rule is derived. This approach was used in the simulations of Section 3.3. The theoretical learning rule uses the stationary distribution of the whole system. This information is not available to the agents as they play but it matches closely the requirements of EEEs. However, the agents can estimate the stationary distribution of the system by observing the play long enough. Hart and Mas-Colell used this two-step approach to prove the convergence of an adaptive no-regret learning rule to correlated equilibria [14]. The adaptive learning rule replaced a matrix inversion step by a simpler maximization one.

4.1.4 Price of Anarchy

Given a global objective, a multiagent system can be controlled by a centralized or a decentralized controller. In a centralized approach, an optimal controller for the objective is computed offline. Each agent is then given to execute a part of this controller. In a decentralized approach using game theory, each agent is given a utility function along

with a learning rule. In this case, the controller corresponds to the equilibrium reached by the learning process. The decentralized approach is more robust and scalable than the centralized approach. However, these advantages come with a cost; the decentralized controller is suboptimal. For systems whose global objective coincide with maximizing the sum of the utility functions, this cost can be evaluated by a metric called the price of anarchy [38]. The sum of the utility functions of the agents is called the social welfare, and the ratio of social welfare between the decentralized and centralized controllers is considered. The price of anarchy is the worst case ratio. In a learning context, the ratio is a random variable and properties other than its minimum value can be computed. This notion, classically defined for Nash equilibria, readily extends to EEEs. What is the price of anarchy for EEEs?

4.1.5 Payoff Folk Theorem

Payoff folk theorems for repeated games prove that all the feasible individually strictly rational payoff profiles are sustainable by subgame-perfect equilibria. This implies that subgame-perfect equilibria sustain almost all payoff profiles. Some of these payoff profiles are undesirable, for example the non-Pareto-optimal ones. Do EEEs sustain such a large set of payoff profiles? If so, can equilibrium selection reduce the size of that set?

4.2 Proposed Work

The proposed research will focus in priority on learning while exploring the implications of using consistency and the effects of a large number of agents. The results will be illustrated on examples inspired from the smart grid and air traffic control.

The smart grid is an heterogeneous multiagent system composed of producers and consumers of power. On the one hand, some agents require a deep understanding of the system to make decisions. Power companies or companies operating data centers are example of such big players. On the other hand, individual home owners only require a basic understanding to make decisions about their power consumption. The EEE framework offers tools to analyze different levels of rationality.

Air traffic control is currently a centralized task relying heavily on air traffic controllers. This approach is reaching its limits and researchers are looking for solutions to give more autonomy to airlines and pilots. Current research mostly falls into two groups, micro modeling and macro modeling. At the micro level each plane is modeled, and at the macro level flows of planes are considered. The EEE framework offers a potential bridge between these two approaches. Planes can be studied individually but the impact of other planes can be aggregated in a consistent manner.

Bibliography

[1] J. v. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 2nd ed. Princeton, NJ: Princeton University Press, 1947.

[2] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ: John Wiley & Sons, 1994.

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, Mar. 1998.

[6] J. F. Nash, "Non-cooperative games," *Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, Sep. 1951.

[7] R. J. Aumann, "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, vol. 55, no. 1, pp. 1–18, Jan. 1987.

[8] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[9] L. S. Shapley, *Some Topics in Two-Person Games*, ser. Memorandum (Rand Corporation). Santa Monica, CA: Rand Corporation, Oct. 1963.

[10] S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium," *The American Economic Review*, vol. 93, no. 5, pp. 1830–1836, Dec. 2003.

[11] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium," in *38th ACM Symposium on Theory of Computing*, May 2006, pp. 71–78.

[12] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.

[13] ——, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, no. 1, pp. 26–54, May 2000.

[14] ——, *Economic Essays*. New York: Springer, 2001, ch. A reinforcement procedure leading to correlated equilibrium, pp. 181–200.

[15] H. P. Young, *Individual Strategy and Social Structure*. Princeton, NJ: Princeton University Press, 1998.

[16] ——, *Strategic Learning and Its Limits*. Oxford, England: Oxford University Press, 2004.

[17] N. Li and J. R. Marden, "Designing games for distributed optimization," in *50th IEEE Conference on Decision and Control*, Dec. 2011, pp. 2434–2440.

[18] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, Jan. 1993.

[19] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," in *51st IEEE Conference on Decision and Control*, Dec. 2012, pp. 7419–7424.

[20] M. J. Fox and J. S. Shamma, "Self-assembly for maximum yields under constraints," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2011.

[21] ——, "Language evolution in finite populations," in *50th IEEE Conference on Decision and Control*, Dec. 2011, pp. 4473–4478.

[22] G. J. Mailath and L. Samuelson, *Repeated Games and Reputations: Long-Run Relationships*. Oxford, England: Oxford University Press, Oct. 2006.

[23] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," in *Journal of Machine Learning Research*, vol. 4, Nov. 2003, pp. 1039–1069.

[24] E. Kalai and E. Lehrer, "Subjective equilibrium in repeated games," *Econometrica*, vol. 61, no. 5, pp. 1231–1240, Sep. 1993.

[25] D. Fudenberg and D. K. Levine, "Self-confirming equilibrium," *Econometrica*, vol. 61, no. 3, pp. 523–545, May 1993.

[26] T. Sugaya, "Folk theorem in repeated games with private monitoring," Nov. 2011, unpublished.

[27] M. Kandori, "Weakly belief-free equilibria in repeated games with private monitoring," *Econometrica*, vol. 79, no. 3, pp. 877–892, May 2011.

[28] P. Jehiel, "Analogy-based expectation equilibrium," *Journal of Economic Theory*, vol. 123, no. 2, pp. 81–104, Aug. 2005.

[29] A. Rubinstein, *Modeling Bounded Rationality*. Cambridge, MA: MIT Press, 1998.

[30] Y.-H. Chang, T. Ho, and L. P. Kaelbling, "All learning is local: Multi-agent learning in global reward games," in *Advances in Neural Information Processing Systems 16*, 2004.

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

[32] G. Y. Weintraub, C. L. Benkard, and B. Van Roy, "Markov perfect industry dynamics with many firms," *Econometrica*, vol. 76, no. 6, pp. 1375–1411, Nov. 2008.

[33] E. Eyster and M. Piccione, "An approach to asset-pricing under incomplete and diverse perceptions," Dec. 2011, unpublished.

[34] V. P.-W. Seah and J. S. Shamma, "Multiagent cooperation through egocentric modeling," J. S. Shamma, Ed. Hoboken, NJ: John Wiley & Sons, Feb. 2008, ch. 9, pp. 213–229.

[35] N. Dudebout and J. S. Shamma, "Empirical evidence equilibria in stochastic games," in *51st IEEE Conference on Decision and Control*, Dec. 2012, pp. 5780–5785.

[36] C. D. Meyer, Jr., "The condition of a Markov chain and perturbation bounds for the limiting probabilities," *SIAM Journal on Algebraic and Discrete Methods*, vol. 1, no. 3, pp. 273–283, Sep. 1980.

[37] M. Benaïm, "A dynamical system approach to stochastic approximations," *SIAM Journal on Control and Optimization*, vol. 34, no. 2, pp. 437–472, Mar. 1996.

[38] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *16th Symposium on Theoretical Aspects of Computer Science*, Mar. 1999, pp. 404–413.