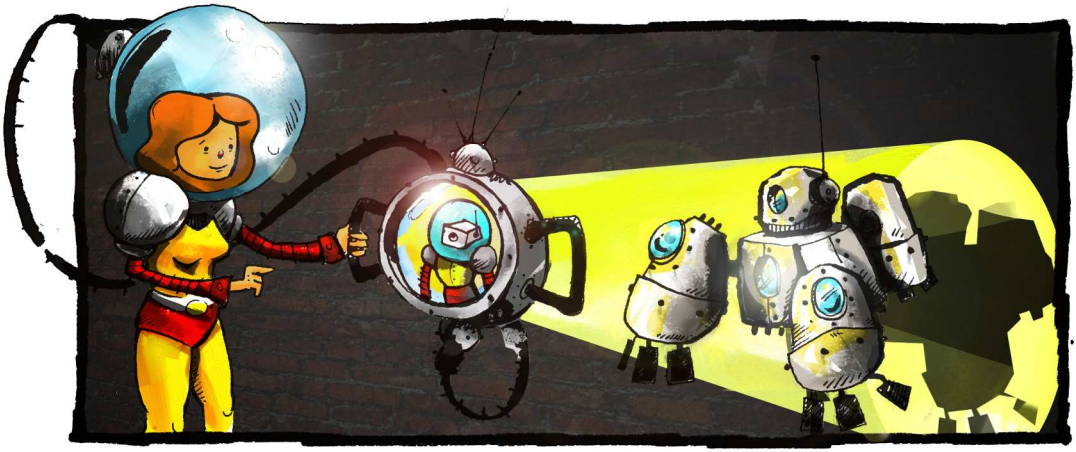
Empirical-evidence Equilibria in Stochastic games



Nicolas Dudebout

Empirical-evidence Equilibria in Stochastic Games

A Thesis
Presented to
the Academic Faculty

by

Nicolas Dudebout

In Partial Fulfillment of
the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
April 2014

Empirical-evidence Equilibria in Stochastic Games

Approved by:

Professor Jeff Shamma, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Éric Féron, Co-advisor
School of Aerospace Engineering
Georgia Institute of Technology

Professor Fumin Zhang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Spiridon Reveliotis
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Maria-Florina Balcan
School of Computer Science
Georgia Institute of Technology

Professor Yorai Wardi
School of Electrical and Computer Engineering
Georgia Institute of Technology

April 14, 2014

We don't see things as they are, we see them as we are.
–Anaïs Nin

To my ever-supportive wife, Laura.

Acknowledgments

I do not spend much time watching reality television, but the Discovery Channel's Gold Rush is an exception. When my wife asks me to turn it off, I explain to her that Gold Rush is helping me finish my PhD. I am not sure she ever bought it, but I can definitely draw some similarities between the show and my pursuit of a doctoral degree. Mining happens on two kinds of claims. Some claims have already been mined. The nuggets have been taken. However, the good areas are known and new tools allow for the extraction of gold flakes left behind.

The second kind of claim is known as *virgin ground*. This area has never been mined, so gold nuggets can potentially be found. My advisor, Jeff Shamma, showed me some *virgin ground* at the onset of my PhD and for this I will forever be thankful.

On *virgin ground*, you spend days, months, and in my case, years, taking the lay of the land. In both the pursuit of gold and the PhD, the unexplored territories are daunting. However, regardless how painful the process, finding gold nuggets feels like magic. Jeff, you gave me the claim and stayed by my side until I struck gold. You never hovered over my shoulder, but you continually guided me towards the treasure. Your ability to help take a step back and think about the core of a problem is amazing. Thank you for seeing me through this opportunity of a lifetime. I know sometimes I was a discouraged miner, but you helped me keep my chin up. It was truly a pleasure working with you.

Secondly, I would like to thank my co-advisor, Éric Féron. Your continued support and industry-oriented approach was the perfect complement to my theoretical thesis. Even though I did not sit in your lab, you always were inclusive and helped me feel like I was part of your team.

Advisors guide you through a PhD, but labmates are there to catch you when you fall. Without them, it would oftentimes be hard to get back up. For those who have been through the process, you understand that getting a PhD takes team work. Thank you for being there and helping me make my way through the mine.

Next, I want to thank those who became close confidants during my doctoral studies. Some of you were in my lab, others of you I met through classes and symposiums. Most of the time we spoke at Starbucks over doppio espressos. Thank you to Francesco Barale, Romain Pelard, Patrick Melet,

Travis Deyle, Pierrick Burgain, William Mantzel, Michael Fox, Malachi Jones, Georgios Kotsalis, Georgios Chasparis, Georgios Piliouras, Aurèle Balavoine, Chris Turnes, Karen Pottin, Sam Shapero, Taylor Shapero, Vlad Popescu, Romain Jobredeaux, Aude Marzuoli, Emmanuel Boidot, Florian Hauer, Tim Wang, Ibrahim al-Shyoukh, Yasin Yazicioglu, Daniel Pickem, Rowland O'Flaherty, Greg Droge, Jean-Pierre de la Croix, Baptiste Coudriller, Behrouz Touri, Sebastian Ruff, Lichun Li, and Kwang-Ki Kim.

I also want to thank my committee members, Fumin Zhang, Spyros Reveliotis, Nina Balcan, and Yorai Wardi, for their invaluable feedback. Sometimes you can get tunnel vision when you are digging for gold. However, your insight helped me take a step back and focus on the big picture. I appreciate your help in letting me chisel the dirt away from my dissertation so it would have a chance to sparkle.

My acknowledgment section would not be complete without recognizing my family. My parents, grandparents, mother-in-law, and siblings were always supportive of me pursuing a doctoral degree. I cannot imagine this journey without their love and guidance. Lastly, I must thank my wife for her many sacrifices so that I could strike gold. Laura was always there for me. She listened when I needed to talk and texted me during the day when I needed encouragement. Laura, thank you for your support. You deserve every bit of your PhT because you really did, *put hubby through*.

Contents

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 $\mathcal{B}$. For $\beta \in \Delta(\mathcal{B})$, $B \sim \beta$ denotes that the random variable B is drawn according to distribution β. $\mathbb{P}_\beta[E]$ is the probability of event E under distribution β. $\beta[b]$ denotes the quantity $\mathbb{P}_\beta[B = b]$. $\mathbb{E}_\beta[B]$ 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 $\left(b_1, b_2, \ldots, b_{|\mathcal{I}|}\right)$.

Acronyms

ABEE analogy-based expectation equilibrium 56

EEE empirical-evidence equilibrium 1, 2, 78, 82, 83, 85–87, 93, 96, 97, 100–103

EEO empirical-evidence optimum 74, 78, 82

MDP Markov decision process 1, 2, 31–36, 39, 41, 43, 44, 48, 49, 51, 52, 54, 57, 59–62, 66–68, 70, 74, 77, 83–86, 101

MFE mean-field equilibrium 57, 102

POMDP partially observable Markov decision process 1, 38–40, 48, 49, 52, 59, 60, 101

xEEE exogenous empirical-evidence equilibrium 87, 89, 90, 92–96

1 Introduction

The objective of this research is to develop the framework of empirical-evidence equilibria (EEEs) in stochastic games. This framework was developed while attempting to design decentralized controllers using learning in stochastic games. The overarching 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 this 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 mockup. The main contribution of this dissertation is the use of modeling to study stochastic games. This approach, while common in engineering, had not been applied to stochastic games. This dissertation is organized as follows.

Chapter 2 presents background material on game theory. The notions

of best response, solution concept for a game, and equilibria are at the heart of this chapter. The, often overlooked, distinction between correlated equilibria and correlated-equilibrium distributions is also made. Finally, a proof of the existence of Nash equilibria is given. This proof has been crafted to make the proof of the existence of EEEs, which is presented later, as intuitive as possible.

Chapter 3 presents repeated games and stochastic games. Their introduction relies on the notions presented in Chapter 2 and on dynamic programming. Using the vocabulary of MDPs makes the topics of sequential rationality and folk theorem in repeated games easier to grasp.

Chapter 4 presents existing results concerning decentralized control and game theory. Three main classes of results are addressed: learning in games, equilibria in repeated games, and use of bounded rationality.

Finally, Chapter 5 introduces EEEs and presents results in this framework. The presentation starts by analyzing a single-agent problem. In this setup, the notions of consistency and optimality are defined. These notions are then extended to encompass stochastic games. The second part of this chapter highlights three important results in the EEE framework. First, the existence of EEEs is proven. Second, a characterization of EEEs in perfect-monitoring repeated games is given in terms of correlated equilibria. Third, a result regarding learning EEEs with a finite observation window is presented.

2 Static Game Theory

2.1 Decision Making

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 also make nondeterministic decisions. Instead of committing to a specific action, it can choose a mixed action. In game-theoretic terms, a mixed action α is a distribution over the action set, i.e. an element of $\Delta(\mathcal{A})$. Similarly, the actions in the original action set are often called pure actions. A mixed action's payoff is the expected value of the payoffs of the pure actions in its support. For example, choosing a with probability $\frac{1}{3}$ and a' with probability $\frac{2}{3}$ yields a payoff $\frac{1}{3}u(a) + \frac{2}{3}u(a')$. As a result, the domain of the utility function can be unambiguously extended from the action set $\mathcal{A}$ to distributions over the action set $\Delta(\mathcal{A})$, i.e. $u\colon \Delta(\mathcal{A}) \to \mathbb{R}$. For an element α in $\Delta(\mathcal{A})$, $u(\alpha) = \mathbb{E}_{A\sim\alpha}[u(A)]$. Therefore, given a utility function, solving a decision-making problem is equivalent to solving a stochastic optimization problem

$$\arg\max_{\alpha\in\Delta(\mathcal{A})} u(a).$$

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 of irrelevant alternatives. 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.

Formally, a preference is a total order on distributions over actions. Such a binary relation is classically represented by the infix operator $\succeq$. Given two distributions α and β, the fact that the agent prefers α to β is denoted by $\alpha \succeq \beta$. This preference is not strict and the agent could in fact be indifferent between α and β if $\beta \succeq \alpha$ is also true. The four axioms of rational preferences are defined as follows:

Completeness *Given two distributions α and β, then $\alpha \succeq \beta$ or $\beta \succeq \alpha$.*

Transitivity *Given three distributions α, β, and γ such that $\alpha \succeq \beta$ and $\beta \succeq \gamma$, then $\alpha \succeq \gamma$.*

Continuity *Given three distributions α, β, and γ such that $\alpha \succeq \beta \succeq \gamma$, there exists $p \in [0, 1]$ such that $\beta = p\alpha + (1 - p)\gamma$.*

Independence of irrelevant alternatives *Given two distributions α and β such that $\alpha \succeq \beta$, a third distribution γ, and $p \in [0, 1]$, then $p\alpha + (1 - p)\gamma \succeq p\beta + (1 - p)\gamma$.*

2.2 Games and Nash 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 for agent i depends on the actions of all the agents and not only its own. The tuple composed of all these utility functions $u = (u_i)_{i \in \mathcal{I}}$ defines a game. 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 $\mathcal{A}_i$ optimal for the fixed actions of $-i$ are called best responses of agent i. Let $a_{-i} = \left(a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{|\mathcal{I}|}\right)$ denote a tuple of $|\mathcal{I}| - 1$ actions corresponding to one action for each opponent of agent i. The set of all such actions $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_{i-1} \times \mathcal{A}_{i+1} \times \cdots \times \mathcal{A}_{|\mathcal{I}|}$ is called the joint action set of the opponents of agent i and is denoted by $\mathcal{A}_{-i}$. For a fixed a_{-i}, the best-response set of agent i is $\mathrm{BR}_i(a_{-i}) = \arg\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i})$, a subset of $\mathcal{A}_i$.

Note 2 (Set-valued Functions). *The mapping* $\mathrm{BR}_i \colon \mathcal{A}_{-i} \to 2^{\mathcal{A}_i}$ *is a set-valued function. It takes an element in* $\mathcal{A}_{-i}$ *and returns a subset of* $\mathcal{A}_i$*. The set of subsets of* $\mathcal{A}_i$ *is called the power set of* $\mathcal{A}_i$*. This power set is commonly denoted by* $2^{\mathcal{A}_i}$*. For each joint actions of agent* i*'s opponents* $a_{-i} \in \mathcal{A}_{-i}$, $\mathrm{BR}_i(a_{-i})$ *contains one or more actions of agent* i *that are optimal against* a_{-i}*. In particular, it never returns an empty set. Set-valued functions with this property are called correspondences.*

Correspondences $f \colon \mathcal{A} \to 2^{\mathcal{B}}$ *from* $\mathcal{A}$ *to subsets of* $\mathcal{B}$ *have similarities with functions from* $\mathcal{A}$ *to* $\mathcal{B}$*. The classical notation for correspondences* $f \colon \mathcal{A} \rightrightarrows \mathcal{B}$ *emphasizes these similarities. With this notation, the best-response correspondence is such that* $\mathrm{BR}_i \colon \mathcal{A}_{-i} \rightrightarrows \mathcal{A}_i$*.*

Theorems about functions often have correspondence counterparts. For example, Kakutani's fixed-point theorem is an extension of Brouwer's fixed-point theorem. Brouwer's theorem proves the existence of fixed points for continuous functions on convex compact sets. Kakutani's theorem replaces continuity by a set of conditions on the graph of a correspondence to reach a similar conclusion. These theorems will be used to prove the existence of equilibria in Section 2.5

2.2.1 Pure Nash Equilibrium

A joint action that is simultaneously a best response for all the agents is a reasonable 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 pure Nash equilibrium if and only if for all $i \in \mathcal{I}$, $a_i \in \mathrm{BR}_i(a_{-i})$.

Defining an optimal action in a single-agent decision-making problem is straightforward and unambiguous. Given a utility function, an action is optimal if and only if it maximizes this utility function. There are many ways to illustrate what optimality means, and the following story is one of them. Consider a rational agent facing a decision-making problem. Suppose the agent knows the utility function associated with its rational preferences. The agent is asked to submit an action. Then, the agent is asked if, given the circumstances, it is satisfied with its choice. Obviously, the answer is yes if and only if the action is a utility maximizer, i.e. an optimal action. This story seems silly and does not add anything to the definition of optimality. In particular, it is unclear why the notion of circumstances is introduced. However, this kind of stories is key in defining game-theoretic solution concepts. In the present context, the story is redundant because the notion of optimal action is intrinsic to the decision-making problem.

Let's now look at the story corresponding to the Nash equilibrium. It will help shed some light on the so-called circumstances mentioned previously. Consider a group of rational agents facing a game. Suppose each agent knows the utility function associated with its rational preferences. Each

agent is asked to submit a pure action without discussing with the other agents. Then, each agent is asked, if given the circumstances, it is satisfied with its action. In this context, the circumstances are the actions of all the other agents. Agent $i \in \mathcal{I}$ is satisfied if and only if $a_i \in \mathrm{BR}_i(a_{-i})$. Therefore, all the agents are simultaneously satisfied if and only if the joint action is a joint utility maximizer, i.e. a pure Nash equilibrium.

This story emphasizes that a pure Nash equilibrium is not an intrinsic solution concept. Modifying some elements of the story would yield a different solution concept. The three main characteristics leading to a Nash equilibrium are the following:

Independent action selection Players communication is proscribed. As a result, they choose their actions independently of each other. Note that preventing communication is not intrinsic to the game.

Unilateral deviation Each agent is asked if it is happy given the other $|\mathcal{I}|-1$ actions are fixed. In other words, each agent is asked if it would prefer to unilaterally deviate. Nothing in the formulation of the game emphasizes these unilateral changes. The agents could, for example, deviate in pairs.

Static concept The solution concept defined is static. Each agent is asked if it is satisfied with its action and the story ends. The agents do not choose a new action and the process does not repeat. Using only a static solution concept is, once again, not intrinsic to the game formulation. It is a common misconception to see some notion of time in the story. This intuition to repeat the process actually prefaces learning in games which is covered later on.

To find a Nash equilibrium, you do not necessarily have to use this story. You can imagine receiving a joint action as the result of an optimization problem, that turns out to be a Nash equilibrium. However, a joint action is a pure Nash equilibrium if and only if it can be cast in this storyline. This last statement will be used shortly to introduce a new solution concept that is not a Nash equilibrium.

2.2.2 Mixed Nash Equilibrium

The definition of best response is readily extended to mixed actions. To do so, remember that the utility of a mixed action $\alpha \in \Delta(\mathcal{A})$ is defined as $u(\alpha) = \mathbb{E}_{A\sim\alpha}[u(A)]$. In the original definition of best response, there was nothing specific to pure actions. Let i be an agent. For a distribution over the joint action set of its opponents $\alpha_{-i} \in \Delta\Big(\prod_{j\in-i} \mathcal{A}_j\Big)$, define $\mathrm{BR}_i(\alpha_{-i}) = \arg\max_{\alpha_i\in\Delta(\mathcal{A}_i)} u_i(\alpha_i, \alpha_{-i})$. This defines a correspondence $\mathrm{BR}_i \colon \Delta\Big(\prod_{j\in-i} \mathcal{A}_j\Big) \rightrightarrows \Delta(\mathcal{A}_i)$. Note that the best-response mapping

for agent i is defined for all distributions over the joint action set of the opponents $\Delta\left(\prod_{j\in -i}\mathcal{A}_j\right)$ and not only product of independent distributions $\prod_{j\in -i}\Delta(\mathcal{A}_j)$.

Note 3 (Distributions over Product Spaces). *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two finite sets associated with two agents 1 and 2. Let $\beta_1 \in \Delta(\mathcal{B}_1)$ and $\beta_2 \in \Delta(\mathcal{B}_2)$ be distributions over these sets. Agent 1 draws a sample b_1 from β_1. Agent 2 independently draws a sample b_2 from β_2. The resulting pair $b = (b_1, b_2)$ is drawn according to distribution β, such that $\beta[b_1, b_2] = \beta_1[b_1]\beta_2[b_2]$. This distribution β is called the product distribution of β_1 and β_2, denoted by $\beta = (\beta_1, \beta_2)$. There are, however, more distributions over $\mathcal{B}_1 \times \mathcal{B}_2$ than product distributions, i.e. $\Delta(\mathcal{B}_1) \times \Delta(\mathcal{B}_2) \subsetneq \Delta(\mathcal{B}_1 \times \mathcal{B}_2)$. A distribution that is not a product distribution cannot be written as a pair, nor as a tuple when working in more than two dimensions.*

Given a distribution $\beta \in \Delta(\mathcal{B}_1 \times \mathcal{B}_2)$ the marginal distributions β_1 and β_2 over $\mathcal{B}_1$ and $\mathcal{B}_2$ are defined as follows:

$$\forall b_1 \in \mathcal{B}_1,\ \beta_1[b_1] = \sum_{b_2\in\mathcal{B}_2} \beta[b_1, b_2],$$
$$\forall b_2 \in \mathcal{B}_2,\ \beta_2[b_2] = \sum_{b_1\in\mathcal{B}_1} \beta[b_1, b_2].$$

In general, the marginal distributions alone are not enough to reconstruct the original distribution. Product of independent distribution $\beta = (\beta_1, \beta_2)$ are the exception since the marginals coincide with β_1 and β_2 and are sufficient to recover β.

Since β_i represents either an element in a product distribution or a marginal, care must be taken when this notation is encountered. The context will clarify which one is meant.

Let's conclude this note by making the meaning of the four most encountered distribution sets in this research explicit:

- *An element α in $\Delta\left(\prod_{i\in\mathcal{I}}\mathcal{A}_i\right) = \Delta(\mathcal{A})$ represents a distribution over the joint action set. In this setting, α_i is the marginal for agent i.*
- *An element $\alpha = \left(\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{I}|}\right)$ in $\prod_{i\in\mathcal{I}}\Delta(\mathcal{A}_i)$ represents a product distributions over the action sets of the agents. It results from each agent independently choosing a distribution over its own action set.*
- *An element α_{-i} in $\Delta\left(\prod_{j\in -i}\mathcal{A}_j\right) = \Delta(\mathcal{A}_{-i})$ represents a distribution over the joint action set of the opponents of agent i.*
- *An element $\alpha_{-i} = \left(\alpha_1, \alpha_2, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_{|\mathcal{I}|}\right)$ in $\prod_{j\in -i}\Delta(\mathcal{A}_j)$ represents a product distribution over the action sets of the opponents*

of agent i. It results from each opponent of agent i independently choosing a distribution over its own action set.

With this definition of the mixed best response correspondence, Nash equilibria can be extended to agents playing mixed actions. To do so, we are going to repeat the Nash equilibrium story making the required changes. Consider a group of rational agents facing a game. Suppose each agent knows the utility function associated with its rational preferences. Each agent is asked to submit a mixed action without discussing with the other agents. Then, each agent is asked, if given the circumstances, it is satisfied with its mixed action. In this context, the circumstances for agent i is the product distributions created by the mixed actions of all the other agents $\alpha_{-i} = \big(\alpha_1, \alpha_2, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_{|\mathcal{I}|}\big) \in \prod_{j \in -i} \Delta(\mathcal{A}_j)$. The answers are all yeses if and only if the joint mixed action is a joint utility maximizer, i.e. a mixed Nash equilibrium. More succintly, a product of independent distributions $\alpha = \big(\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{I}|}\big) \in \prod_{i \in \mathcal{I}} \Delta(\mathcal{A}_i)$ is a mixed Nash equilibrium if and only if for all i in $\mathcal{I}$, $\alpha_i \in \mathrm{BR}_i(\alpha_{-i})$. A mixed Nash equilibrium is often simply called a Nash equilibrium. Similarly, an either pure or mixed Nash equilibrium is called a potentially mixed Nash equilibrium.

Nash proved the following fact which is a cornerstone of game theory. Any game with a finite number of players choosing from finite actions sets has at least one, potentially mixed, Nash equilibrium.

2.3 A Game Example

Let's illustrate the game-theoretic concepts exposed so far on the following game known as battle of the sexes. A couple, composed of a man ♂ and a woman ♀, is planning a date. Each one chooses between two actions: going to a football game F or going to an opera performance O. The joint action of the couple is represented by an ordered pair $\big(a_{♂}, a_{♀}\big)$, where $a_{♂}$ is the action chosen by the man and $a_{♀}$ 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 $u_{♂}$ for the man and $u_{♀}$ for the woman with the following values:

$$u_{♂}(\mathrm{F},\mathrm{F}) = 2, \quad u_{♂}(\mathrm{O},\mathrm{O}) = 1, \quad u_{♂}(\mathrm{F},\mathrm{O}) = 0, \quad u_{♂}(\mathrm{O},\mathrm{F}) = 0,$$
$$u_{♀}(\mathrm{O},\mathrm{O}) = 3, \quad u_{♀}(\mathrm{F},\mathrm{F}) = 2, \quad u_{♀}(\mathrm{F},\mathrm{O}) = 1, \quad u_{♀}(\mathrm{O},\mathrm{F}) = 0.$$

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

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 mixed action of the woman. A mixed action of the woman randomizes between F and O. Since there are only two actions, a single number $p_{♀} \in [0,1]$ is enough to describe the mixed action where she chooses F with probability $p_{♀}$ and O with probability $\big(1 - p_{♀}\big)$. When the context makes the distinction clear, $p_{♀}$ denotes either this probability or the mixed action. Note that she chooses a pure action for $p_{♀} = 0$ or $p_{♀} = 1$. When the woman plays $p_{♀}$, the utility received by the man is

$$u_{♂}\big(\mathrm{F}, p_{♀}\big) = p_{♀} u_{♂}(\mathrm{F}, \mathrm{F}) + \big(1 - p_{♀}\big) u_{♂}(\mathrm{F}, \mathrm{O}) = 2p_{♀},$$

if he plays F, and

$$u_{♂}\big(\mathrm{O}, p_{♀}\big) = p_{♀} u_{♂}(\mathrm{O}, \mathrm{F}) + \big(1 - p_{♀}\big) u_{♂}(\mathrm{O}, \mathrm{O}) = 1 - p_{♀},$$

if he plays O. His optimal action depends on $p_{♀}$ with a critical value of $\frac{1}{3}$. If $p_{♀} > \frac{1}{3}$, he strictly prefers F to O. If $p_{♀} < \frac{1}{3}$, he strictly prefers O to F. If $p_{♀} = \frac{1}{3}$, he is indifferent between F and O; any combination of F and O is a best response to $p_{♀} = \frac{1}{3}$. Grouping all of these preferences yields the following best-response correspondence for the man:

$$\mathrm{BR}_{♂}\big(p_{♀}\big) = \begin{cases} \{\mathrm{F}\} & \text{if } p_{♀} > \frac{1}{3}, \\ \{\mathrm{O}\} & \text{if } p_{♀} < \frac{1}{3}, \\ \{\, p_{♂}\mathrm{F} + \big(1 - p_{♂}\big)\mathrm{O} \,|\, p_{♂} \in [0,1] \,\} & \text{if } p_{♀} = \frac{1}{3}. \end{cases}$$

The best response of the woman to the man's action $p_{♂}$ is computed in a similar fashion. The critical value of $p_{♂}$ making her indifferent is $\frac{3}{4}$ and

the best-response correspondence is the following:

$$\mathrm{BR}_{\female}(p_{\male}) = \begin{cases} \{\mathrm{F}\} & \text{if } p_{\male} > \frac{3}{4}, \\ \{\mathrm{O}\} & \text{if } p_{\male} < \frac{3}{4}, \\ \{\, p_{\female}\mathrm{F} + (1 - p_{\female})\mathrm{O} \,|\, p_{\female} \in [0,1] \,\} & \text{if } p_{\male} = \frac{3}{4}. \end{cases}$$

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 $p_{\male} = \frac{3}{4}$ and $p_{\female} = \frac{1}{3}$.

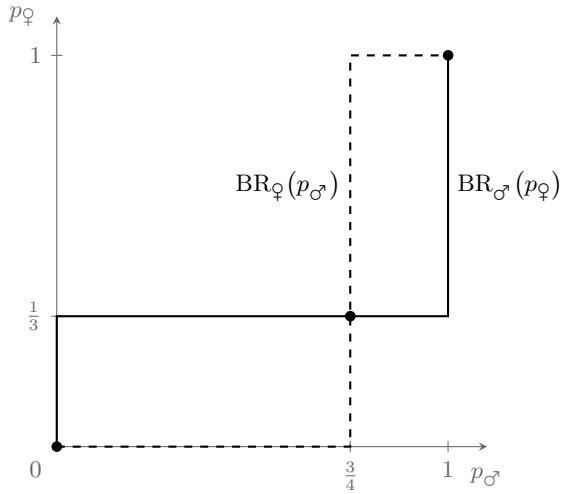


Figure 1. *Best responses and Nash equilibria for battle of the sexes. The man plays* F *with probability* $p_{\male}$*. The woman plays* F *with probability* $p_{\female}$*. The solid line is the man's best response. The dashed line is the woman's best response. The filled circles indicate the Nash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

2.3.1 A Different Story

In battle of the sexes, only one person is truly happy in each pure Nash equilibrium. The man is not thrilled to be at an opera performance, neither is the woman at the football game. The mixed Nash equilibrium seems more fair than the two pure ones. Each person has a chance to go on his or her preferred date. However, they end up in different locations most of the time. Figure 2 illustrates that more than half of the mixed Nash equilibrium distribution is focused on joint actions yielding low utility to both players.

$$\underbrace{\begin{matrix}\frac{3}{4}\mathbf{F}\\ \frac{1}{4}\mathbf{O}\end{matrix}}_{\alpha_{\male}} \times \underbrace{\begin{matrix}\frac{1}{3}\mathrm{F} & \frac{2}{3}\mathrm{O}\end{matrix}}_{\alpha_{\female}} = \underbrace{\begin{matrix}\frac{1}{4}(\mathbf{F},\mathrm{F}) & \frac{1}{2}(\mathbf{F},\mathrm{O})\\ \frac{1}{12}(\mathbf{O},\mathrm{F}) & \frac{1}{6}(\mathbf{O},\mathrm{O})\end{matrix}}_{\alpha} \implies u_{\male} = \tfrac{2}{3}, u_{\female} = \tfrac{3}{2}$$

Figure 2. *Distribution of the mixed Nash equilibrium for the battle of the sexes. The mixed Nash equilibrium α is the product of two independent distributions $\alpha_{\male}$ and $\alpha_{\female}$. More than half of the weight of distribution α is on (F,O) and (O,F) which are low-utility joint actions for both players.*

When facing the kind of incompatible decisions modeled by battle of the sexes, humans sometimes have recourse to a coin toss. The man and the woman agree that on heads they go to the football game and on tails they go to the opera performance. Doing this induces a probability distribution focused on the high-utility joint actions, as illustrated in Figure 3.

$$\begin{matrix}\frac{1}{2}(\mathrm{F},\mathrm{F}) & 0(\mathrm{F},\mathrm{O})\\ 0(\mathrm{O},\mathrm{F}) & \frac{1}{2}(\mathrm{O},\mathrm{O})\end{matrix} \implies u_{\male} = \tfrac{3}{2}, u_{\female} = \tfrac{5}{2}$$

Figure 3. *Distribution for the battle of the sexes using a coin toss. This distribution puts all its weight on high-utility joint actions. The utility for both agents is higher than in the mixed Nash equilibrium.*

Introducing this coin toss in our equilibrium stories goes as follows. Consider two rational agents facing a battle of the sexes. Suppose each agent knows the utility function associated with its rational preferences. The agents agree to flip a coin to decide of their action. They both know that the coin is unbiased and produces heads and tails with probability one half. They both agree to play F if the coin comes out as heads. They both agree to play O if the coin comes out as tails. Each agent is asked two questions. First, given the circumstances, if the coin toss yields heads, are you satisfied with your agreed action F. Second, given the circumstances, if the coin toss yields tails, are you satisfied with your agreed action O. In this context, the circumstances are the distribution over heads and tails induced by the coin and the actions agreed upon by the opponent in case of heads and tails. The answers are all yeses, which is the characteristic of another form of equilibrium introduced by Aumann under the name correlated equilibrium [2,3]. Correlated equilibria will be described formally

in the next section. Observe that, in this example, the correlated equilibrium maintains the fairness of the mixed Nash equilibrium and yields a higher utility for both agents.

However, this course of actions does not induce a Nash equilibrium. Recall that the Nash equilibrium story requires independence in the choice of actions. This distribution over actions cannot be obtained as the product of two independent distributions. Therefore, this is not a Nash equilibrium.

2.4 Equilibria

In the previous sections, Nash and correlated equilibria have been introduced and illustrated. In the present section, these concepts are being more formally defined and some of their properties proven.

A game is described by a set of agents, the action set of each agent, and a utility function for each agent. All this information is encoded in the function $u = (u_i)_{i \in \mathcal{I}}$ defined as follows:

$$\begin{aligned} u \colon \mathcal{A} &\to \mathbb{R}^{|\mathcal{I}|} \\ a &\mapsto \begin{pmatrix} u_1(a) \\ u_2(a) \\ \vdots \\ u_{|\mathcal{I}|}(a) \end{pmatrix}. \end{aligned}$$

Since each agent is trying to maximize its one-time payoff, this type of games is called one-shot games. As mentioned previously, one-shot games allow for mixed actions. In this case, the utility function of each agent is canonically extended through the use of expectation. The following note exposes a few canonical extensions allowing the application of functions to distributions. From now on, these canonical extensions are used implicitly when needed. For example, we will say that $u \colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describes a one-shot game. We will not mention explicitly mixed actions even though they are allowed and their utility unambiguously defined through the canonical extension.

Note 4 (Canonical Extensions and Category Theory)**.** *We defined the utility function for pure actions and later extended it to handle mixed actions through the use of expectation. This extension is not restricted to utility functions. Let $u \colon \mathcal{A} \to \mathcal{K}$, where $\mathcal{K}$ is an $\mathbb{R}$ vector space. The domain of u can be extended to accommodate distributions over $\mathcal{A}$ and yield $\tilde{u} \colon \Delta(\mathcal{A}) \to \mathcal{K}$. For $\alpha \in \Delta(\mathcal{A})$, define $\tilde{u}(\alpha) = \sum_{a \in \mathcal{A}} \alpha[a] u(a)$. This extension only relies on the fact that $\mathcal{K}$ is a vector space on $\mathbb{R}$ and that probabilities are real numbers. In particular, when u represents a utility, the vector space $\mathcal{K}$ is $\mathbb{R}$ itself. Previously, we used the symbol u to represent both the original*

function and the extension. This abuse of notation is common since this is the canonical extension.

We now explicit two other commonly used canonical extensions.

Let f be a function from $\mathcal{A}$ to $\mathcal{B}$. The function f cannot readily be applied to distributions over $\mathcal{A}$. However, it can be extended to a function $\tilde{f}$ on distributions. Function $f\colon \mathcal{A} \to \mathcal{B}$ is extended to function $\tilde{f}\colon \Delta(\mathcal{A}) \to \Delta(\mathcal{B})$. Let α be a distribution over $\mathcal{A}$. The extension works by associating to each element $b \in \mathcal{B}$ a probability equal to the sum of probabilities of its preimages in α, i.e.

$$\forall b \in \mathcal{B},\ \tilde{f}(\alpha)[b] = \sum_{\substack{a \in \mathcal{A}, \\ f(a)=b}} \alpha[a].$$

By abuse of notation, the symbol f represents the function from $\mathcal{A}$ to $\mathcal{B}$, as well as the extension to $\Delta(\mathcal{A})$.

Let g be a function from $\mathcal{A}$ to $\Delta(\mathcal{B})$, and α be a distribution over $\mathcal{A}$. Define $\tilde{g}\colon \Delta(\mathcal{A}) \to \Delta(\mathcal{B})$ such that

$$\forall b \in \mathcal{B},\ \tilde{g}(\alpha)[b] = \sum_{a \in \mathcal{A},} \alpha[a] g(a)[b].$$

as the extension to g. By abuse of notation, the symbol g represents the function from $\mathcal{A}$ to $\Delta(\mathcal{B})$, as well as the extension to $\Delta(\mathcal{A})$. In fact, this derivation can be seen as an application of the extension through expectation, since $\Delta(\mathcal{B})$ is an $\mathbb{R}$ vector space.

The extensions of f and g have been called canonical. Looking at probability distributions through the eye of category theory backs this claim. In a nutshell, category theory is an extension of set theory. Set-theoretical algebraic structures, such as monoids, have category-theoretical counterparts. In particular, two structures shed light on the extensions. The extension of f is explained by the fact that probability distributions form a functor. The extension of g by the fact that they form a monad. The details about functors and monads are beyond the scope of this research, but the interested reader is referred to [4, 5] for more information.

Category theory is mentioned here for two reasons. First, it is a tool making reasoning about probabilities easier. Second, this theoretical tool has practical implications for programming with probability distributions. The programming implications are explored in the following example.

Example 1 (Category Theory in Haskell)**.** *Haskell [6] is a programming language with strong mathematical roots. As such, it is a very good tool for applied mathematics. In particular, category theory is baked at the heart of Haskell; functors and monads are handled natively. Below is a toy example demonstrating this fact.*

Let $\mathcal{A} = \{a_1, a_2\}$ *and* $\mathcal{B} = \{b_1, b_2, b_3\}$ *be two finite sets. Let*

$$\begin{aligned} f\colon \mathcal{A} &\to \mathcal{B} \\ a_1 &\mapsto b_3 \\ a_2 &\mapsto b_1, \end{aligned} \qquad \text{and} \qquad \begin{aligned} g\colon \mathcal{A} &\to \Delta(\mathcal{B}) \\ a_1 &\mapsto b_1 \\ a_2 &\mapsto 0.15\, b_1 + 0.6\, b_2 + 0.25\, b_3 \end{aligned}$$

be two functions with domain $\mathcal{A}$*, and* $\alpha = \frac{1}{5}a_1 + \frac{4}{5}a_2$ *a distribution over* $\mathcal{A}$*.*

This setup is translated in Haskell as follows:

```
data A = A1 | A2
data B = B1 | B2 | B3

f :: A  -> B
f    A1 =  B3
f    A2 =  B1

g :: A  -> Δ(B)
g    A1 =  [(B1, 1)]
g    A2 =  [(B1, 0.15), (B2, 0.6), (B3, 0.25)]

α :: Δ(A)
α =  [(A1, 0.2), (A2, 0.8)]
```

Function application in Haskell is denoted by `$`*. The following shows the result of applying* f *to* a_1 *and* g *to* a_2*:*

```
> f $ A1
B3

> g $ A2
[(B1, 0.15), (B2, 0.6), (B3, 0.25)]
```

Applying a function in a functor context is represented by `<$>`*. Applying it in a monad context is represented by* `=<<`*. With these two notations, the functions* f *and* g *can be applied to the distribution* α *as follows:*

```
> f <$> α
[(B1, 0.8), (B3, 0.2)]

> g =<< α
[(B1, 0.32), (B2, 0.48), (B3, 0.2)]
```

Working with distributions in Haskell is easy. Note in particular that extensions of f *and* g *were not defined by the user. It might seem surprising since Haskell does not know anything about probabilities. However, distributions form a functor and a monad, therefore, Haskell was able to compute*

the extensions automatically. This example shows how readable code is and how closely it follows mathematical notation.

A one-shot game is a model of interacting decision makers. When presented with such a model, two questions come to mind.

The first is a descriptive one. *What happens when rational agents play a game?* Game theory seeks to answer this question by defining and analyzing solution concepts. In economics, stories, similar to the ones mentioned previously, are used to convince the reader that a given solutions concept is appropriate for rational agents. Furthermore, emphasis is placed on characterizing the set of payoffs achievable at equilibrium.

The second question is a prescriptive one. *How should the game and the agents be designed to reach a desired goal?* This question is prevalent in the control and systems' approach to game theory. Solution concept are once again central but are not enough. Algorithms reaching these solution concepts are also needed.

Most answers to these questions are framed in the context of non-cooperative game theory. Non-cooperative game theory is a subset of game theory in which agents are selfish and only interested in maximizing their own utility. The agents understand the impact of other agents through actions. However, they would not think of collaborating with other agents in order to jointly increase their utility. This is why, the other agents are called opponents. In this context, a solution concept corresponds to action profiles with no profitable unilateral deviation, also known as equilibria. In this section, we show how different definitions of profitability give rise to the most common equilibria for one-shot games. To start, we get back to the most basic notion of profitability, best response.

Definition 1 (Best Response). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $i \in \mathcal{I}$ be an agent. The mapping*

$$\begin{aligned} \mathrm{BR}_i\colon \Delta(\mathcal{A}_{-i}) &\rightrightarrows \Delta(\mathcal{A}_i) \\ \alpha_{-i} &\mapsto \operatorname*{arg\,max}_{\alpha_i \in \Delta(\mathcal{A}_i)} u_i(\alpha_i, \alpha_{-i}) \end{aligned}$$

is called agent i's best-response correspondence for u.

The mapping

$$\begin{aligned} \mathrm{BR}\colon \Delta(\mathcal{A}) &\rightrightarrows \prod_{i \in \mathcal{I}} \Delta(\mathcal{A}_i) \\ \alpha &\mapsto \begin{pmatrix} \mathrm{BR}_1(\alpha_{-1}) \\ \mathrm{BR}_2(\alpha_{-2}) \\ \vdots \\ \mathrm{BR}_{|\mathcal{I}|}\left(\alpha_{-|\mathcal{I}|}\right) \end{pmatrix} \end{aligned}$$

is called the joint best-response correspondence for u.

This definition is a case where Note 3 is relevant. In the context of Definition 1, α_{-1} represents the marginal distribution of α for agent 1. Armed with this definition of best response, we can formally define Nash equilibria.

Definition 2 (Nash Equilibrium). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\alpha_i \in \Delta(\mathcal{A}_i)$ be a distribution over action space $\mathcal{A}_i$ for agent i.*

The distribution $\alpha = (\alpha_i)_{i\in\mathcal{I}} \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is a Nash equilibrium for u if and only one of the three following equivalent conditions is verified:

- *The distribution α is a fixed point of the joint best-response correspondence for u, i.e. $\alpha \in \mathrm{BR}(\alpha)$.*
- *For all i in $\mathcal{I}$, $\alpha_i \in \mathrm{BR}_i(\alpha_{-i})$.*
- *For all i in $\mathcal{I}$ and a_i' in $\mathcal{A}_i$, $\mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i})] \geq \mathbb{E}_{A\sim\alpha}[u_i(a_i', A_{-i})]$.*

The second classical solution concept developed for one-shot games is the correlated equilibrium. In the battle of the sexes example, we mentioned that correlated equilibria expand the notion of Nash equilibria from product distributions in $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ to distributions over the joint action space $\Delta(\mathcal{A}) = \Delta\big(\prod_{i\in\mathcal{I}} \mathcal{A}_i\big)$. Before introducing the concept of correlated equilibrium, we need to introduce the closely related concept of correlated-equilibrium distribution. The intuition behind correlated equilibria is made clear in the upcoming Proposition 1.

Definition 3 (Correlated-equilibrium Distribution). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\alpha \in \Delta(\mathcal{A})$ be a distribution over joint actions.*

The distribution α is a correlated-equilibrium distribution for u if

$$\forall i \in \mathcal{I},\ a_i \in \mathcal{A}_i \text{ such that } \alpha_i[a_i] > 0,\ a_i' \in \mathcal{A}_i,$$
$$\mathbb{E}_{A\sim\alpha}[u_i(a_i, A_{-i}) \mid A_i = a_i] \geq \mathbb{E}_{A\sim\alpha}[u_i(a_i', A_{-i}) \mid A_i = a_i].$$

Every Nash equilibrium is a correlated-equilibrium distribution. However, the converse is not true.

Definition 4 (Correlated Equilibrium). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\mathcal{T}_i$ be a set of types for agent i, and $\mathcal{T} = \prod_{i\in\mathcal{I}} \mathcal{T}_i$ be the resulting joint type space. Let $\pi \in \Delta(\mathcal{T})$ be a distribution over joint types. Let $\sigma_i\colon \mathcal{T}_i \to \Delta(\mathcal{A}_i)$ be a strategy for agent i, and σ be the resulting joint strategy. Consider a random variable Θ drawn according to π. Construct the random vector $A = (A_i)_{i\in\mathcal{I}}$ such that for all $i \in \mathcal{I}$, $A_i = \sigma_i(\Theta_i)$. The distribution of A is α, such that, for any joint action $a \in \mathcal{A}$, $\alpha[a] = \sum_{\theta\in\mathcal{T}} \pi[\theta] \cdot \sigma(\theta)[a]$.*

The pair (π, σ) is a correlated equilibrium for u if α is a correlated-equilibrium distribution for u.

The types are sometimes called signals. The key feature separating Nash equilibria from correlated equilibria is the potential correlation of the types. This is the correlation of the types that make the conditional expectation in Definition 3 different from the expectation in Definition 2. In the example of the coin toss for battle of the sexes, the types were actually extremely correlated since they were identical. The following example describes more interesting correlated signals and highlights how the conditional distributions are computed.

Example 2 (Correlated Signals). *Consider the following protocol to generate correlated signals for two agents. A third party, independent of both agents, rolls a die. It then sends a signal to each agent. Agent* 1*'s signal* s_1 *is binary. It tells agent* 1 *if the die's value is Low (1, 2, or 3) or High (4, 5, or 6). Agent* 2*'s signals*$_2$ *is ternary. It tells agent* 2 *if the die's value is Small (1 or 2), Medium (3 or 4), or Large (5 or 6). The signals received by each agent are illustrated in Figure 4.*

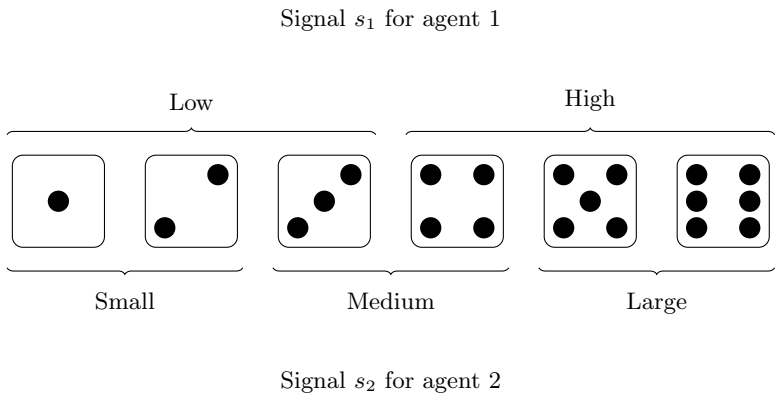


Figure 4. *A pair of coupled signals generated from a die roll. Observing the signal received by agent* 1 *gives some information concerning the signal received by agent* 2*. This information is recovered through the application of Bayes' rule.*

This protocol induces the following joint distribution over the pair (s_1, s_2)*:*

	Low	High
Small	$\frac{1}{3}$	0
Medium	$\frac{1}{6}$	$\frac{1}{6}$
Large	0	$\frac{1}{3}$

.

Suppose both agents know the protocol. Therefore, they know the distribu-

tion of the die and of the pair of signals. When an agent receives a signal it infers something about the signal received by the other one. This inference is done through the application of Bayes' rule. The following facts are inferred:

$$\begin{aligned}
&\mathbb{P}[s_1 = \text{Low} \,|\, s_2 = \text{Small}] = 1, && \mathbb{P}[s_2 = \text{Small} \,|\, s_1 = \text{Low}] = \tfrac{2}{3},\\
&\mathbb{P}[s_1 = \text{High} \,|\, s_2 = \text{Small}] = 0, && \mathbb{P}[s_2 = \text{Medium} \,|\, s_1 = \text{Low}] = \tfrac{1}{3},\\
& && \mathbb{P}[s_2 = \text{Large} \,|\, s_1 = \text{Low}] = 0,\\
&\mathbb{P}[s_1 = \text{Low} \,|\, s_2 = \text{Medium}] = \tfrac{1}{2}, && \\
&\mathbb{P}[s_1 = \text{High} \,|\, s_2 = \text{Medium}] = \tfrac{1}{2}, && \\
& && \mathbb{P}[s_2 = \text{Small} \,|\, s_1 = \text{High}] = 0,\\
&\mathbb{P}[s_1 = \text{Low} \,|\, s_2 = \text{Large}] = 0, && \mathbb{P}[s_2 = \text{Medium} \,|\, s_1 = \text{High}] = \tfrac{1}{3},\\
&\mathbb{P}[s_1 = \text{High} \,|\, s_2 = \text{Large}] = 1, && \mathbb{P}[s_2 = \text{Large} \,|\, s_1 = \text{High}] = \tfrac{2}{3}.
\end{aligned}$$

The distinction between correlated equilibria and correlated-equilibrium distributions is not emphasized in the literature. This blurring is due to a previously mentioned fact; the main focus of game theory in economics is to determine the set of payoffs achievable at equilibrium. Since a correlated equilibrium is defined by the correlated-equilibrium distribution it induces, the payoff sets of the two concepts are identical. The following proposition reinforces that point and gives useful characterizations of correlated equilibria.

Proposition 1 (Characterization of Correlated Equilibria). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\mathcal{T}_i$ be a set of types for agent i, and $\mathcal{T} = \prod_{i\in\mathcal{I}} \mathcal{T}_i$ be the resulting joint type space. Let $\pi \in \Delta(\mathcal{T})$ be a distribution over joint types. Let $\sigma_i\colon \mathcal{T}_i \to \Delta(\mathcal{A}_i)$ be a strategy for agent i, and σ be the resulting joint strategy.*

The pair (π, σ) is a correlated equilibrium for u if and only if one of the following three equivalent conditions is true:

(i) $\forall i \in \mathcal{I}$, $\theta_i \in \mathcal{T}_i$, $a'_i \in \mathcal{A}_i$,

$$\mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\theta_i), \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i] \geq \mathbb{E}_{\Theta\sim\pi}[u_i(a'_i, \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i]$$

(ii) $\forall i \in \mathcal{I}$, $\sigma'_i\colon \mathcal{T}_i \to \Delta(\mathcal{A}_i)$,

$$\mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))] \geq \mathbb{E}_{\Theta\sim\pi}[u_i(\sigma'_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))]$$

(iii) $\forall i \in \mathcal{I}$, $\varphi_i\colon \mathcal{A}_i \to \mathcal{A}_i$,

$$\mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))] \geq \mathbb{E}_{\Theta\sim\pi}[u_i(\varphi_i \circ \sigma_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))]$$

Condition (i), introduced in [2], tests that for every signal, the action prescribed by the strategy is optimal. This is the definition used in the battle of the sexes example. Condition (ii), introduced in [7], tests the

strategy against all the other possible strategies. Condition (iii), introduced in [3], tests the strategy against the restricted set of swap strategies. Swap functions φ_is make recommendations based on the action prescribed by the strategies and not the full signal received.

Proof. The proof is split in the four following steps:

1. Condition (i) implies condition (ii).

2. Condition (ii) implies condition (iii).

3. Any pair (π, σ) verifying condition (iii) is a correlated equilibrium, i.e. induces a correlated-equilibrium distribution.

4. Any correlated-equilibrium distribution α can be induced by a pair (π, σ) satisfying condition (i).

The actual proofs are the following:

1. Let (π, σ) be a pair satisfying condition (i). Let $i \in \mathcal{I}$, $\sigma_i' \colon \mathcal{T}_i \to \Delta(\mathcal{A}_i)$, $\theta_i \in \mathcal{T}_i$, and $a_i' \in \mathcal{A}_i$. By definition, the following inequality holds:

$$\mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\theta_i), \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i] \geq \mathbb{E}_{\Theta\sim\pi}[u_i(a_i', \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i].$$

Multiply both sides of the inequality by $\mathbb{P}_{\Theta\sim\pi}[\Theta_i = \theta_i]$, then sum over θ_i in $\mathcal{T}_i$. The left-hand side becomes

$$\sum_{\theta_i \in \mathcal{T}_i} \mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\theta_i), \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i] \cdot \mathbb{P}_{\Theta\sim\pi}[\Theta_i = \theta_i] = \mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))].$$

Similarly, the right hand side yields

$$\sum_{\theta_i \in \mathcal{T}_i} \mathbb{E}_{\Theta\sim\pi}[u_i(a_i', \sigma_{-i}(\Theta_{-i})) \,|\, \Theta_i = \theta_i] \cdot \mathbb{P}_{\Theta\sim\pi}[\Theta_i = \theta_i] = \mathbb{E}_{\Theta\sim\pi}[u_i(a_i', \sigma_{-i}(\Theta_{-i}))].$$

Therefore,

$$\mathbb{E}_{\Theta\sim\pi}[u_i(\sigma_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))] \geq \mathbb{E}_{\Theta\sim\pi}[u_i(a_i', \sigma_{-i}(\Theta_{-i}))].$$

This inequality holds for any a_i'. Therefore, the left-hand side is greater than any convex combinations of the right-hand side. In particular, it

is greater than

$$\sum_{a'_i \in \mathcal{A}_i} \mathbb{E}_{\Theta\sim\pi}[u_i(a'_i, \sigma_{-i}(\Theta_{-i}))] \cdot \mathbb{P}_{\Theta\sim\pi}[\sigma'_i(\Theta_i) = a'_i] = \mathbb{E}_{\Theta\sim\pi}[u_i(\sigma'_i(\Theta_i), \sigma_{-i}(\Theta_{-i}))],$$

which yields condition (ii).

2. Let (π, σ) be a pair satisfying condition (ii). Let $i \in \mathcal{I}$ and $\varphi_i \colon \mathcal{A}_i \to \mathcal{A}_i$. Using the functor extension from Note 4, the composition $\varphi_i \circ \sigma_i$ is an element of $\mathcal{T}_i \to \Delta(\mathcal{A}_i)$. Therefore, condition (iii) is a direct consequence of condition (ii).

3. Let (π, σ) be a pair satisfying condition (iii). Let Θ be a random variable drawn according to π and A the induced random variable over the joint action set. Denote by α the distribution of A. Let $i \in \mathcal{I}$, $a_i \in \mathcal{A}_i$ such that $\alpha_i[a_i] > 0$, and $a'_i \in \mathcal{A}_i$. Define φ by $\varphi(a_i) = a'_i$ and $\varphi(a''_i) = a''_i$ for all $a''_i \in \mathcal{A}_i \setminus \{a_i\}$. Condition (iii) translates to

$$\mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i})] \geq \mathbb{E}_{A\sim\alpha}[u_i(\varphi_i(A_i), A_{-i})].$$

Applying the law of total probability for the left-hand side yields

$$\mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i}) \mid A_i = a_i] \cdot \mathbb{P}_{A\sim\alpha}[A_i = a_i] + \mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i}) \mid A_i \neq a_i] \cdot \mathbb{P}_{A\sim\alpha}[A_i \neq a_i],$$

or, more concisely,

$$\mathbb{E}_{A\sim\alpha}[u_i(a_i, A_{-i}) \mid A_i = a_i] \cdot \alpha_i[a_i] + \mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i}) \mid A_i \neq a_i] \cdot (1 - \alpha_i[a_i]).$$

Similarly, by using the definition of φ_i and the law of total probability, the right-hand side is equal to

$$\mathbb{E}_{A\sim\alpha}[u_i(a'_i, A_{-i}) \mid A_i = a_i] \cdot \alpha_i[a_i] + \mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i}) \mid A_i \neq a_i] \cdot (1 - \alpha_i[a_i]).$$

Subtracting $\mathbb{E}_{A\sim\alpha}[u_i(A_i, A_{-i}) \mid A_i \neq a_i] \cdot (1 - \alpha_i[a_i])$ from each side and then dividing by $\alpha[a_i]$, which is positive, gives the correlated-equilibrium distribution condition for α.

4. Let α be a correlated-equilibrium distribution. For agent $i \in \mathcal{I}$, define the types to be its actions, $\mathcal{T}_i = \mathcal{A}_i$, and σ_i to be the identity function over $\mathcal{A}_i$. Furthermore, let $\pi = \alpha$. By definition, the distribution α satisfies the inequality in Definition 3. This immediately translates in the pair (π, σ) verifying condition (i).

□

To conclude this section, we look at equilibria from a different perspective. In a game setting, the actions taken by a set of agents induce a distribution $\alpha \in \Delta(\mathcal{A})$ over the joint action set. Equilibria are distributions in which no rational agent has an incentive to unilaterally deviate. This means that defining an equilibrium concept boils down to choosing the two following elements:

- A set of feasible distributions over the joint action set.
- A set of incentive constraints for each agent.

For example, a Nash equilibrium is a product distribution with the simple incentive constraint from Definition 2. Similarly, a correlated-equilibrium distribution is an unrestricted distribution with the conditional incentive constraint from Definition 3.

It is natural to wonder what happens when looking at different combinations of feasible distributions and incentive constraints. On the one hand, a product distribution with the conditional incentive constraint is a Nash equilibrium, since the independence renders the conditional superfluous. On the other hand, an unrestricted distribution with the simple incentive constraint yields a new equilibrium concept called a coarse correlated equilibrium.

Definition 5 (Coarse Correlated Equilibrium)**.** *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\alpha \in \Delta(\mathcal{A})$ be a distribution over joint actions.*

The distribution α is a coarse correlated equilibrium for u if

$$\forall i \in \mathcal{I},\ \forall a_i' \in \mathcal{A}_i,\ \mathbb{E}_{A\sim\alpha}[u_i(A)] \geq \mathbb{E}_{A\sim\alpha}[u_i(a_i', A_{-i})].$$

Every correlated equilibrium distribution is a coarse correlated equilibrium. Indeed, the inequality defining a coarse correlated equilibrium follows from the one defining a correlated-equilibrium distributions by multiplying each side of the inequality by $\mathbb{P}_{A\sim\alpha}[A_i = a_i]$ and summing over $a_i \in \mathcal{A}_i$. This observation, along with the relationship between Nash equilibria and correlated-equilibrium distributions, is captured in the following proposition.

Proposition 2 (Hierarchy of Equilibria)**.** *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\alpha \in \Delta(\mathcal{A})$ be a distribution over joint actions.*

If α is a Nash equilibrium for u, then α is also a correlated-equilibrium distribution for u.

If α is a correlated-equilibrium distribution for u, then α is also a coarse correlated equilibrium for u.

With the standard abbreviations for Nash equilibria (NE), correlated-equilibrium distributions (CE), and coarse correlated equilibria (CCE), this proposition is written concisely as follows:

$$\text{NE} \subset \text{CE} \subset \text{CCE}.$$

This last proposition explains the importance of knowing that Nash equilibria always exist. Indeed, the existence of Nash equilibria implies the existence of correlated-equilibrium distributions and coarse correlated equilibria. Therefore, the following section is dedicated to the proof of Nash's seminal result. Elements of this proof will be used to prove the existence of the solution concept introduced in this research.

2.5 Nash's Existence Theorem

The existence of Nash equilibria was previously mentioned. Here is the formal statement of this result.

Theorem 1 (Nash's Existence [8]). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game.*

There exists a product distribution $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ which is a Nash equilibrium for u.

To prove it, we will use the definition of Nash equilibria in term of fixed points. The best-response correspondence restricted to product distributions $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is as an element of $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i) \rightrightarrows \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$. The set $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is a product of finite simplices and as such is non-empty, compact and a convex subset of an Euclidean space. Brouwer's fixed-point theorem is a classical result guaranteeing the existence of fixed point for functions over this kind of sets.

Theorem 2 (Brouwer's Fixed-point Theorem). *Let $\mathcal{X}$ be a non-empty, compact and convex subset of some Euclidean space. Let $f\colon \mathcal{X} \to \mathcal{X}$ be a continuous function. Then f has a fixed point x^* such that $x^* = f(x^*)$.*

An illustration of this theorem is given in Figure 5. Kakutani's fixed point theorem is an extension to Brouwer's fixed-point theorem dealing with correspondences instead of functions. Figure 6 provides the corresponding illustration.

Theorem 3 (Kakutani's Fixed-point Theorem). *Let $\mathcal{X}$ be a non-empty, compact and convex subset of some Euclidean space. Let $f\colon \mathcal{X} \rightrightarrows \mathcal{X}$ be a correspondence on $\mathcal{X}$ with a closed graph and the property that for all $x \in \mathcal{X}$, $f(x)$ is non empty and convex. Then f has a fixed point, x^* such that $x^* \in f(x^*)$.*

The following definition explicits what a closed graph for a correspondence means.

Definition 6 (Correspondence with a Closed Graph). *Let $\mathcal{X}$ be a non-empty, compact and convex subset of some Euclidean space. Let $f\colon \mathcal{X} \rightrightarrows \mathcal{X}$ be a correspondence on $\mathcal{X}$.*

The graph of f is closed if and only if, for all converging sequence $((x^t, y^t))_{t\in\mathbb{N}}$, such that $y^t \in f(x^t)$, with limit point (x^*, y^*), $y^* \in f(x^*)$.

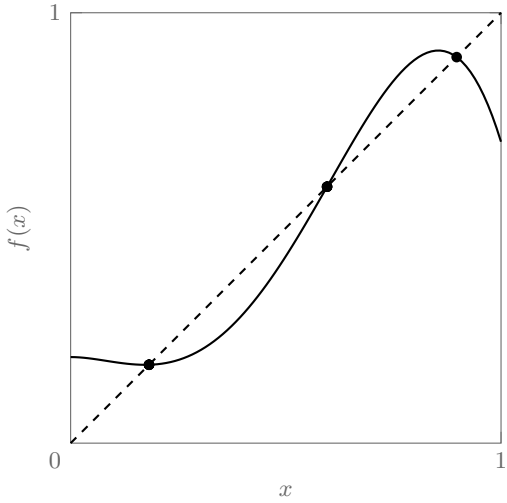


Figure 5. *Illustration of Brouwer's fixed-point theorem. The solid line is the graph of a continuous function f from the interval $[0, 1]$ to itself. The interval $[0, 1]$ is a non-empty, compact and convex subset of the Euclidean space $\mathbb{R}$. The dashed line is the graph of the identity function on the same interval. The filled circles correspond to fixed points of f.*

Brouwer's fixed-point theorem uses continuous functions on which we have a strong grasp. It is a good stepping stone to understand Kakutani's fixed-point theorem. The core elements of these two fixed-point theorems are present in Brouwer's theorem. Kakutani's theorem irons out the details for correspondences. Similarly, Theorem 1 is proven in two step. First, we use a function approximating the best-response correspondence and apply Brouwer's theorem to prove the existence of approximate Nash equilibria. Then, we make the necessary adjustments to apply the full-fledged Kakutani's theorem and prove the existence of exact Nash equilibria. This approach helps building insight about the Nash existence theorem.

2.5.1 Existence of Approximate Nash Equilibria

Every equilibrium definition includes an incentive constraint, which takes the form of a maximization. Sometimes, the exact maximization is not required, and approximate maximization is acceptable. For example, the definition of an approximate Nash equilibrium is the following. Let $\varepsilon > 0$ be a small additive factor. A product distribution $\alpha \in \prod_{i \in \mathcal{I}} \Delta(\mathcal{A}_i)$ is an ε Nash equilibrium for u if

$$\forall i \in \mathcal{I},\, a'_i \in \mathcal{A}_i,\, \mathbb{E}_{A \sim \alpha}[u_i(A_i, A_{-i})] \geq \mathbb{E}_{A \sim \alpha}[u_i(a'_i, A_{-i})] - \varepsilon.$$

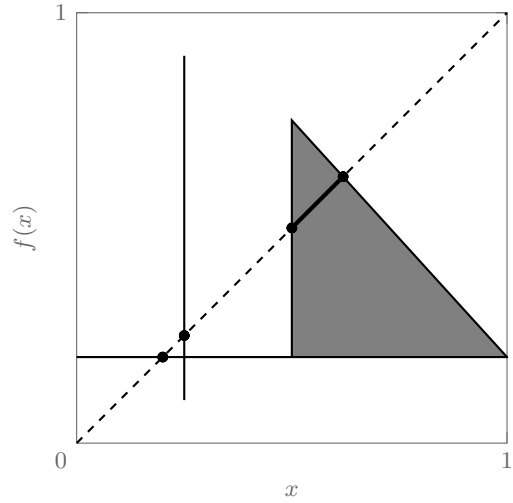


Figure 6. *Illustration of Kakutani's fixed-point theorem. The solid lines and the shaded area represent the closed graph of a correspondence f from the interval $[0,1]$ to itself. For all $x \in [0,1]$, $f(x)$ is non empty and convex. The interval $[0,1]$ is a non-empty, compact and convex subset of the Euclidean space $\mathbb{R}$. The dashed line is the graph of the identity function on the same interval. The filled circles and the bold segment correspond to fixed points of f.*

Correlated equilibria and coarse correlated equilibria are similarly extended to ε correlated equilibrium and ε coarse correlated equilibria by relaxing the incentive constraints by ε.

To prove the existence of an approximate Nash equilibrium, we use Brouwer's fixed-point theorem for a Gibbs-smoothed best-response function.

Definition 7 (Gibbs-smoothed Best Response)**.** *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $i \in \mathcal{I}$ be an agent. The function*

$$\mathrm{GBR}_i^\tau \colon \Delta(\mathcal{A}_{-i}) \to \Delta(\mathcal{A}_i),$$

such that, for all $\alpha_{-i} \in \Delta(\mathcal{A}_{-i})$ and $a_i \in \mathcal{A}_i$,

$$\mathrm{GBR}_i^\tau(\alpha_{-i})[a_i] = \frac{\mathrm{e}^{\frac{1}{\tau}u(a_i,\alpha_{-i})}}{\sum_{a_i' \in \mathcal{A}_i} \mathrm{e}^{\frac{1}{\tau}u\left(a_i',\alpha_{-i}\right)}},$$

is called agent i's Gibbs-smoothed best response with parameter τ for u.

The following note gives a little background about the Gibbs distribution.

Note 5 (Gibbs Distribution). *The Gibbs distribution arises in statistical physics. Consider a system made of a large number of particles. The system, as a whole, can take configurations from the set $\mathcal{X}$. In configuration $x \in \mathcal{X}$, the energy of the system is $E(x)$. Nature seeks to minimize the energy of the system. However, the presence of thermal noise creates a stochastic disturbance in this minimization process. The Gibbs distribution characterizes this disturbance. Let T be the temperature of the system. The probability that the system is in configuration x is $G_T[x] = \frac{\mathrm{e}^{-\frac{1}{\mathrm{k}T}E(x)}}{\sum_{x' \in \mathcal{X}} \mathrm{e}^{-\frac{1}{\mathrm{k}T}E(x')}}$, where k is the Boltzmann constant. For any non-zero temperature, the Gibbs distribution assigns a positive probability to every configuration. As the temperature goes to infinity, the distribution converges to the uniform distribution over $\mathcal{X}$. As the temperature decreases, the distribution puts more weight on the configurations of minimal energy. In the limit, as the temperature approaches zero, the Gibbs distribution converges to the uniform distribution over the configurations of minimal energy.*

By making a couple of changes, the Gibbs distribution is relevant for decision making. In a decision-making problem, the agent seeks an action a maximizing its utility function u. Therefore, the following Gibbs-shaped distribution is of interest $G_\tau[a] = \frac{\mathrm{e}^{\frac{1}{\tau}u(a)}}{\sum_{a' \in \mathcal{A}} \mathrm{e}^{\frac{1}{\tau}u(a')}}$. As the parameter τ goes to zero, distribution G_τ concentrates its weight on utility-maximizing actions. This property explains why, for small τ, the Gibbs distribution is used to define a function approximating the best-response correspondence.

The following proposition formalizes the fact that the Gibbs distribution approaches an optimal distribution as the parameter goes to zero.

Proposition 3 (Approximate Optimality of the Gibbs Distribution). *Let $u\colon \mathcal{A} \to \mathbb{R}$ be a utility function over finite action set $\mathcal{A}$ with cardinality $n = |\mathcal{A}|$. Let $\mathcal{A}^* = \arg\max_{a \in \mathcal{A}} u(a)$ be the set of maximizers of u with cardinality $n^* = |\mathcal{A}^*|$. Define the four following quantities:*

$$u_{\max} = \max_{a \in \mathcal{A}} u(a), \qquad u_\Delta = u_{\max} - u_{\min},$$
$$u_{\min} = \min_{a \in \mathcal{A}} u(a), \qquad u_\delta = u_{\max} - \max_{a \in \mathcal{A} \setminus \mathcal{A}^*} u(a).$$

Let $0 < \varepsilon < u_\Delta\left(\frac{n}{n^*} - 1\right)$, $0 \leq \tau \leq \frac{u_\delta}{\ln\left(\frac{u_\Delta}{\varepsilon}\left(\frac{n}{n^*}-1\right)\right)}$, and α the Gibbs distribution with parameter τ, i.e. for $a \in \mathcal{A}$, $\alpha[a] = \frac{\mathrm{e}^{\frac{1}{\tau}u(a)}}{\sum_{a' \in \mathcal{A}} \mathrm{e}^{\frac{1}{\tau}u(a')}}$.

The distribution α is ε optimal for u.

Proof. We only consider the case where u is not a constant function. Therefore, the set $\mathcal{A}^*$ is strictly included in $\mathcal{A}$ which guarantees $u_\Delta\left(\frac{n}{n^*} - 1\right) > 0$, and $u_\delta > 0$. This, in turn, guarantees the existence of ε and τ satisfying

the aforementioned constraints. If the utility function is constant, any distribution is optimal and therefore ε optimal for u.

The proof is straightforward. We compute $u(\alpha)$, compare it to $u_{\max}$, and show that the difference $u_{\max} - u(\alpha)$ is smaller than ε. Let's first, explicit $u(\alpha)$ and split the optimal actions from the rest of them,

$$
\begin{aligned}
u(\alpha) &= \sum_{a\in\mathcal{A}} \frac{\mathrm{e}^{\frac{1}{\tau}u(a)}}{\sum_{a'\in\mathcal{A}} \mathrm{e}^{\frac{1}{\tau}u(a')}} u(a) \\
&= \sum_{a\in\mathcal{A}} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a)-u_{\max}]}}{\sum_{a'\in\mathcal{A}} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u(a) \\
&= \sum_{a\in\mathcal{A}} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a)-u_{\max}]}}{\sum_{a'\in\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u_{\max}-u_{\max}]} + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u(a) \\
&= \sum_{a\in\mathcal{A}} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a)-u_{\max}]}}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u(a) \\
&= \sum_{a\in\mathcal{A}^*} \frac{1}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u_{\max} + \\
&\qquad\qquad \sum_{a\in\mathcal{A}\setminus\mathcal{A}^*} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a)-u_{\max}]}}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u(a) \\
&= \underbrace{\frac{n^*}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u_{\max}}_{u_{\text{opt}}} + \\
&\qquad\qquad \underbrace{\sum_{a\in\mathcal{A}\setminus\mathcal{A}^*} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a)-u_{\max}]}}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u(a)}_{u_{\text{rest}}} .
\end{aligned}
$$

Rewrite $u_{\max}$ as a sum with a similar structure,

$$
\begin{aligned}
u_{\max} &= \frac{n^*}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u_{\max} + \\
&\qquad\qquad \frac{\sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}\left[u\left(a'\right)-u_{\max}\right]}}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u_{\max} \\
&= u_{\text{opt}} + \frac{\sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}\left[u\left(a'\right)-u_{\max}\right]}}{n^* + \sum_{a'\in\mathcal{A}\setminus\mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a')-u_{\max}]}} u_{\max}.
\end{aligned}
$$

Therefore,

$$\begin{aligned}
u_{\max} - u(\alpha) &= u_{\max} - u_{\text{rest}} \\
&= \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a) - u_{\max}]}}{n^* + \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a') - u_{\max}]}} (u_{\max} - u(a)) \\
&\leq \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a) - u_{\max}]}}{n^* + \sum_{a' \in \mathcal{A} \setminus \mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a') - u_{\max}]}} u_\Delta \\
&\leq \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \frac{\mathrm{e}^{\frac{1}{\tau}[u(a) - u_{\max}]}}{n^*} u_\Delta \\
&= \frac{u_\Delta}{n^*} \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \mathrm{e}^{\frac{1}{\tau}[u(a) - u_{\max}]} \\
&\leq \frac{u_\Delta}{n^*} \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \mathrm{e}^{-\frac{1}{\tau} u_\delta} \\
&\leq \frac{u_\Delta}{n^*} (n - n^*) \mathrm{e}^{-\frac{1}{\tau} u_\delta} \\
&= u_\Delta \Big(\frac{n}{n^*} - 1 \Big) \mathrm{e}^{-\frac{1}{\tau} u_\delta}.
\end{aligned}$$

Expand the inequality satisfied by τ as follows:

$$\begin{aligned}
\tau &\leq \frac{u_\delta}{\ln\big(\frac{u_\Delta}{\varepsilon}\big(\frac{n}{n^*} - 1\big)\big)} \\
\frac{1}{\tau} &\geq \frac{1}{u_\delta} \ln\Big(\frac{u_\Delta}{\varepsilon} \Big(\frac{n}{n^*} - 1 \Big) \Big) \\
-\frac{u_\delta}{\tau} &\leq \ln\left(\frac{\varepsilon}{u_\Delta\big(\frac{n}{n^*} - 1\big)} \right) \\
\mathrm{e}^{-\frac{1}{\tau} u_\delta} &\leq \frac{\varepsilon}{u_\Delta\big(\frac{n}{n^*} - 1\big)}.
\end{aligned}$$

Plugging the result in $u_{\max} - u(\alpha) \leq u_\Delta\big(\frac{n}{n^*} - 1\big)\mathrm{e}^{-\frac{1}{\tau} u_\delta}$ proves that α is ε optimal for u. □

Combining Definition 7 and Proposition 3 yields one possible definition for an approximate best-response function.

Definition 8 (Approximate Best Response). *Let* $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ *describe a*

one-shot game. Define the following quantities, for agent $i \in \mathcal{I}$:

$$\begin{aligned}
\mathcal{A}_i^*(a_{-i}) &= \arg\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}),\\
n_i &= |\mathcal{A}_i|,\\
n_i^* &= \min_{a_{-i} \in \mathcal{A}_{-i}} |\mathcal{A}_i^*(a_{-i})|,\\
u_i^\Delta &= \max_{a_{-i} \in \mathcal{A}_{-i}} \left[\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}) - \min_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}) \right],\\
u_i^\delta &= \min_{a_{-i} \in \mathcal{A}_{-i}} \left[\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}) - \max_{a_i \in \mathcal{A}_i \setminus \mathcal{A}_i^*(a_{-i})} u_i(a_i, a_{-i}) \right].
\end{aligned}$$

Let $0 < \varepsilon < \min_{i \in \mathcal{I}} u_i^\Delta \left(\frac{n_i}{n_i^*} - 1 \right)$. From now on, we will refer to this technical condition as ε small enough for u. *Define agent i's ε best-response function for u by $\mathrm{BR}_i^\varepsilon = \mathrm{GBR}_i^{\tau_i}$, for $\tau_i = \frac{u_i^\delta}{\ln\left(\frac{u_i^\Delta}{\varepsilon} \left(\frac{n_i}{n_i^*} - 1 \right) \right)}$. As the name indicates, for $\alpha_{-i} \in \Delta(\mathcal{A}_{-i})$, the distribution $\mathrm{BR}_i^\varepsilon(\alpha_{-i})$ is ε optimal for $u_i(\cdot, \alpha_{-i})$. Accordingly, the function*

$$\begin{aligned}
\mathrm{BR}^\varepsilon \colon \Delta(\mathcal{A}) &\to \prod_{i \in \mathcal{I}} \Delta(\mathcal{A}_i)\\
\alpha &\mapsto \begin{pmatrix} \mathrm{BR}_1^\varepsilon(\alpha_{-1}) \\ \mathrm{BR}_2^\varepsilon(\alpha_{-2}) \\ \vdots \\ \mathrm{BR}_{|\mathcal{I}|}^\varepsilon\left(\alpha_{-|\mathcal{I}|}\right) \end{pmatrix}
\end{aligned}$$

is called the joint ε best response for u.

By definition, the joint ε best response is approximately optimal. This is the first condition required for proving the existence of approximate Nash equilibria. The second condition needed is its continuity.

Proposition 4 (Continuity of the Approximate Best-response Function)**.** *Let $u \colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game and ε small enough for u.*

The joint ε best response for u is continuous.

Proof. The function BR^ε is vector valued. It is continuous if and only if it is componentwise continuous. Let $i \in \mathcal{I}$ be an agent. We need to prove that $\alpha \mapsto \mathrm{BR}_i^\varepsilon(\alpha_{-i})$ is continuous. The function $\alpha \mapsto \alpha_{-i}$ is continuous. Therefore, it is sufficient to prove that $\mathrm{BR}_i^\varepsilon$ is continuous in order to prove that BR^ε is continuous.

The function $\mathrm{BR}_i^\varepsilon$ takes values in $\Delta(\mathcal{A}_i)$. It can be interpreted as a vector-valued function, with as many entries as elements in $\mathcal{A}_i$. Therefore, it is sufficient to prove that $\alpha_{-i} \mapsto \mathrm{BR}_i^\varepsilon(\alpha_{-i})[a_i]$ is continuous for a fixed $a_i \in \mathcal{A}_i$.

By definition

$$\mathrm{BR}_i^{\varepsilon}(\alpha_{-i})[a_i] = \frac{\mathrm{e}^{\frac{1}{\tau_i}u_i(a_i,\alpha_{-i})}}{\sum_{a_i'\in\mathcal{A}_i}\mathrm{e}^{\frac{1}{\tau_i}u_i\left(a_i',\alpha_{-i}\right)}}.$$

The mapping is therefore continuous as it is a composition of continuous functions: expectation, division, exponential, and sum. □

With the definition and proposition in place, we can now prove the existence of approximate Nash equilibria.

Theorem 4 (Existence of Approximate Nash Equilibria). *Let* $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ *describe a one-shot game and* $\varepsilon > 0$.

The exists an ε *Nash equilibrium for* u.

Proof. A distribution $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ forming a fixed point of the joint ε best response for u, i.e. verifying $\alpha = \mathrm{BR}^{\varepsilon}(\alpha)$, is an ε Nash equilibrium for u. Therefore, proving the existence of such a fixed point is a sufficient condition to proving the theorem.

As was previously mentioned, the set $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is a product of finite simplices and as such is non-empty, compact and a convex subset of an Euclidean space. The joint ε best response for u is continuous. Therefore, by applying Brouwer's fixed-point theorem, we conclude that such a fixed point exist. □

2.5.2 Existence of Exact Nash Equilibria

This section is dedicated to the proof of Nash's existence theorem. The proof of Theorem 4 gives some intuition regarding the existence of Nash equilibria. The following proof contains the details.

Proof of Theorem 1. A distribution $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ forming a fixed point of the joint best response for u, i.e. verifying $\alpha = \mathrm{BR}(\alpha)$, is a Nash equilibrium for u. Therefore, proving the existence of such a fixed point is a sufficient condition to proving the theorem.

To apply Kakutani's fixed point theorem we need to prove the four following facts:

- The set $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is non-empty, compact and a convex subset of an Euclidean space.
- For $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$, $\mathrm{BR}(\alpha)$ is non-empty.
- For $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$, $\mathrm{BR}(\alpha)$ is convex.
- The best-response correspondence has a closed graph.

The first two facts are immediately proven. The set $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ is a product of finite simplices and as such is non-empty, compact and a convex subset of an Euclidean space. Let $\alpha \in \prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$. There exists a best response for each agent which guarantees that $\mathrm{BR}(\alpha)$ is non-empty.

We now prove that the set $\mathrm{BR}(\alpha)$ is convex, for α a distribution in $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$. Let i be an agent, β_i and γ_i be two elements of $\mathrm{BR}_i(\alpha_{-i})$, and $\theta \in [0, 1]$. Since β_i and γ_i are both best responses to α_{-i}, it is the case that $u_i(\beta_i, \alpha_{-i}) = u_i(\gamma_i, \alpha_{-i})$. By linearity of the expectation, we conclude that

$$u_i(\beta_i, \alpha_{-i}) = u_i(\theta\beta_i + (1-\theta)\gamma_i, \alpha_{-i}) = u_i(\gamma_i, \alpha_{-i}).$$

As a result, the convex combination $\theta\beta_i + (1-\theta)\gamma_i$ is also an element of $\mathrm{BR}_i(\alpha_{-i})$. Therefore, the set $\mathrm{BR}_i(\alpha_{-i})$ is convex. The set $\mathrm{BR}(\alpha)$ is the Cartesian product of convex sets and as such is convex.

We finally prove that the best-response correspondence has a closed graph. Let $\alpha = (\alpha^t)_{t\in\mathbb{N}}$ and $\beta = (\beta^t)_{t\in\mathbb{N}}$ be sequences in $\prod_{i\in\mathcal{I}} \Delta(\mathcal{A}_i)$ such that for all t in $\mathbb{N}$, $\beta^t \in \mathrm{BR}(\alpha^t)$. Suppose that α converges to α^* and β converges to β^*. Let i be an agent and a_i an action for this agent. For $t \in \mathbb{N}$, the fact that $\beta^t \in \mathrm{BR}(\alpha^t)$ implies that $\beta_i^t \in \mathrm{BR}_i\big(\alpha_{-i}^t\big)$. This translates to $u_i\big(\beta_i^t, \alpha_{-i}^t\big) \geq u_i\big(a_i, \alpha_{-i}^t\big)$. The utility function is continuous in the joint action. Therefore, in the limit $u_i\big(\beta_i^*, \alpha_{-i}^*\big) \geq u_i\big(a_i, \alpha_{-i}^*\big)$ which proves that $\beta_i^* \in \mathrm{BR}_i\big(\alpha_{-i}^*\big)$. This fact is true for any agent and therefore BR has a closed graph.

Everything is now in place to apply Kakutani's fixed-point theorem and to conclude that an exact Nash equilibrium always exists. □

At first glance, it is easier to prove the existence of an exact equilibrium rather than an approximate one. However, note that most of the approximate equilibrium section deals with defining the ε best response. The actual proof is shorter and uses a simpler fixed-point theorem.

3 Dynamic Game Theory

3.1 Markov Decision Processes

3.1.1 Setup

The problems analyzed up to this point were static; there was no notion of time. We are now switching gears and turning to problems with 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. The dynamic is described as follows:

$$\forall t \in \mathbb{N},\, x^{t+1} \sim f\big(x^t, a^t\big),$$

where x^t and x^{t+1} are states in a finite state space $\mathcal{X}$, a^t is an action in a finite action set $\mathcal{A}$, and f is a state-transition function in $\mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$. This dynamic is alternatively represented by the short notation

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

where x and a are the state at the action at a given time step and x^+ is the state at the next time step.

At each time step t, the agent observes the state and chooses an action. Over time, the agent accumulates some information. This sequence of states and actions is called the history. The history up to time t is $h^t = \big(x^0, a^0, x^1, a^1, \ldots, x^{t-1}, a^{t-1}, x^t\big)$. Denote by $\mathcal{H}^t$ the set of histories up to time t and by $\mathcal{H} = \cup_{t\in\mathbb{N}}\mathcal{H}^t$ the set of all possible histories. The information observable over an infinite run is called an infinite history. The set of infinite histories is denoted by $\mathcal{H}^\infty$.

In state x, choosing action a yields a payoff $v(x, a)$. The agent is interested in maximizing its expected sum of discounted payoffs for a given discount factor $\delta \in [0, 1)$. For a given infinite history $h = \big(x^0, a^0, x^1, a^1, \ldots\big)$, the agent receives a sum of discounted payoffs

$$V(h) = \sum_{t=0}^{\infty} \delta^t v\big(x^t, a^t\big). \tag{2}$$

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.

Note 6 (Different Flavors of MDPs). *MDPs are amongst the most studied dynamical systems since Bellman's seminal work on dynamic programming [9]. Multiple books are devoted to their analysis [10, 11]. As a result, MDPs come in a variety of flavors. These different flavors are described below with an emphasis on the one used in this research:*

Discrete Time *In this research, the agent chooses an action at discrete-time steps. There also exist continuous-time Markov decision processes. This shift to uncountable spaces requires the use of more advanced measure theoretic tools to define probabilities.*

Finite State Space and Action Set *In this research, the state space and the action set of the agent are finite. This restriction guarantees that small enough problems can be simulated and solved on a computer. Some MDP results carry over from finite sets to countable sets. Some other problems use uncountable sets, such as the continuous real line. As mentioned previously, analyzing these problems require more advanced measure theoretic tools.*

Single Action Set for All States *In this research, at each time step and in each state, the agent is allowed to use any of the actions in its action set. In some problems, the action set is indexed by the state. In state x, the agent can choose an action in the set $\mathcal{A}_x$. The analysis with a single action set is not more restrictive but requires less notation.*

Unconstrained Optimization *In this research, the optimization performed by the agent is unconstrained. The addition of constraints requires the use of additional tools, such as Lagrange multipliers, to analyze the problems.*

Infinite Horizon *In this research, the cost is aggregated over an infinite time horizon. Other classes of MDPs predetermine a final time T at which the process stops. With a finite horizon, optimal strategies are computed by using backwards induction. In the infinite horizon setup, backwards induction is not applicable. However, a fixed-point property, described in the next section, replaces backwards induction. The game-theoretic literature strongly favors the use of infinite horizon.*

Absence of Termination State *In this research, the process goes on forever. A variation considers processes with a special state. If the process reaches this state, the payoffs are tallied and everything stops. This is, once again, a notational issue and the setup used in this research supersedes this variation.*

Discounted Payoff *In this research, the infinite stream of payoffs is aggregated through the use of a discounted sum. The objective in some MDPs is the average payoff. For finite-horizon problems this does not change anything. However, in the infinite horizon case, a non discounted sum might not converge and more technicalities have to be dealt with. Discounted payoffs are predominant in the game-theoretic literature.*

From now on, these characteristics are implied. Therefore, they will not be made explicit for each MDP encountered.

3.1.2 Strategies

In a static decision-making problem, as described in Section 2.1, the agent seeks to maximize its one-time payoff. This payoff is the utility associated with its action. Therefore, the agent faces an optimization problem of the form

$$\arg\max_{a\in\mathcal{A}} u(a).$$

In an MDP, the equivalent of this static utility function is the function $V\colon \mathcal{H}^\infty \to \mathbb{R}$ defined by (2). To determine the payoff of the agent, the entire infinite history $h \in \mathcal{H}^\infty$ is required. Therefore, the agent faces an optimization problem of the form

$$\arg\max_{h\in\mathcal{H}} V(h).$$

However, the agent cannot influence the history at will. The dynamic (1) imposes some constraints on the possible histories. Instead of choosing directly a history, the agent chooses a strategy, which is a plan of action for all the possible outcomes of the process. A strategy σ determines at time t an action a^t depending on h^t, the information available to the agent at time t. As was the case in the previous chapter, this action can also be mixed instead of pure. Therefore, a strategy is an element $\sigma\colon \mathcal{H} \to \Delta(\mathcal{A})$. An agent using strategy σ with initial state x receives an expected sum of discounted payoffs

$$U_\sigma(x) = \mathbb{E}_\sigma\big[V(h) \,\big|\, x^0 = x\big] = \mathbb{E}_\sigma\left[\sum_{t=0}^{\infty} \delta^t v\big(x^t, a^t\big) \,\middle|\, x^0 = x\right]. \tag{3}$$

Therefore, for a given initial state x, the agent faces an optimization problem of the form

$$\arg\max_{\sigma\in\Sigma} U_\sigma(x).$$

3.1.3 Agent Knowledge

Section 2.3.1 illustrates that equilibria are not intrinsic to a given static game. Depending on the story used, different solution concepts arise. For dynamic problems, some side information is similarly required. It is crucial to know the information available to the agent at every time step.

In an MDP, the information available to an agent is two fold. First, the agent knows some information a priori and keeps this knowledge all along. It knows the dynamic of the world f and the causality relations in place. The main causality relation is the impact of its action on the state evolution. Second, it accumulates some information along the way. At each time step, the agent observes the action played and the resulting state. At time t, it has accumulated the history h^t. The information available to the agent is represented in Figure 7.

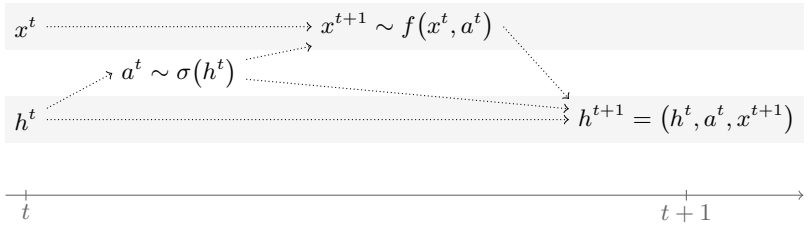


Figure 7. *Agent information in an MDP. The dotted arrows materialize causality. A value at the start of an arrow impacts the value at the end of this arrow. The agent knows all of these causality relations, the transition function f, and at time step t it has observed h^t. The purpose of the gray highlights is solely to improve readability. They do not emphasize specific values.*

The type of diagram introduced in Figure 7 is central in this research. Indeed, the solution concept introduced relies on tweaking the information available to the agents, and these diagrams help visualizing the process.

3.1.4 Bellman's Principle of Optimality

Recall that, for a given initial state x, the agent faces an optimization problem of the form

$$\arg\max_{\sigma \in \Sigma} U_\sigma(x).$$

It is actually possible to look for a single strategy that is simultaneously optimal for every initial state. As such, 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 nor that the intersection is not empty. Furthermore, recall that a strategy is a function from $\mathcal{H}$ to $\Delta(\mathcal{A})$. The domain of a strategy $\mathcal{H}$ is infinite; therefore,

the set of strategies Σ is infinite. As a result, looking for a solution with an exhaustive-search method is in vain.

Bellman was the first to observe that the Markovian structure of the problem gives structure to optimal strategies. He described this structure in his principle of optimality [9]:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

This simple principle has far-reaching consequences. The most important one guarantees that for any unconstrained discounted finite MDP, there exists a stationary deterministic optimal strategy [10, Theorem 6.2.10]. 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 (3) for each of the $|\mathcal{A}|^{|\mathcal{X}|}$ stationary deterministic strategies and finding the maximum is prohibitively expensive. Once again, the Markovian structure of the problem helps. There are more efficient ways to explore the solution space. As the existence of a stationary policy indicates, the only information that really matters in an MDP is the current state. The fact that there is no need to consider the entire history is captured in Figure 8, which simplifies the agent's knowledge diagram.

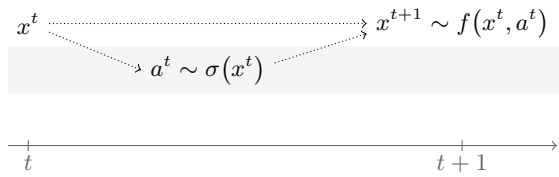


Figure 8. *Minimal agent information in an MDP. Bellman's principle of optimality guarantees it is enough to know the current state to act optimally. Tracking the entire history of state-action pairs cannot yield a higher expected sum of discounted payoffs.*

Bellman also gave a characterization of stationary deterministic optimal strategies. This characterization relies on two concepts. First, for an MDP, there exists a function $U^* \colon \mathcal{X} \to \mathbb{R}$ called the value function of the problem. When using an optimal strategy from the initial state x, the agent receives

a payoff $U^*(x)$. Second, we define the Bellman operator

$$\begin{aligned}\mathscr{B}\colon (\mathcal{X} \to \mathbb{R}) &\to (\mathcal{X} \to \mathbb{R}) \\ U &\mapsto \left\{ x \mapsto \max_{a \in \mathcal{A}} \left\{ u(x,a) + \delta \mathbb{E}_f \left[U\left(x^+\right) \,\middle|\, x, a \right] \right\} \right\},\end{aligned}$$

which takes a function that looks like a value function and returns another such function. Intuitively speaking, given an estimate of the value function U, a better estimate is $\mathscr{B}U$. Thanks to the discount factor δ being smaller than 1, the Bellman operator $\mathscr{B}$ is a contraction mapping. It therefore has a unique fixed point. This unique fixed point is the value function U^*. This result is known as the Bellman equation

$$U^* = \mathscr{B}U^*.$$

Given the value function U^*, a stationary deterministic strategy σ^* satisfies, for all x in $\mathcal{X}$,

$$\sigma^*(x) \in \arg\max_{a \in \mathcal{A}} \left\{ u(x,a) + \delta \mathbb{E}_f \left[U^*\left(x^+\right) \,\middle|\, x, a \right] \right\}.$$

This characterization is known as the one-shot deviation principle in the repeated games literature. As this name suggests, it is sufficient to verify that the one-shot action taken at each state is optimal to guarantee global optimality of the strategy.

3.1.5 Dynamic Programming

The astute reader noticed that the Bellman operator was not used in the characterization of optimal strategies. However, it is central in the actual computation of such strategies through dynamic programming. Indeed, 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 polynomially with the sizes of the state space and action set. However, the size of MDPs solvable in practice is limited. The two main dynamic-programming algorithms, value iteration and policy iteration, are presented below.

Value Iteration

The value iteration algorithm uses the fact that the Bellman operator $\mathscr{B}$ is a contraction mapping. On top of guaranteeing the existence of a fixed point, the contraction mapping property also guarantees that the fixed point is found by repeated application of the Bellman operator. For any initial

value $U^0 \colon \mathcal{X} \to \mathbb{R}$,

$$\lim_{t \to \infty} \mathscr{B}^t U^0 = U^*.$$

An actual algorithm yielding an ε optimal strategy is exposed in Algorithm 1. The stopping condition guarantees that the returned strategy is ε optimal. See [10, Theorem 6.3.1] for a detailed proof.

Algorithm 1 Value Iteration

```
procedure VALUE ITERATION(U^0, ε)
    t ← 0
    repeat
        for all x ∈ 𝒳 do
            U^{t+1}(x) ← ℬU^t(x)
        end for
        t ← t + 1
    until ‖U^t − U^{t−1}‖_∞ ≤ ε(1−δ)/(2δ)
    for all x ∈ 𝒳 do
        σ(x) ← arg max_{a∈𝒜} {u(x, a) + δ𝔼_f[U^t(x⁺) | x, a]}
    end for
    return σ
end procedure
```

As the name suggests, the algorithm computes successive approximation of the value function. The actual strategy is only computed at then end, once the approximation of the value function is satisfactory.

Policy Iteration

The policy iteration algorithm takes a different approach by computing successive strategies, also called policies. For a given strategy σ, the algorithm computes the expected payoff from each state, encoded in the function $U_\sigma \colon \mathcal{X} \to \mathbb{R}$. The next strategy is computed by taking U_σ as the approximation of the value function.

To compute U_σ, the following operator is used:

$$\begin{aligned} \mathscr{B}_\sigma \colon (\mathcal{X} \to \mathbb{R}) &\to (\mathcal{X} \to \mathbb{R}) \\ U &\mapsto \big\{ x \mapsto \mathbb{E}_{f,\sigma}\big[u(x,a) + \delta U\big(x^+\big) \,\big|\, x\big] \big\}. \end{aligned}$$

This operator is related to the Bellman operator. For the same reasons, it is a contraction mapping, and U_σ is computed by solving the equation

$$U_\sigma = \mathscr{B}_\sigma U_\sigma. \tag{4}$$

Solving Bellman's equation is difficult because of the maximization in the

Bellman operator. The lack of maximization makes solving (4) equivalent to a matrix inversion. The resulting algorithm is presented in Algorithm 2.

Algorithm 2 Policy Iteration

procedure POLICY ITERATION(σ^0)
  $t \leftarrow 0$
  repeat
    $U_{\sigma^t} \leftarrow$ the solution of the equation $U_{\sigma^t} = \mathscr{B}_{\sigma^t} U_{\sigma^t}$
    for all $x \in \mathcal{X}$ **do**
      $\sigma^{t+1}(x) \leftarrow \arg\max_{a \in \mathcal{A}} \{u(x,a) + \delta \mathbb{E}_f [U_{\sigma^t}(x^+) \,|\, x, a]\}$
    end for
    $t \leftarrow t + 1$
  until $\sigma^t = \sigma^{t-1}$
  return σ^t
end procedure

3.1.6 Online Learning

When the dynamic (1) of the system is not known but can be easily simulated, reinforcement-learning algorithms can be used [12, 13]. A reinforcement-learning algorithm learns the value function 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 using new strategies in order to get a better estimate of the value function. Exploitation refers to using a strategy maximizing the current estimate of the value function. Dynamic programming is an offline approach, whereas reinforcement learning is an online approach.

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

$$Q(x,a) = u(x,a) + \delta \mathbb{E}_f \big[U^*\big(x^+\big) \,\big|\, x, a \big].$$

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

3.2 Partially Observable Markov Decision Processes

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 drawn according to

$$y \sim g(x),$$

where y is a signal in finite state space $\mathcal{Y}$ and $g\colon \mathcal{X} \to \Delta(\mathcal{Y})$ is an observation function. In this setup, the information available to the agent is called private history and is denoted by p. At time t, the agent has observed $p^t = \big(y^0, a^0, y^1, a^1, \ldots, y^{t-1}, a^{t-1}, y^t\big)$. The information available to the agent is pictured in Figure 9. This small change in the information available to the agent has big consequences: POMDPs are intractable.

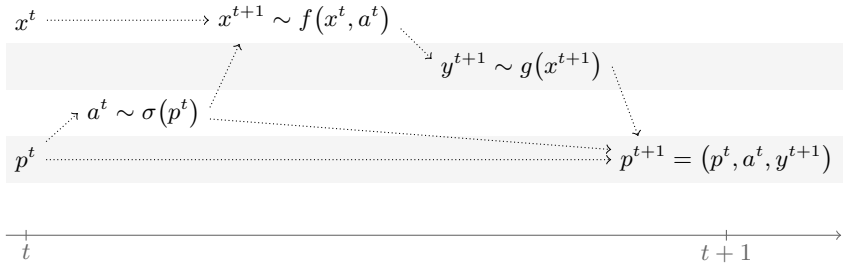


Figure 9. *Agent information in a POMDP. The agent knows all of the causality relations, the transition function f, the observation function g and at time step t it has observed private history p^t. In particular, the state is never observed and therefore is not part of the private history.*

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 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. This difficulty is partly visible in the minimal agent knowledge diagram of Figure 10.

3.3 Repeated Games

MDPs are the simplest dynamic decision-making problems. Similarly, repeated games are the simplest dynamic games. In a repeated game, agents play a one-shot game at discrete time steps and accumulate their payoffs with a discount factor. This section introduces repeated games with an emphasis on their similarities with MDPs.

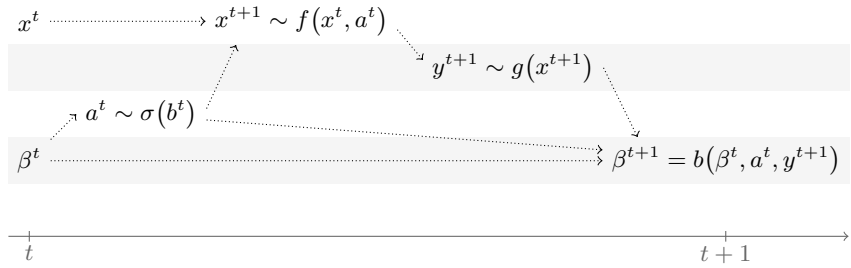


Figure 10. *Minimal agent information in a POMDP. Bellman's principle of optimality guarantees it is enough to know the distribution β over the states to act optimally. However, the belief space $\mathcal{B} = \Delta(\mathcal{X})$ is uncountable, hence, the problem is intractable. Bayesian inference is denoted by b.*

Consider a set of agents $\mathcal{I}$ and a one-shot game described by utility functions $u = (u_i)_{i\in\mathcal{I}}$ where $u_i \colon \mathcal{A} \to \mathbb{R}$. At each time step t, agent i chooses an action $a_i \in \mathcal{A}_i$. Over time, the agents accumulate some information. In the simplest class of repeated games, called perfect-monitoring repeated games, each agent observes the joint action played at each time step. The sequence of joint actions is called the public history, or simply history when there is no risk of confusion. The history up to time t is $h^t = (a^0, a^1, \dots, a^{t-1})$. Denote by $\mathcal{H}^t$ the set of histories up to time t and by $\mathcal{H} = \cup_{t\in\mathbb{N}}\mathcal{H}^t$ the set of all possible histories. The information observable over an infinite run is called an infinite history. The set of infinite histories is denoted by $\mathcal{H}^\infty$.

The joint action a yields a payoff $u_i(a)$ for agent i. Agent i is interested in maximizing its expected sum of discounted payoffs for a given discount factor $\delta_i \in [0, 1)$. For a given infinite history $h = (a^0, a^1, \dots)$, the agent receives a sum of discounted payoffs

$$U_i(h) = \sum_{t=0}^{\infty} \delta_i^t u_i(a^t).$$

The same way a one-shot game is described by a tuple of utility functions u, a perfect-monitoring repeated game is described by a pair (u, δ), where $\delta = (\delta_i)_{i\in\mathcal{I}}$.

Agent i's choice of action a_i at time t only depends on the observed sequence of joint actions h^t up to time t. Therefore a strategy for agent i is an element σ_i such that $\sigma_i \colon \mathcal{H} \to \Delta(\mathcal{A}_i)$. The set of strategies for agent i are denoted by Σ_i and the joint strategy set by $\Sigma = \prod_{i\in\mathcal{I}} \Sigma_i$. Given a joint

strategy $\sigma = (\sigma_i)_{i \in \mathcal{I}}$, agent i receives an expected discounted payoff

$$U_i(\sigma) = \mathbb{E}_\sigma \left[\sum_{t=0}^{\infty} \delta_i^t u_i\big(a^t\big) \right].$$

For the sake of rigor, when introducing MDPs, two symbols were used for the payoffs associated with a history, $V(h)$, and with a strategy, $U(\sigma)$. When unambiguous, U_i represents either payoff.

Repeated games include a time component which is not present in one-shot games. However, a repeated game can be viewed as a one shot game with action set Σ and utilities $U = (U_i)_{i \in \mathcal{I}}$. Therefore the notions of best response and Nash equilibria in one-shot games directly translate to repeated games. For an agent i and fixed strategies of its opponents σ_{-i}, agent i faces an MDP with state h. Its best response strategy is therefore characterized by Bellman's equation. A Nash equilibrium for a repeated game is therefore a tuple of strategies each satisfying a Bellman equation induced by the other ones. Note that the state space $\mathcal{H}$ is in this case is countable and not finite.

The knowledge of a pair of agents in a repeated game with perfect-monitoring is presented in Figure 11. Notice the strong resemblance with Figure 7 when looking from agent 1's perspective or from agent 2's perspective.

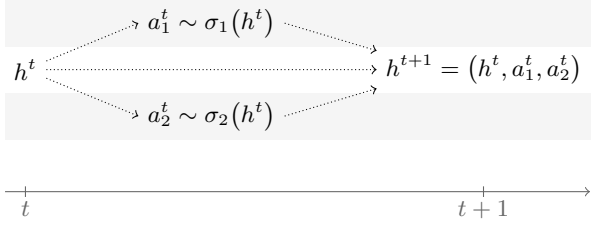


Figure 11. *Agent knowledge in a two-player perfect-monitoring repeated game. The public history is the sequence of joint actions. It is shared as it is observed by both agents.*

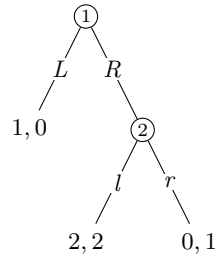
This section emphasized the similarities between repeated games and MDPs. However, there is one major difference in the treatment of repeated games. This difference is the subject of the following subsection.

3.3.1 Sequential Rationality

In a two-player repeated game, a Nash equilibrium is a pair of strategies (σ_1, σ_2). Agent 1's strategy σ_1 is an optimal strategy for the MDP induced by σ_2. However, the conditions for a Nash equilibrium do not

address the case where agent 2 does not follow σ_2. The following example illustrates this difficulty introduced by the notion of time.

Example 3 (Non-credible Threat). *Consider the following diagram:*



. (5)

It describes a game in its so-called extensive form. Play starts with the state being the root of the tree. When the state is a non-terminal node, the number in the circle determines which agent is to take an action. The branches of this node correspond to the actions available to this agent. The action taken determines the next state. When the state is a leaf, the game is over. The numbers in this leaf correspond to the payoffs for both agents. Therefore, the game described by (5) *has the following interpretation:*

- *Agent* 1 *chooses between actions* L *and* R.
  - *If* L *is chosen, the game is over; agent* 1 *receives a payoff of* 1 *and agent* 2 *a payoff of* 0.
  - *If* R *is chosen, it is now agent* 2*'s turn to choose between actions* l *and* r. *Irrespective of which action is chosen, the game ends after it.*
    - *If* l *is chosen, both agents receive a payoff of* 2.
    - *If* r *is chosen, agent* 1 *receives a payoff of* 0 *and agent* 2 *a payoff of* 1.

Note 7 (Extensive-form Games). *Extensive-form games can always be redefined in a one-shot form. For example, the extensive-form game* (5) *admits the following one-shot definition:*

	l	r
L	1, 0	1, 0
R	2, 2	0, 1

.

Note that agent 2*'s action does not impact the payoffs when agent* 1 *plays* L. *Therefore, the agents' incentives are preserved and rational agents exhibit identical behaviors in the one-shot game or the extensive-form game. Since the incentives are preserved, Nash equilibria are also preserved.*

Example 4 (Non-credible Threat [continued]). *Extensive-form games are not repeated games. However, they introduce a notion of time that is sufficient to illustrate the problem at hand.*

Note 7 shows us that this game has two pure Nash equilibria: (R, l) *and* (L, r)*. Let's focus on* (L, r)*. Under the joint action* (L, r)*, agent* 1 *plays and the game ends immediately. Therefore, agent* 2*'s action does not affect the payoffs and agent* 2 *has no incentive to unilaterally deviate. If agent* 1 *switches its action to* R *its payoff goes from* 1 *to* 0 *so it has no incentive to unilaterally deviate either. However, this switch to* R *brings an interesting situation to the table. If the deviation occurs, it becomes agent* 2*'s turn to play. Agent* 2 *has committed to playing* r*. However, if the game ever reaches that stage, a rational agent would always play* l*. The problem is that agent* 2 *makes a non-credible threat. Agent* 2 *is threatening agent* 1 *with a low payoff. However, enforcing that threat requires agent* 2 *to act irrationally by taking a smaller payoff,* 1 *instead of* 2*. This equilibrium is said to lack sequential rationality.*

In the normal-form representation the problem is not as apparent since both agents play at the same time. However, since anything is a best response to L*, agent* 2 *could move to* l *which would prompt agent* 1 *to move to* R *leading to the other Nash equilibrium and both agents' payoff increases. Therefore, even in the normal-form representation, the weakness of* (L, r) *is observable.*

When time is involved, strategies have the potential to exclude some states. In this example, agent 2 *never gets to play. A Nash equilibrium does not impose anything on these states. A sequentially-rational equilibrium, however, imposes no profitable unilateral deviation even on these unreachable states. This condition is equivalent to forbidding non-credible threats.*

Backwards induction is used to verify if a Nash equilibrium of an extensive-form game is sequentially rational. In this example, start at agent 2*'s turn. The only rational action is* l*. Propagate this information backwards and analyze agent* 1*'s turn. At this point, agent* 1 *has to play* R*. This proves the only sequentially-rational equilibrium of this game is* (R, l)*.*

In a repeated game, agents do not alternate playing turns. However, the action of an agent eliminates some possible histories, creating some unreachable states. The Nash equilibrium condition in repeated games only verifies Bellman's equation on the reachable states. Sequential rationality in repeated games verifies Bellman's equation on all the possible states. In an MDP, unreachable states are simply ignored as the model guarantees that these states will never be seen.

Let's now give the formal definition of sequential rationality for perfect-monitoring repeated games. In this context, a sequentially-rational equilibrium is called a subgame-perfect equilibrium.

Definition 9 (Subgame-perfect Equilibrium). *Let* $u \colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ *describe a*

one-shot game and $\delta = (\delta_i)_{i\in\mathcal{I}}$ be discount factors. Let $\sigma_i \in \Sigma_i$ be a strategy for agent i.

The joint strategy $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ is a subgame-perfect equilibrium if for every agent $i \in \mathcal{I}$ and every history $h \in \mathcal{H}$, the strategy σ_i is optimal with respect to the MDP induced by σ_{-i} with initial state h.

The game-theoretic literature often mentions the one-shot deviation principle as a key result in verifying subgame perfection. It only restates that Bellman's equation has to be verified at every state h, including unreachable ones.

3.3.2 Folk Theorem

Sequentially-rational equilibria are the logical extension of Nash equilibria for repeated games. As mentioned previously, in the static-game setting, economists are interested in characterizing the set of achievable payoffs at equilibrium. A similar characterization is studied for repeated games. The results are more difficult to obtain but a folk theorem has guided this line of research. The name folk theorem comes from the community believing it to be true before a proof existed. To state the folk theorem, the concepts of feasibility and individual rationality of payoffs are described below.

Definition 10 (Feasible Payoff). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. A joint payoff $p \in \mathbb{R}^{|\mathcal{I}|}$ is feasible for u if there exist convex coefficients $(\theta_a)_{a\in\mathcal{A}}$ indexed by the joint actions, such that $p = \sum_{a\in\mathcal{A}} \theta_a u(a)$. The tuple $(\theta_a)_{a\in\mathcal{A}}$ forms convex coefficients if $\theta_a \in [0,1]$ for all $a \in \mathcal{A}$ and $\sum_{a\in\mathcal{A}} \theta_a = 1$. Simply put, a payoff is feasible if it is a convex combination of pure payoffs.*

Definition 11 (Minmax Value). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game and $i \in \mathcal{I}$ be an agent. Suppose the opponents of agent i have fixed their, potentially mixed, actions. Agent i can guarantee for itself a payoff, called its minmax value, defined as*

$$\text{minmax}_i = \min_{\alpha_{-i}\in\prod_{j\in-i}\Delta(\mathcal{A}_j)} \max_{a_i\in\mathcal{A}_i} u_i(a_i, a_{-i}).$$

Definition 12 (Individually-rational Payoff). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. A joint payoff $p \in \mathbb{R}^{|\mathcal{I}|}$ is individually rational for u if for all agents $i \in \mathcal{I}$, p_i is greater or equal than the minmax value of agent i. A joint payoff $p \in \mathbb{R}^{|\mathcal{I}|}$ is strictly individually rational for u if for all agents $i \in \mathcal{I}$, p_i is strictly greater than the minmax value of agent i.*

The folk theorem states that every feasible individually-rational payoff of the one-shot game is achievable as the expected sum of discounted payoffs

of a sequentially-rational equilibrium for a discount factor close enough to one.

The following example illustrates the concepts of feasible and individually-rational payoffs.

Example 5 (Feasible Individually-rational Payoffs in the Battle of the Sexes)**.** *Recall the battle of the sexes game described by the following normal form:*

		♀ F	♀ O
♂ F		2, 2	0, 1
♂ O		0, 0	1, 3

The minmax value for the man is

$$\begin{aligned}
\text{minmax}_{♂} &= \min_{\alpha_{♀} \in \Delta(\{\text{F},\text{O}\})} \max_{a_{♂} \in \{\text{F},\text{O}\}} u_{♂}\big(a_{♂}, a_{♀}\big) \\
&= \min_{p_{♀} \in [0,1]} \max_{a_{♂} \in \{\text{F},\text{O}\}} \big\{u_{♂}\big(a_{♂}, \text{F}\big)p_{♀} + u_{♂}\big(a_{♂}, \text{O}\big)\big(1 - p_{♀}\big)\big\} \\
&= \min_{p_{♀} \in [0,1]} \max\big\{2p_{♀} + 0\big(1 - p_{♀}\big), 0p_{♀} + 1\big(1 - p_{♀}\big)\big\} \\
&= \min_{p_{♀} \in [0,1]} \max\big\{2p_{♀}, 1 - p_{♀}\big\} \\
&= \frac{2}{3}.
\end{aligned}$$

The minmax value for the woman is

$$\begin{aligned}
\text{minmax}_{♀} &= \min_{\alpha_{♂} \in \Delta(\{\text{F},\text{O}\})} \max_{a_{♀} \in \{\text{F},\text{O}\}} u_{♀}\big(a_{♀}, a_{♂}\big) \\
&= \min_{p_{♂} \in [0,1]} \max_{a_{♀} \in \{\text{F},\text{O}\}} \big\{u_{♀}\big(a_{♀}, \text{F}\big)p_{♂} + u_{♀}\big(a_{♀}, \text{O}\big)\big(1 - p_{♂}\big)\big\} \\
&= \min_{p_{♂} \in [0,1]} \max\big\{2p_{♂} + 0\big(1 - p_{♂}\big), 1p_{♂} + 3\big(1 - p_{♂}\big)\big\} \\
&= \min_{p_{♂} \in [0,1]} \max\big\{2p_{♂}, 3 - 2p_{♂}\big\} \\
&= \frac{3}{2}.
\end{aligned}$$

The feasible payoffs are contained in the convex hull of the pure payoffs. The individually-rational payoffs are those above both minmax values. The feasible and individually-rational payoffs for the battle of the sexes are illustrated in Figure 12.

The folk theorem is not a single result. It takes different forms, each

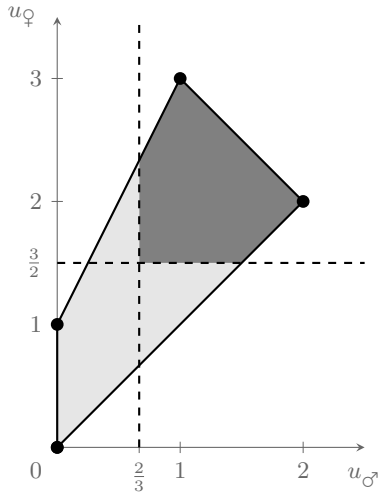


Figure 12. *Feasible enforceable payoffs in the battle of the sexes. The dotted circles corresponds to the payoffs of the four pairs of pure actions. The union of the light gray and the dark gray areas represents the feasible payoffs. The dashed lines represent the minmax values. The dark gray area corresponds to the feasible and individually-rational payoffs.*

targeting a specific scenario. The simplest result concerns perfect-monitoring repeated games.

Theorem 5 (Perfect-monitoring Folk Theorem). *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $p \in \mathbb{R}^{|\mathcal{I}|}$ be a feasible strictly-individually-rational payoff for u.*

There exist strategies $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ and discount factors $\delta = (\delta_i)_{i\in\mathcal{I}}$ close enough to 1 such that σ form a subgame-perfect equilibrium for the perfect-monitoring repeated game (u, δ) yielding an expected sum of discounted payoff of p.

For a proof of this result, see in [14, Proposition 3.8.1]

Apart from perfect-monitoring repeated games, there are two other categories of repeated games. Public imperfect-monitoring and private monitoring repeated games are explored next. Their associated folk theorems are brushed upon.

3.3.3 Public Imperfect Monitoring

Consider a set of agents $\mathcal{I}$ and a one-shot game described by utility functions $u = (u_i)_{i\in\mathcal{I}}$ where $u_i\colon \mathcal{A}_i \to \mathbb{R}$. At each time step t, agent i chooses

an action $a_i \in \mathcal{A}_i$. Over time, the agents accumulate some information. The joint action induces a signal s, from finite signal space $\mathcal{S}$, distributed according to

$$s^+ \sim n(a),$$

where $n \colon \mathcal{A} \to \Delta(\mathcal{S})$. The sequence of signals is called the public history, or simply history when there is no risk of confusion. The history up to time t is $h^t = \left(s^0, s^1, \ldots, s^{t-1}\right)$. Denote by $\mathcal{H}^t$ the set of histories up to time t and by $\mathcal{H} = \cup_{t \in \mathbb{N}} \mathcal{H}^t$ the set of all possible histories. The information observable over an infinite run is called an infinite history. The set of infinite histories is denoted by $\mathcal{H}^\infty$. Each agent also observes the actions it has played. This sequence of actions is called the its private history. Agent i' private history up to time t is $p_i^t = \left({a_i}^0, {a_i}^1, \ldots, {a_i}^{t-1}\right)$.

The knowledge of the agents in a two-player public imperfect-monitoring repeated game is pictured in Figure 13.

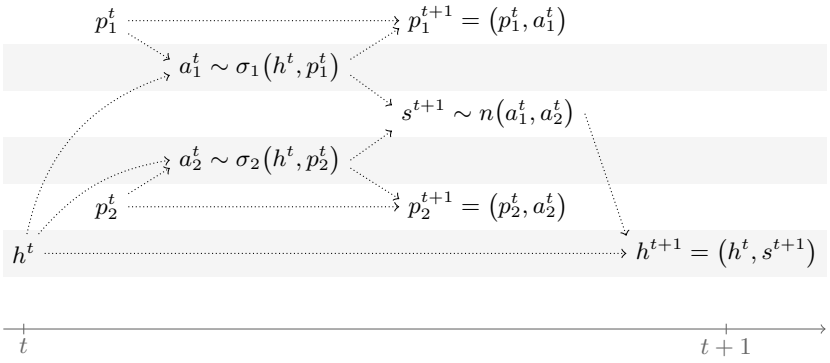


Figure 13. *Agent knowledge in a two-player public-monitoring repeated game. The public history is shared as it is observed by both agents.*

Public-monitoring folk theorem results exist. Under some technical conditions for the signal, all the feasible strictly-individually-rational payoff are achievable by sequentially-rational equilibria in public strategies. A strategy is public if it only uses the public history to compute actions. The knowledge of the agents in a two-player public imperfect-monitoring repeated game with public strategies is pictured in Figure 14.

3.3.4 Private Monitoring

Consider a set of agents $\mathcal{I}$ and a one-shot game described by utility functions $u = (u_i)_{i \in \mathcal{I}}$ where $u_i \colon \mathcal{A}_i \to \mathbb{R}$. At each time step t, agent i chooses an action $a_i \in \mathcal{A}_i$. The joint action induces for agent i a signal s_i from finite signal space $\mathcal{S}_i$. The signals $s = (s_i)_{i \in \mathcal{I}}$ are potentially correlated and

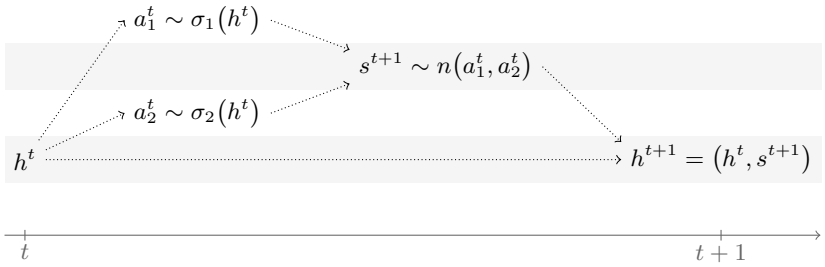


Figure 14. *Agent knowledge in a two-player public-monitoring repeated game with public strategies. The private histories are not kept and only the public history remains.*

distributed according to

$$s^+ \sim n(a),$$

where $n \colon \mathcal{A} \to \Delta(\mathcal{S})$. The sequence of signals observed by agent i is called its private history, or simply history when there is no risk of confusion. Agent i's history up to time t is $h_i^t = (s_i{}^0, s_i{}^1, \dots, s_i{}^{t-1})$. Denote by $\mathcal{H}_i^t$ the set of agent i's histories up to time t and by $\mathcal{H}_i = \cup_{t \in \mathbb{N}} \mathcal{H}_i^t$ the set of all possible histories. The information observable over an infinite run is called an infinite history. The set of infinite histories for agent i is denoted by $\mathcal{H}_i^\infty$.

The knowledge of the agents in a two-player private-monitoring repeated game is pictured in Figure 15.

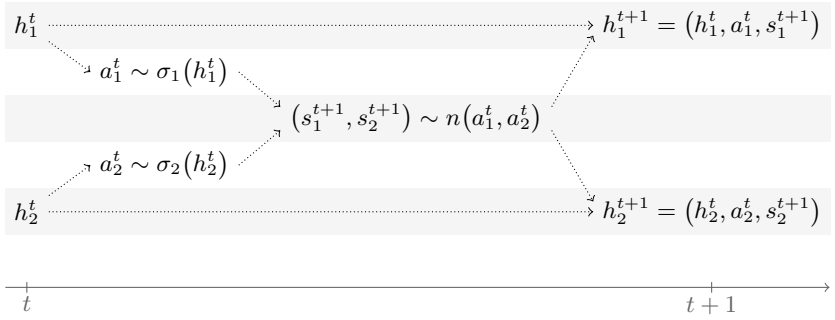


Figure 15. *Agent knowledge in a two-player private-monitoring repeated game.*

Perfect-monitoring repeated games are closely related to MDPs. Therefore, the sequential-rationality condition in a perfect-monitoring repeated game requires the Bellman equation to be satisfied after every possible history. Similarly, private-monitoring repeated games are related to POMDPs. Agent i faces a POMDP with state $(h_i)_{i \in \mathcal{I}}$ and observation (a_i, s_i). Accord-

ingly, in the private-monitoring setting, the sequential-rationality condition requires that each agent's strategy be optimal for a POMDP. This requires that the Bellman equation involving beliefs over the state be satisfied after every possible tuple of histories $(h_i)_{i\in\mathcal{I}}$. The fact that the agent do not share a public signal makes this setting incredibly more complicated.

The folk theorem for private-monitoring repeated games has recently been derived [15]. Before this 200-page long achievement, some partial results relied on the existence of subsets of strategies with a nice recursive structure. For example, belief-free equilibria [16, 17] and weakly belief-free equilibria [18] are two solution concepts that were used to derive some partial folk theorems. In a belief-free equilibrium, agents must only use actions that are optimal no matter what their belief about the last action played by their opponents is. In a weakly belief-free equilibrium, agents only need to have correct beliefs about the last action played by their opponents.

3.4 Stochastic Games

Stochastic games [19] are the most general 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.

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

$$x^+ \sim f(x, a),$$

where x and x^+ are states in a finite state space $\mathcal{X}$ and $a = \big(a_1, \ldots, a_{|\mathcal{I}|}\big)$ is a joint action in the finite joint action set $\mathcal{A} = \prod_{i\in\mathcal{I}} \mathcal{A}_i$. In state x, the joint action a yields for agent i a payoff $u_i(x, a)$.

A variety of stochastic games are defined by varying the monitoring structure. As an example, the agent-knowledge diagrams for perfect and private monitoring are presented in Figures 16 and 17.

The main results in repeated games characterize the payoffs achievable at equilibrium. There are virtually no result to actually compute equilibrium strategies. This lack of result is not surprising. In the simplest setting of perfect-monitoring repeated game, at equilibrium an agent need to solve an MDP. This MDP depends on the strategies of its opponents. However, the strategies of the opponents is usually not available to the agent, especially in a learning setting where the strategies evolve. Stochastic games form

a strict superset of repeated games. The added complexity explains that there are very few results available.

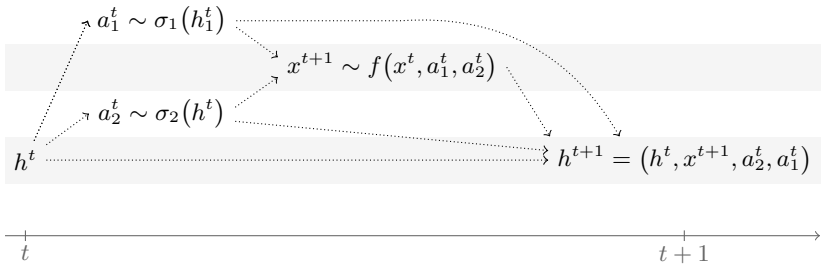


Figure 16. *Agent knowledge in a two-player perfect-monitoring stochastic game*

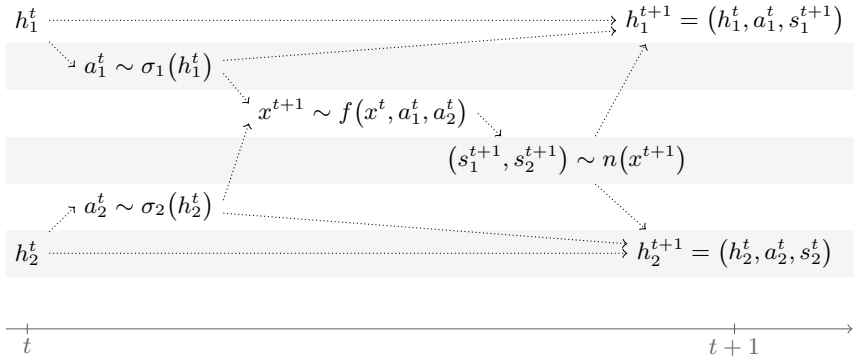


Figure 17. *Agent knowledge in a two-player private-monitoring stochastic game*

4 Decentralized Control and Games

4.1 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 in the sense of [20, 21] 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. This research uses bounded rationality to make each agent face an MDP instead of a POMDP.

The rest of this chapter introduces work paving the way towards using game-theoretic learning to design decentralized controllers. These results are grouped under three main themes: learning in games, equilibria in repeated games, and bounded rationality.

4.2 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 this 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 [22].

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 this 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 [23]. In fact, no adaptive learning rule converges to Nash equilibria for all games [24]. This result is in part due to the fact that computing a Nash equilibrium is PPAD complete [25]. PPAD, which stands for Polynomial Parity Arguments on Directed graphs, is a complexity class contained between P and NP. PPAD complete problems are believed to be hard to solve but the exact relationship to P and NP is not known.

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.

4.2.1 Correlated Equilibria

Hart and Mas-Colell proved that a family of adaptive learning rules converge to the set of correlated equilibria [20, 21, 26]. 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 $\frac{1}{t}\sum_{\tau=1}^{t} a^{\tau}$ is different from the joint action a^t and that convergence to a set is less constraining than convergence to a point.

4.2.2 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 [27], insures that better-reply learning algorithms converge to a Nash equilibrium in weakly acyclic games [28]. Weakly acyclic games are an extension of potential games, a class of games used to model congestion problems and to systematically design decentralized controllers [29].

4.2.3 Stochastically Stable States

Young introduced the notion of stochastically stable states to characterize the long-run behavior of a Markov chain subject to a diminishing random noise [30]. 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 [31], to characterize the yield of self-assembly mechanisms [32], and to study language evolution [33].

4.3 Equilibria in Repeated Games

This section presents results pertaining to repeated games. The first result extends the parallel existing between MDPs and repeated games by adapting reinforcement learning to a multiagent setting. The next three results are equilibrium concepts lowering the requirements on the beliefs imposed by sequential rationality.

4.3.1 Multiagent Reinforcement Learning

Hu and Wellman attempted to apply results from reinforcement learning to the multiagent setting with the Nash–Q-learning algorithm [34]. 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.

When agents have identical utility functions, the problem is identical to a single-agent problem. Therefore, classical reinforcement-learning results carry over. When agents have opposite utility functions, they are facing a zero-sum game. In a zero-sum game, one can define *the* solution by using the minimax theorem. The lack of ambiguity in defining rational solution concepts explains the convergence.

4.3.2 Subjective and Self-confirming Equilibria

Subjective equilibria, introduced by Kalai and Lehrer, lower the requirements on the beliefs in repeated games [35]. 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 [36]. 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 this research.

4.3.3 Belief-free and Weakly Belief-free Equilibria

Belief-free equilibria and weakly belief-free equilibria are solution concepts for private-monitoring repeated games, presented in Section 3.3.4. They lower the rationality requirements by not requiring the agents to carry full beliefs about the state of their opponents.

4.3.4 Analogy-based Expectation Equilibria

For games of perfect information, Jehiel introduced the concept of analogy-based expectation equilibria (ABEEs) to keep the belief space size constant [37]. 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 $\mathcal{A}_j$, $\alpha^{i,\kappa_i}_j[a_j] = \mathbb{P}[\sigma_j(h) = a_j \mid h \in \kappa_i]$.

The ABEE concept is a substantial step in the direction of this 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. This 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.

4.4 Bounded Rationality and Consistency

In classical game theory, agents are assumed to be fully rational. Bounded rationality studies scenarios where agents have limited computation power or make mistakes [38]. 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 this 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 this research lifts that restriction. Other differences with this research are highlighted.

4.4.1 Linear Modeling

Chang, Ho and Kaelbling used modeling to simplify multiagent learning [39]. 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.

4.4.2 Mean-field Games

Lasry and Lions studied a setting where a very large number of agents faces identical copies of an MDP [40]. 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. This research aims for a different equilibrium concept in general games with any number of players. Furthermore, MFEs focuses on stationary models, whereas this 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 [41]. They named their equilibrium concept oblivious equilibrium.

4.4.3 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 [42]. 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 this 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.

4.4.4 Egocentric Modeling

Seah and Shamma analyzed a specific game where two agents share a one-dimensional signal [43]. 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 this research, an inaccurate simplified model is used to lower the complexity of computations.

5 Empirical-evidence Equilibria

Chapter 3 showed that equilibria in repeated games and stochastic games are complicated entities. At best, at equilibrium in a perfect-monitoring repeated game, each agent solves an MDP. At worst, in a private-monitoring stochastic game, each solves a POMDP. This chapter introduces a new equilibrium concept for stochastic games in which each agent only solves an MDP. As was the case in Chapters 2 and 3, the concept is first introduced through a single-agent point of view. Then, the full-fledged multiagent case is exposed.

The following presentation of this research was first developed in [44].

5.1 Single-agent Setup

Consider a discrete-time dynamical system governed by

$$x^+ \sim f\big(x, a, s^+\big), \tag{6}$$

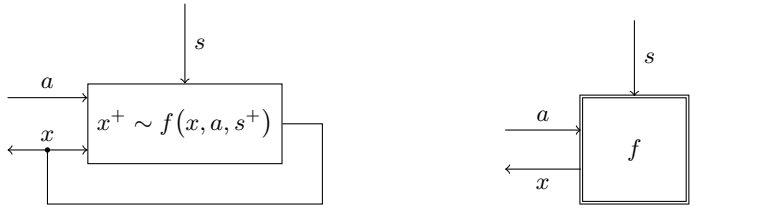
where x is a state, a is an action, and s is a signal. Variables x, a, and s take values in finite sets $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{S}$, respectively. The agent picks the action a. Nature determines the signal s^+ according to

$$w^+ \sim n(w, x, a), \tag{7a}$$

$$s^+ \sim \nu\big(w^+\big), \tag{7b}$$

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 (6) and (7). Think of this system as a perturbed MDP. The block diagram associated with (6) will be used later in this chapter. It is represented in Figure 18.

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 $p^t = \big(o^0, o^1, \dots, o^t\big) \in \mathcal{P}^t$ and the true history is $h^t = \big(r^0, r^1, \dots, r^t\big)$. Denote by $\mathcal{P} = \cup_{t\in\mathbb{N}}\mathcal{P}^t$ 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. The agent knowledge of this perturbed MDP is represented in Figure 19, using the diagram format introduced in Chapter 3.



(a) *Full representation* **(b)** *Condensed representation*

Figure 18. *Block diagram for the MDP with perturbation signal. The contour of the block in the condensed representation is doubled. This serves as a visual reminder that the output variable is fed back into the block.*

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 $\sum_{t=0}^{\infty} \delta^t u\big(x^t, a^t, s^{t+1}\big)$, where $\delta \in [0, 1)$ is a discount factor. The agent wants to find a strategy maximizing its expected sum of discounted payoffs

$$U_{\mathbf{N},\sigma}(x) = \mathbb{E}_{\mathbf{N},\sigma}\left[\sum_{t=0}^{\infty} \delta^t u\big(x^t, a^t, s^{t+1}\big) \,\middle|\, x^0 = x\right].$$

When the agent knows (6) and (7), it is facing a POMDP. To see why, compare Figure 19 with Figure 9. The POMDP has state (w, x) and observed value (x, s). A natural solution concept for this type of problems is an optimal policy for the POMDP. As described in Section 3.2, the agent computes an optimal policy using beliefs, which are probability distributions over states. Beliefs are obtained from the private histories p^t, the signaling structure (7), and the application of Bayes' rule. Belief computation is intractable because the size of the belief space grows with time.

The present research is interested in the case where the agent knows (6) but does not know (7). As previously mentioned, this is an MDP with a perturbation. The agent fully understands the effect of the perturbation s on (6) but does not know how this perturbation is generated. This block diagram for this setting is shown in Figure 20. The information known to the agent is highlighted in Figure 21. 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.

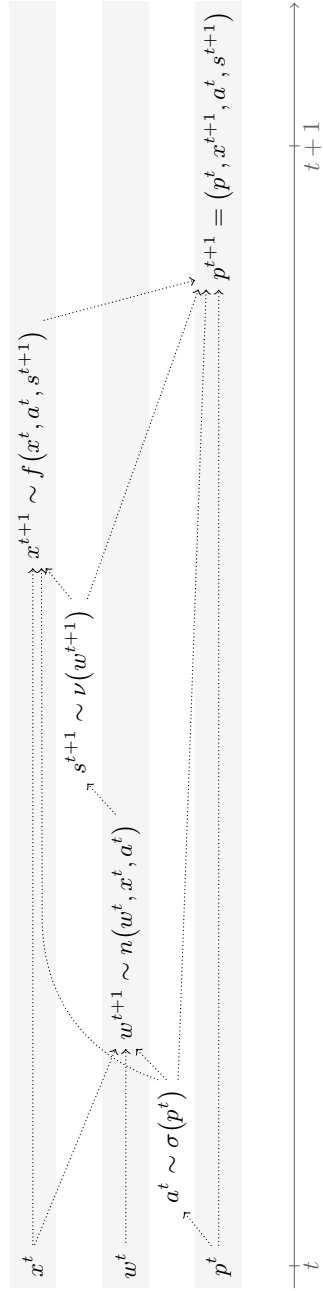


Figure 19. *Agent knowledge in the MDP with perturbation signal.*

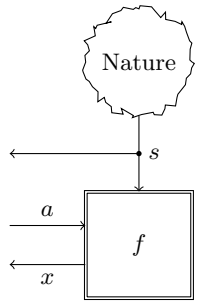


Figure 20. *Block diagram for the single-agent empirical-evidence setup* **N**. *The agent is aware of facing an MDP with an unknown perturbation signal. The noisy contour of Nature emphasizes that the agent does not know anything about it. Furthermore, Nature has access to all the variables to compute s, but these dependencies are not represented.*

5.2 Depth-k Consistency

The notion of consistency used in this research is best introduced through an example.

Example 6 (Binary Signal)**.** *Suppose an agent receives a binary signal. Furthermore, suppose the agent has no information about the underlying generating process. When the agent observes a realization of this binary signal, it can compute certain parameters.*

One of the simplest set of parameters is the probabilities of 0*s and* 1*s. For example, if the agent observes the following sequence:*

$$S_{\mathrm{a}} = 011011011011011011\ldots,$$

it would compute the following parameters:

$$\mathbb{P}_{\mathrm{a}}[0] = \frac{1}{3} \qquad \textit{and} \qquad \mathbb{P}_{\mathrm{a}}[1] = \frac{2}{3}.$$

Similarly, if the agent observes this other sequence:

$$S_{\mathrm{b}} = 001111001111001111\ldots,$$

it would compute

$$\mathbb{P}_{\mathrm{b}}[0] = \frac{1}{3} \qquad \textit{and} \qquad \mathbb{P}_{\mathrm{b}}[1] = \frac{2}{3}.$$

If the probabilities of 0*s and* 1*s are the only parameters used by the agent, it would not differentiate* S_{a} *and* S_{b}*. Sequences* S_{a} *and* S_{b} *are said to be*

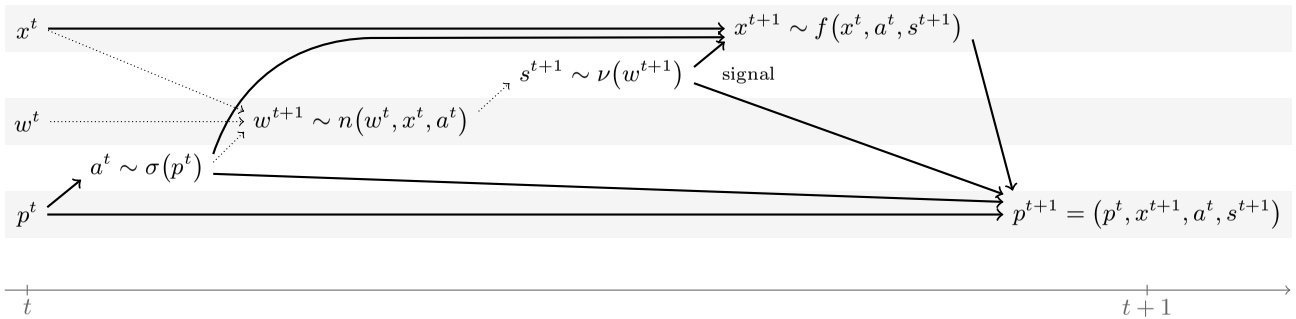


Figure 21. *Agent knowledge in the single-agent empirical-evidence setup* $\mathbf{N}$*. The agent only acknowledges the causality materialized by the bold arrows. In particular, its signal is underspecified. From the agent's perspective, the signal arrows have an unknown source.*

*depth-*0 *consistent.*

The agent can use more parameters to characterize the signal. For example, the probabilities of 00*s,* 01*s,* 10*s, and* 11*s. Using the same sequences, the agent would compute*

$$\mathbb{P}_{\mathrm{a}}[00] = 0, \qquad \mathbb{P}_{\mathrm{a}}[10] = \frac{1}{3},$$
$$\mathbb{P}_{\mathrm{a}}[01] = \frac{1}{3}, \qquad \mathbb{P}_{\mathrm{a}}[11] = \frac{1}{3},$$

for S_{a}*, and*

$$\mathbb{P}_{\mathrm{b}}[00] = \frac{1}{6}, \qquad \mathbb{P}_{\mathrm{b}}[10] = \frac{1}{6},$$
$$\mathbb{P}_{\mathrm{b}}[01] = \frac{1}{6}, \qquad \mathbb{P}_{\mathrm{b}}[11] = \frac{1}{2},$$

for S_{b}*. Using these parameters, which correspond to a deeper analysis, the agent is able to differentiate* S_{a} *and* S_{b}*. Sequences* S_{a} *and* S_{b} *are not depth-*1 *consistent.*

Consider C, an $\mathcal{S}$-valued 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 $\mathcal{S}^{k+1}$

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

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

$$z^+ = m^k\big(z, s^+\big), \tag{8a}$$
$$s^+ \sim \mu(z), \tag{8b}$$

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

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

and whose block diagram is depicted in Figure 22. Under some technical assumptions, described in Section 5.3, the observed signal and the Markov chain described by (8) are ergodic processes. Furthermore, the Markov chain is depth-k consistent with the true signal when the following equality holds:

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

Example 7 (Binary Signal [continued])**.** *The agent has observed a realization*

(a) *Full representation* **(b)** *Condensed representation*

Figure 22. *Block diagram for the depth-k mockup.*

of the signal. It has computed the parameters of interest. The agent now creates a mockup with similar parameters. It fixes, a priori, a deterministic Markov chain, as described by (8a)*. Then, it identifies the values of* μ *making the mockup consistent with the observed sequence.*

*For depth-*0 *consistency, it is sufficient to use a Markov chain with a single state, as depicted in Figure 23. Remember that sequences* S_a *and* S_b *are depth-*0 *consistent. Therefore, a single distribution* $\mu(z_\emptyset)$ *makes the mockup depth-*0 *consistent with* S_b *and* S_b*. This distribution is the following:*

$$\mu(z_\emptyset)[0] = \mathbb{P}_\mathrm{a}[0] = \mathbb{P}_\mathrm{b}[0] = \frac{1}{3},$$
$$\mu(z_\emptyset)[1] = \mathbb{P}_\mathrm{a}[1] = \mathbb{P}_\mathrm{b}[1] = \frac{2}{3}.$$

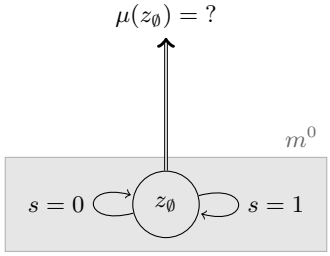


Figure 23. *Mockup enabling depth-*0 *consistency with binary signals. The mockup is composed of a fixed deterministic Markov chain* m^0 *with a single state and a distribution* $\mu(z_\emptyset)$*. By adjusting the distribution* $\mu(z_\emptyset)$*, this mockup can be made depth-*0 *consistent with any given ergodic binary signal.*

*For depth-*1 *consistency, a Markov chain with two state, portrayed in Figure 24, is required. The distributions making the mockup depth-*1 *consistent*

with the sequence S_{a} *are*

$$\begin{aligned}\mu(z_0)[0] &= \mathbb{P}_{\mathrm{a}}[0 \,|\, z_0] = 0,\\ \mu(z_0)[1] &= \mathbb{P}_{\mathrm{a}}[1 \,|\, z_0] = 1,\end{aligned} \qquad \textit{and} \qquad \begin{aligned}\mu(z_1)[0] &= \mathbb{P}_{\mathrm{a}}[0 \,|\, z_1] = \frac{1}{2},\\ \mu(z_1)[1] &= \mathbb{P}_{\mathrm{a}}[1 \,|\, z_1] = \frac{1}{2}.\end{aligned}$$

*Similarly,the distributions making the mockup depth-*1 *consistent with the sequence* S_{b} *are*

$$\begin{aligned}\mu(z_0)[0] &= \mathbb{P}_{\mathrm{b}}[0 \,|\, z_0] = \frac{1}{2},\\ \mu(z_0)[1] &= \mathbb{P}_{\mathrm{b}}[1 \,|\, z_0] = \frac{1}{2},\end{aligned} \qquad \textit{and} \qquad \begin{aligned}\mu(z_1)[0] &= \mathbb{P}_{\mathrm{b}}[0 \,|\, z_1] = \frac{1}{4},\\ \mu(z_1)[1] &= \mathbb{P}_{\mathrm{b}}[1 \,|\, z_1] = \frac{3}{4}.\end{aligned}$$

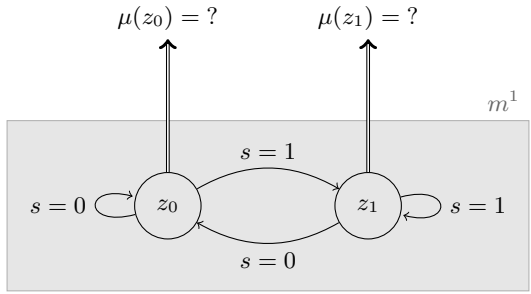


Figure 24. *Mockup enabling depth-*1 *consistency with binary signals. The mockup is composed of a fixed deterministic Markov chain* m^1 *with two states and two distributions* $\mu(z_0)$ *and* $\mu(z_1)$*. By adjusting the distributions* $\mu(z_0)$ *and* $\mu(z_1)$*, this mockup can be made depth-*1 *consistent with any given ergodic binary signal.*

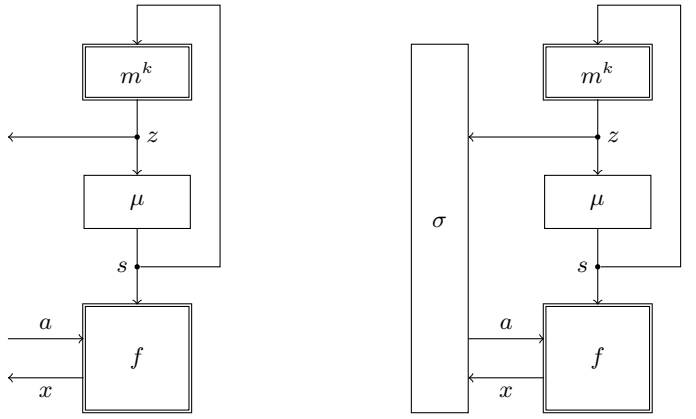
Note that the mockup is split in two parts. First, since the agent does not know anything about the signal, it assumes that it is generated by a parametric model. Second, it computes observation distributions matching the empirical evidence provided by the signal. The mockup built this way is consistent with the signal. Nothing in the empirical evidence contradicts the assumption made by the agent.

Denote by $\mathbf{M}^k$ the dynamical system described by (6) and (8). This system is obtained from the original system $\mathbf{N}$ by substituting Nature with a depth-k consistent mockup. It is depicted in Figure 25. The system $\mathbf{M}^k$ induces an MDP with state (x, z), action a, strategy $\sigma\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{A}$, and

the objective function

$$U_{\mathbf{M}^k,\sigma}(x^0, z^0) = \mathbb{E}_{\mathbf{M}^k,\sigma}\left[\sum_{t=0}^{\infty} \delta^t u(x^t, a^t, s^{t+1})\right].$$

A strategy σ for the MDP can be implemented in the real system by building z with (8a). This trick, allowing to use a strategy designed for the simpler system $\mathbf{M}^k$ in the more complicated system $\mathbf{N}$, is illustrated in Figure 26. The agent knowledge in system $\mathbf{M}^k$ is depicted in Figure 27. The simplification to the agent knowledge system $\mathbf{N}$, induced by the use of the length-k–memory function, is presented in Figure 28.

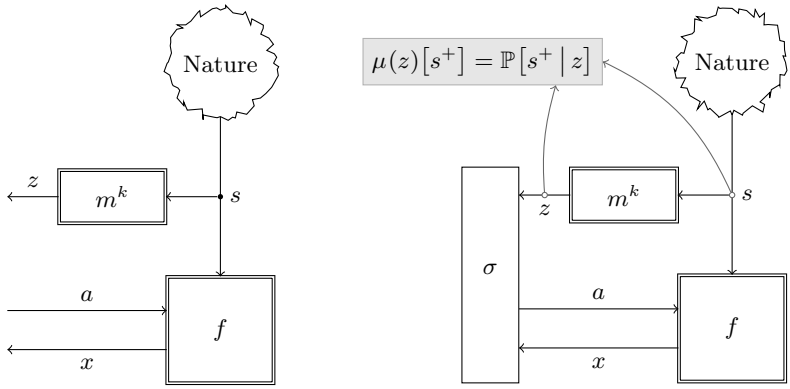


(a) *Open loop* **(b)** *Closed loop with a strategy*

Figure 25. *Block diagram for the single-agent empirical-evidence setup with mockup* $\mathbf{M}^k$. *For a given distribution* μ, *the system forms a finite MDP and an optimal strategy can be computed.*

By using depth-k consistency, we went from an arbitrarily complicated system $\mathbf{N}$ to a finite MDP $\mathbf{M}^k$. The next idea is to compute optimal strategies in the simpler system and measure their impact in the real system.

Consider the following iterative process, illustrated in Figure 29. 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.



(a) *Open loop* **(b)** *Closed loop with a strategy*

Figure 26. *Block diagram for the single-agent empirical-evidence setup* $\mathbf{N}$ *with length-k–memory function. For a given strategy, the closed-loop system generates a probability distribution* $\mathbb{P}[s^+ \mid z]$*. Using this distribution for* μ *along with* m^k *yields a depth-k–consistent mockup.*

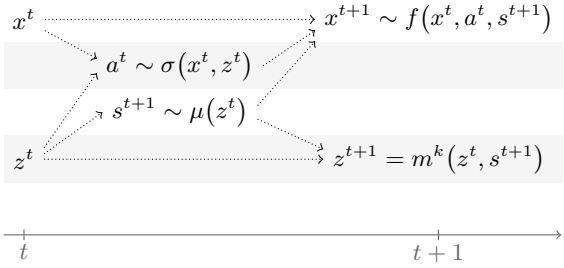


Figure 27. *Agent knowledge in the single-agent empirical-evidence setup with mockup* $\mathbf{M}^k$*. The agent knows all of the causality relations and all the parameters of the MDP, namely* f *and* μ*.*

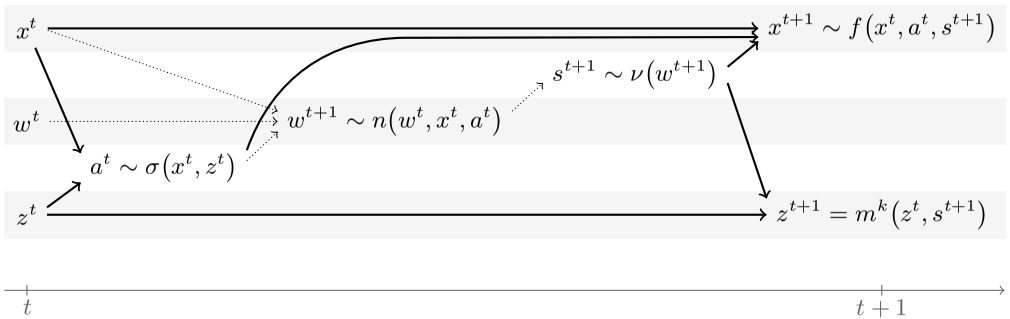
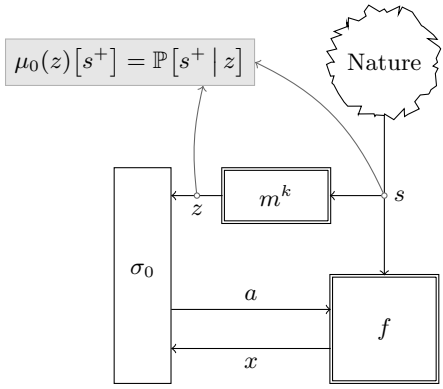
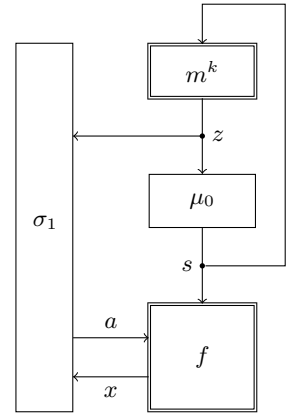


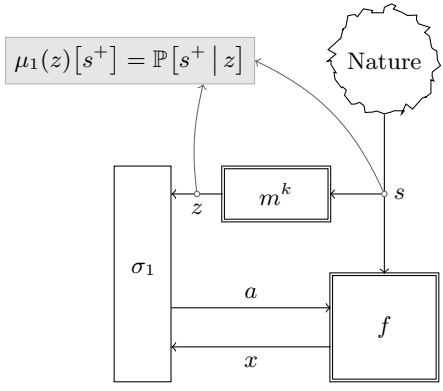
Figure 28. *Agent knowledge in the single-agent empirical-evidence setup* $\mathbf{N}$*, simplified by the use of the length-k–memory function. The full private information is not retained anymore. The pair* (x, z) *is the information used by the agent to compute the next action.*



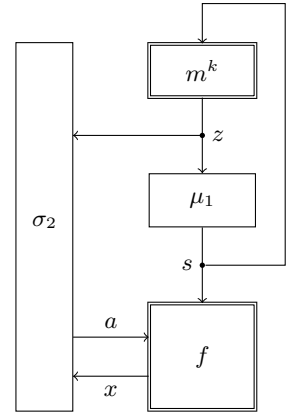
(a) *Use σ_0 in $\mathbf{N}$ and measure μ_0 that makes the mockup depth-k consistent.*



(b) *Use μ_0 in $\mathbf{M}^k$ to setup an MDP. Compute σ_1 optimal for this MDP.*



(c) *Use σ_1 in $\mathbf{N}$ and measure μ_1 that makes the mockup depth-k consistent.*



(d) *Use μ_1 in $\mathbf{M}^k$ to setup an MDP. Compute σ_2 optimal for this MDP.*

Figure 29. *Iterative process alternating between the real and the mockup system. Using a strategy in the real system yields a measurement. Using this measurement in the mockup system yields an optimal strategy.*

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.

5.3 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:

$$z^+ \sim m\big(z, x^+, a, s^+\big), \tag{9a}$$

$$s^+ \sim \mu(z). \tag{9b}$$

The model m represents the assumption the agent makes about the system. The predictor μ is the set of parameters the agent adjusts to obtain a signal resembling its observations. The pair (m, μ) is called a mockup. Denote by $\mathbf{M}$ the dynamical system described by (6) and (9). The block diagram for this system is depicted in Figure 30. The block diagram for the associated system $\mathbf{N}$ with the model m is depicted in Figure 31. The agent knowledge in systems $\mathbf{M}$ and $\mathbf{N}$ are presented in Figure 32 and Figure 33 respectively.

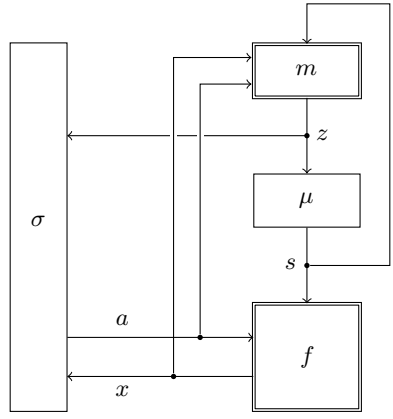


Figure 30. *Block diagram for the single-agent empirical-evidence setup with mockup* $\mathbf{M}$.

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

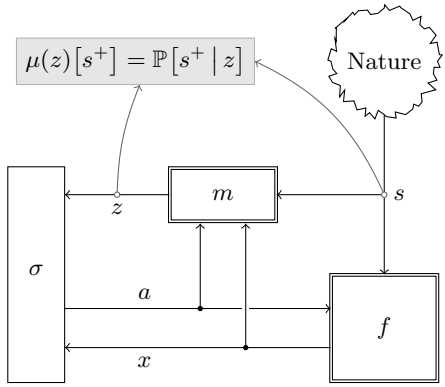


Figure 31. *Block diagram for the single-agent empirical-evidence setup* **N** *with model* m.

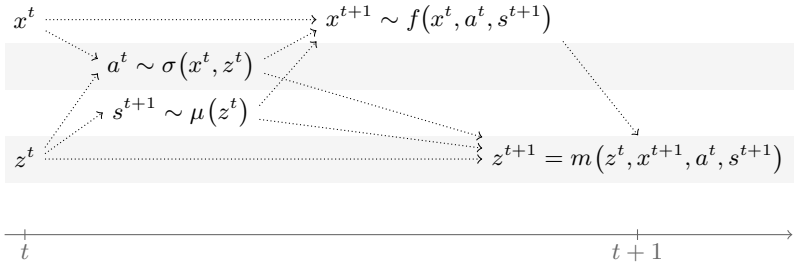


Figure 32. *Agent knowledge in the single-agent empirical-evidence setup with mockup* **M**.

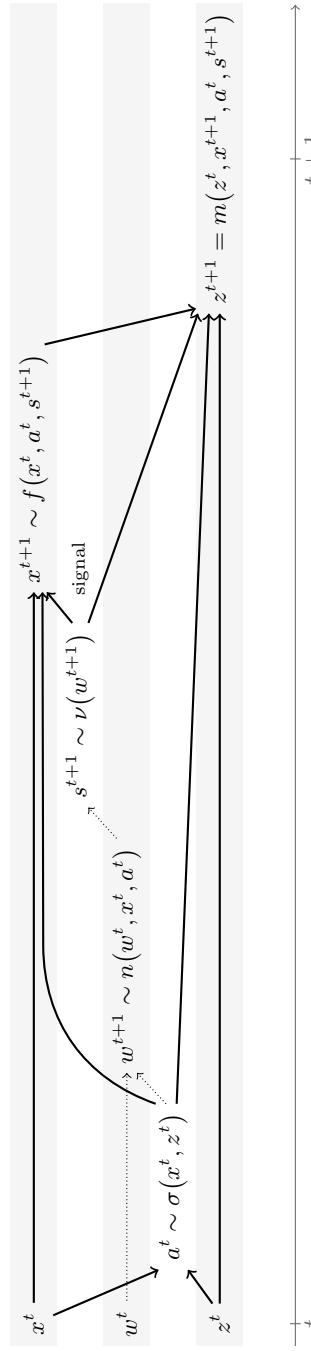


Figure 33. *Agent knowledge in the single-agent empirical-evidence setup* **N**, *simplified by the use of the model* m.

Definition 13. *Let σ be a strategy and (m, μ) be a mockup. Predictor μ is (σ, m) consistent with $\mathbf{N}$ if*

$$\forall z \in \mathcal{Z} \text{ and } s^+ \in \mathcal{S},\ \mu(z)\big[s^+\big] = \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s^+ \,\big|\, z^t = z\big].$$

The notion of optimality used is the following.

Definition 14. *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 15. *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 13 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 irreducible and aperiodic. In this case, some authors say that the Markov chain T_σ is ergodic.*

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

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

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 13 is well defined for all z and s as follows:

$$\begin{aligned}
\mu(z)\big[s^+\big] &= \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s^+ \,\big|\, z^t = z\big] \\
&= \mathbb{P}_{\pi_\sigma}\big[s^+ \,\big|\, z\big] \\
&= \sum_{w^+\in\mathcal{W}} \mathbb{P}_{\pi_\sigma}\big[s^+ \,\big|\, z, w^+\big] \cdot \mathbb{P}_{\pi_\sigma}\big[w^+ \,\big|\, z\big] \\
&= \sum_{w^+\in\mathcal{W}} \mathbb{P}_{\pi_\sigma}\big[s^+ \,\big|\, w^+\big] \frac{\mathbb{P}_{\pi_\sigma}[w^+, z^+]}{\mathbb{P}_{\pi_\sigma}[z]} \\
&= \sum_{w^+\in\mathcal{W}} \nu\big(w^+\big)\big[s^+\big] \cdot \\
&\qquad \frac{\displaystyle\sum_{\substack{w\in\mathcal{W}\\ x\in\mathcal{X}\\ a\in\mathcal{A}}} \pi_\sigma[w,x,z]\cdot\sigma(x,z)[a]\cdot n(w,x,a)\big[w^+\big]}{\displaystyle\sum_{\substack{w\in\mathcal{W}\\ x\in\mathcal{X}}} \pi_\sigma[w,x,z]}
\end{aligned} \tag{10}$$

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.

5.4 Weak Consistency and Eventual Consistency

The notion of consistency exposed in Definition 13 is fairly strong. It requires that the quantity $\mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s^+ \,\big|\, z^t = z\big]$ converges for all $z \in \mathcal{Z}$ and $s^+ \in \mathcal{S}$. As a result, the associated Assumption 1 is constraining. This section, highlights two slightly less stringent notions of consistency with their associated assumptions. The first one, weak consistency, uses a weaker notion of convergence. The second one, eventual consistency, only requires convergence on a subset of states $z \in \mathcal{Z}$.

First, let us define the notion of weak consistency. It relies on the fact that convergence of the average of a sequence is weaker than convergence of the sequence.

Definition 16. *Let σ be a strategy and (m,μ) be a mockup. Predictor μ is weakly (σ,m) consistent with $\mathbf{N}$ if*

$$\forall z \in \mathcal{Z} \text{ and } s^+ \in \mathcal{S},\ \mu(z)\big[s^+\big] = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\mathbf{N},\sigma}\big[s^{t+1} = s^+ \,\big|\, z^t = z\big].$$

The associated assumption is the following.

Assumption 2. *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 irreducible.*

Assumption 2 insures that T_σ has a unique stationary distribution π_σ such that

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

Furthermore, Assumption 2 guarantees that π_σ has full support. Using the same reasoning as in (10) guarantees that μ in Definition 16 is well defined.

Second, let us define the notion of eventual consistency. It relies on the fact that the value of $\mu(z)$ for a state z with zero probability in the limit is irrelevant.

Definition 17. *Let σ be a strategy and (m, μ) be a mockup. Predictor μ is eventually (σ, m) consistent with $\mathbf{N}$ if for all states $z \in \mathcal{Z}$ such that $\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}[z^t = z] > 0$*

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

For $z \in \mathcal{Z}$ such that $\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}[z^t = z] = 0$, there is no requirement on $\mu(z)$. Its value is totally arbitrary.

If a state z is such that $\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}[z^t = z] = 0$, it will not be seen in the long run. Therefore, there is no need to impose some constraints on μ for this state. The following assumption is associated with this definition.

Assumption 3. *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 unichain and aperiodic.*

A Markov chain is unichain if it contains a single communication class. Assumption 3 insures that T_σ has a unique stationary distribution π_σ with full support on its communication class, and no support outside of it. Therefore, for $z \in \mathcal{Z}$ such that $\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}[z^t = z] > 0$, it is the case that $\mathbb{P}_{\pi_\sigma}[z] > 0$. Furthermore, for that same z,

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

Once again, using the same reasoning as in (10) guarantees that μ in Definition 17 is well defined. The division by $\mathbb{P}_{\pi_\sigma}[z]$ only occurs when this quantity is strictly positive. When it is zero, μ is defined arbitrarily.

Combining Definitions 16 and 17 yields a third notion, eventual weak consistency. The associated assumption requires the Markov chain to be unichain.

The following list summarizes the different notions of consistency and their associated assumptions on the Markov chain:

Consistent Irreducible and aperiodic.

Weakly consistent Irreducible.

Eventually consistent Unichain and aperiodic.

Eventually weakly consistent Unichain.

For conciseness reasons, the rest of this presentation only mentions the first notion of consistency defined in Definition 13. However, the results are applicable to all the notions of consistency defined in the present section. Extra care needed to accommodate a specific notion of consistency will be addressed on a case by case basis.

5.5 Predictors and Strategies

Given a strategy σ, there is a unique predictor μ which is (σ, m) consistent with $\mathbf{N}$. This predictor can be measured in the system $\mathbf{N}$, as depicted in Figure 31. Note that (10) guarantees that μ is a continuous function of σ and π_σ. The mapping associating a predictor to each strategy is a function. This function is denoted by $F^{\mathrm{M},m}$, where M stands for modeling.

Given a predictor μ, there might be multiple strategies σ that are (μ, m) optimal. Such strategies are the optimal strategies for the MDP induced by system $\mathbf{M}$, depicted in Figure 30. Therefore, the mapping associating a predictor to the corresponding optimal strategies is a correspondence. This correspondence is denoted $F^{\mathrm{O},m}$, where O stands for optimization. As in Chapter 2, we define a function approximating this correspondence. This will allow us later to once again use Brouwer's fixed-point theorem to gain intuition before using Kakutani's. 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

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

and σ by

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

The astute reader recognizes the Gibbs distribution, described in details in Note 5. Recall that, 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^*. This function approximating the optimization correspondence is denoted $F^{\mathrm{O},m,\varepsilon}$.

The composition of an optimization mapping and a modeling mapping gives a mapping from the space of strategies to itself. Two such mappings can be defined, the correspondence $F^m = F^{\mathrm{O},m} \circ F^{\mathrm{M},m}$ and the function $F^{m,\varepsilon} = F^{\mathrm{O},m,\varepsilon} \circ F^{\mathrm{M},m}$.

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

5.6 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 $\big(x_1, x_2, \dots, x_{|\mathcal{I}|}\big)$. Define a and s similarly. Agent i is controlling the system described by

$$x_i^+ \sim f_i\big(x_i, a_i, s_i^+\big). \tag{11}$$

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

$$x_{-i}^+ \sim f_{-i}\big(x_{-i}, a_{-i}, s_{-i}^+\big). \tag{12}$$

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

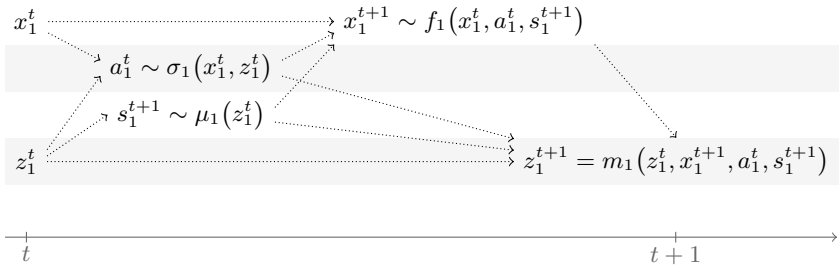
$$w^+ \sim n(w, x, a), \tag{13a}$$

$$s^+ \sim \nu\big(w^+\big). \tag{13b}$$

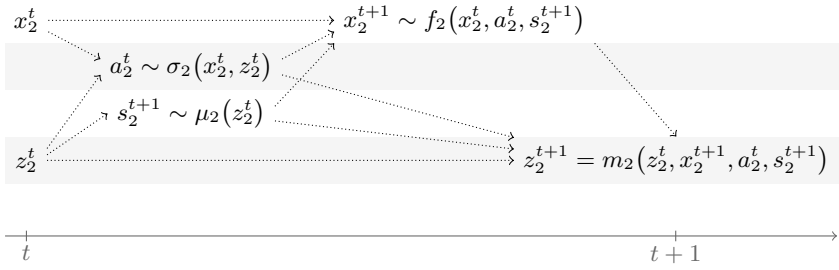
The rest of this section extends the notions of consistency and optimality to this setting. The block diagrams and agent-knowledge figures for a two-agent setup, available in Figures 34 to 37, help in following the discussion. In particular, they highlight the fact that there are very few differences in the treatment of the single-agent problem and the multiagent one. By design, the notions of consistency and optimality used remove the differences between the two settings.

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 (6) and an unknown part (7). Similarly, $\mathbf{N}_i$ has a known part (11) and an unknown part (12) and (13).

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



(a) *Agent 1's knowledge in* $\mathbf{M}_1$



(b) *Agent 2's knowledge in* $\mathbf{M}_2$

Figure 34. *Agent knowledge in the two-agent empirical-evidence setup with mockup* $\mathbf{M}_1$ *and* $\mathbf{M}_2$*. Once again, the diagrams for the two agents are identical and not different from the single agent setting.*

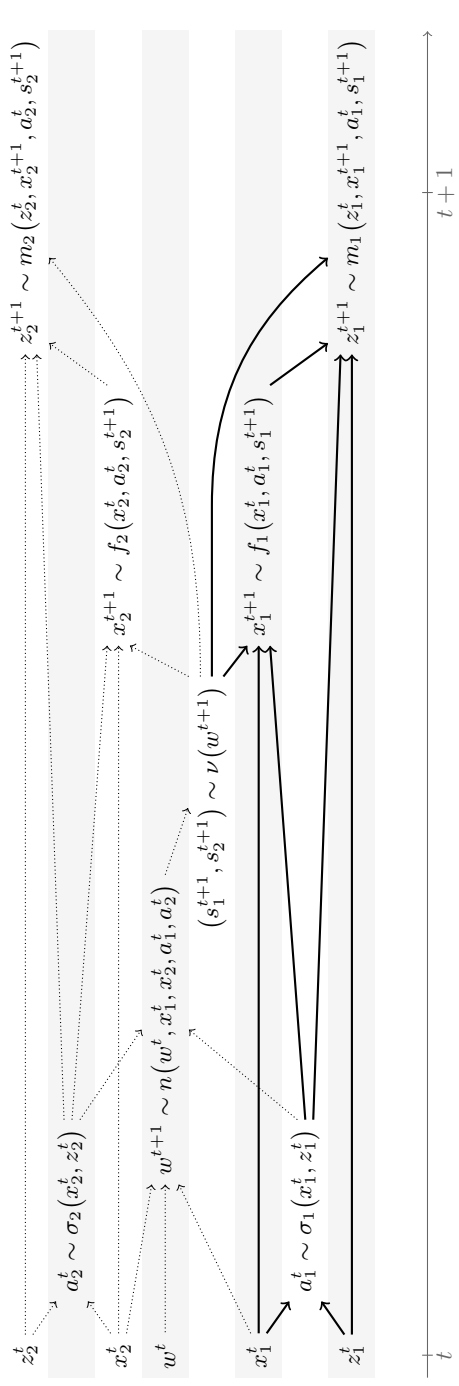


Figure 35. *Agent knowledge for the two-agent empirical-evidence setup* $\mathbf{N}$ *with models* m_1 *and* m_2*. Agent* 1*'s knowledge of the causality relations is highlighted with bold arrows. Notice that, from agent* 1*'s point of view, nothing has changed from the single-agent setup. Its signal is underspecified but it does not realize that there is another agent along with Nature.*

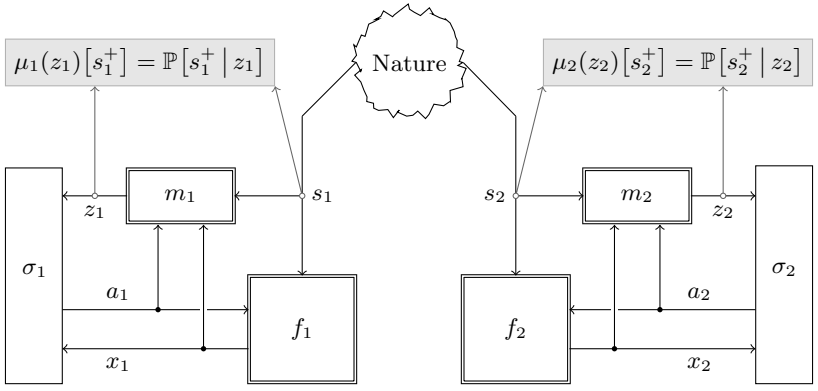
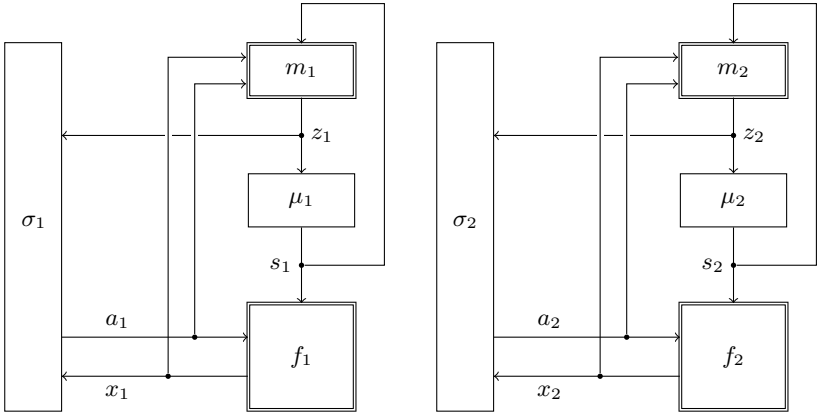


Figure 36. *Block diagram for the two-agent empirical-evidence setup* $\mathbf{N}$ *with models* m_1 *and* m_2*.*



(a) *Agent* 1*'s block diagram* **(b)** *Agent* 2*'s block diagram*

Figure 37. *Block diagrams in the two-agent empirical-evidence setup with mockup* $\mathbf{M}_1$ *and* $\mathbf{M}_2$*. Notice that the two block diagrams are identical. Furthermore, from the agents' perspective, there is no difference between the single-agent and the multiagent setting.*

described by a state z_i, a model m_i, and a predictor μ_i. This forms the system $\mathbf{M}_i$.

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 18. *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 $F^{\mathrm{O},m}$ the optimization correspondence from predictors to strategies, by $F^{\mathrm{O},m,\varepsilon}$ its approximating function, 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, a correspondence from the space of strategies to itself, by $F^m = F^{\mathrm{O},m} \circ F^{\mathrm{M},m}$. Similarly define $F^{m,\varepsilon}$, a function from the space of strategies to itself, by $F^{m,\varepsilon} = F^{\mathrm{O},m,\varepsilon} \circ F^{\mathrm{M},m}$.

It is sometimes easier to work with a function from the space of predictors to itself. In these cases, we use $G^{m,\varepsilon} = F^{\mathrm{M},m} \circ F^{\mathrm{O},m,\varepsilon}$.

Now that the setup has been established, it is time to prove some results. The first result tackles the existence of EEEs.

5.7 Existence of Empirical-evidence Equilibria

The proof of existence of empirical-evidence equilibria follows along the line of the proof of existence of Nash equilibria presented in Section 2.5. It starts by proving the existence of approximate equilibria through Brouwer's fixed-point theorem, before proving the existence of exact equilibria using Kakutani's fixed-point theorem. The proof of the existence of Nash equilibria used the best-response correspondence. The best response correspondence mapped the set of independent mixed actions to itself. In the empirical-evidence setting, these actions are replaced by strategies, and F^m plays a similar role to the best-response correspondence.

5.7.1 Existence of Approximate Equilibria

Theorem 6. *Let $m = (m_i)_{i\in\mathcal{I}}$ be models and $\varepsilon = (\varepsilon_i)_{i\in\mathcal{I}}$ be positive numbers. There exists an (ε, m) EEE.*

Proof. First, show that $F^{m,\varepsilon}$ has a fixed point. The set of strategies is representable by a product of simplices. Therefore, $F^{m,\varepsilon}$ is a mapping from a convex and compact set to itself. By Propositions 6 and 7, $F^{\mathrm{O},m,\varepsilon}$ 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.

The upcoming Proposition 5 therefore implies that an (ε, m) EEE exists. □

Proposition 5. *Let $m = (m_i)_{i\in\mathcal{I}}$ be models and $\varepsilon = (\varepsilon_i)_{i\in\mathcal{I}}$ be positive numbers. Let σ^* be a fixed point of $F^{m,\varepsilon}$. Define μ^* by $\mu^* = F^{\mathrm{M},m}(\sigma^*)$. The pair (μ^*, σ^*) is an (ε, m) EEE.*

Proof. Fix $i \in \mathcal{I}$. By definition predictor μ^* is (σ^*, m_i) consistent with $\mathbf{N}_i$. Note that

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

This implies that strategy σ^* is $(\varepsilon_i, \mu_i^*, m_i)$ optimal. Therefore, (μ^*, σ^*) is an (ε, m) EEE. □

Proposition 6. *Let $m = (m_i)_{i\in\mathcal{I}}$ be models and $\varepsilon = (\varepsilon_i)_{i\in\mathcal{I}}$ be positive numbers.*

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.

The upcoming 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 5.5 that (b) is continuous. □

Proposition 7. *Let $m = (m_i)_{i\in\mathcal{I}}$ be models.*

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.

The elements $(\sigma_i(x_i, z_i))_{x_i \in \mathcal{X}_i, z_i, \in \mathcal{Z}_i}$ are entries in the matrix T_σ. Therefore (c) is linear, hence continuous. [45, Theorem 4.1] shows that the stationary distribution of a finite irreducible Markov chain is a continuous function of the elements of its transition matrix, which proves that (d) is continuous. It was noted in Section 5.5 that (e) is continuous. □

The result in [45, Theorem 4.1] targets finite irreducible Markov chains. Therefore, the proof of Proposition 7 immediately holds when using weak consistency. The following observation allows to accommodate eventual consistency as well. Consider a finite unichain Markov chain with communication class $\mathcal{C}$. Label the states such that the states in $\mathcal{C}$ come before the other ones. Let T be the transition matrix for this Markov chain. It has the following block structure:

$$\begin{pmatrix} T' & 0 \\ A & B \end{pmatrix}$$

where T' is the transition matrix for an irreducible Markov chain on $\mathcal{C}$. Denote by π and π' the stationary distributions of T and T'. These stationary distributions coincide on $\mathcal{C}$, which in block notation is denoted by $\pi = (\pi' 0)$. First, the function $T \mapsto T'$ is a projection and therefore continuous. Then, the function $T' \mapsto \pi'$ is continuous according to [45, Theorem 4.1]. Finally, the function $\pi' \mapsto \pi = (\pi' 0)$ is trivially continuous. This guarantees that Proposition 7 holds when using eventual consistency or eventual weak consistency.

Lemma 1 (Continuity of the Value Function in the Parameters of an MDP)**.** *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 parameters of the problem. It corresponds to all the entries in f and u. Let $\mathscr{B}_\theta$ be the Bellman operator associated with the problem. By definition, the value function of the problem U^*_θ is the fixed point of $\mathscr{B}_\theta$, $U^*_\theta = \mathscr{B}_\theta U^*_\theta$.*

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

Proof. Let θ and θ' be two vectors of parameters corresponding to two MDPs. Let U^*_θ and $U^*_{\theta'}$ be the value functions associated with these MDPs. We will show that U^*_θ converges to $U^*_{\theta'}$ as θ converges to θ' by showing that $\|U^*_\theta - U^*_{\theta'}\|$ converges to 0.

The value function U^*_θ is a fixed point of $\mathscr{B}_\theta$. The Bellman operator $\mathscr{B}_\theta$ is a contraction mapping with Lipschitz constant δ. As a result,

$$\begin{aligned}\|U^*_\theta - U^*_{\theta'}\| &= \|\mathscr{B}_\theta U^*_\theta - U^*_{\theta'}\| \\ &\leq \|\mathscr{B}_\theta U^*_\theta - \mathscr{B}_\theta U^*_{\theta'}\| + \|\mathscr{B}_\theta U^*_{\theta'} - U^*_{\theta'}\| \\ &\leq \delta\|U^*_\theta - U^*_{\theta'}\| + \|\mathscr{B}_\theta U^*_{\theta'} - U^*_{\theta'}\| \\ &\leq \frac{1}{(1-\delta)}\|\mathscr{B}_\theta U^*_{\theta'} - U^*_{\theta'}\|.\end{aligned} \tag{14}$$

We will now prove that the function $\theta \mapsto \mathscr{B}_\theta U^*_{\theta'}$ is continuous because each of its finitely many components is continuous. By definition, $(\mathscr{B}_\theta U^*_{\theta'})(x) = \max_{a\in\mathcal{A}} v(x,a,\theta)$, where $v(x,a,\theta) = u(x,a) + \delta f(x,a)^\top U^*_{\theta'}$. For fixed x and a, $\theta \mapsto v(x,a,\theta)$ is linear and therefore continuous. For a fixed x, $\theta \mapsto \mathscr{B}_\theta U^*_{\theta'}(x)$ is the maximum of a finite number of continuous functions and as such is continuous. Therefore, the function $\theta \mapsto \mathscr{B}_\theta U^*_{\theta'}$ is continuous.

As a result, as θ converges to θ', $\mathscr{B}_\theta U^*_{\theta'}$ converges to $\mathscr{B}_{\theta'} U^*_{\theta'}$. Since $U^*_{\theta'}$ is a fixed point of $\mathscr{B}_{\theta'}$, $\mathscr{B}_\theta U^*_{\theta'}$ converges to $U^*_{\theta'}$. We have proven that the limit of $\mathscr{B}_\theta U^*_{\theta'} - U^*_{\theta'}$ as θ goes to θ' is zero. Finally, (14) implies that $\|U^*_\theta - U^*_{\theta'}\|$ goes to zero as θ goes to θ' which concludes the proof. □

5.7.2 Existence of Exact Equilibria

Theorem 7. *Let $m = (m_i)_{i\in\mathcal{I}}$ be models.*
There exists an exact m EEE.

Proof. This proof follows closely the proof of the existence of an exact Nash equilibrium.

An exact optimization counterpart to Proposition 5 is easily established. It guarantees that strategies $\sigma \in \Sigma$ forming a fixed point of F^m, corresponds to an m EEE. Therefore, proving the existence of such a fixed point is a sufficient condition to proving the theorem.

To apply Kakutani's fixed point theorem we need to prove the four following facts:

- The set Σ is non-empty, compact and a convex subset of an Euclidean space.
- For $\sigma \in \Sigma$, $F^m(\sigma)$ is non-empty.
- For $\sigma \in \Sigma$, $F^m(\sigma)$ is convex.
- The F^m correspondence has a closed graph.

The first fact was already proven. The following two are immediate application of dynamic-programming results for finite MDPs with discounted cost.

The definition of the set $F^m(\sigma)$ through the Bellman equation guarantees its non-emptiness and convexity.

Let us prove that the F^m correspondence has a closed graph. The function $F^{\mathrm{M},m}$ is continuous. The composition of a continuous function with a correspondence with a closed graph is also a correspondence with a closed graph. Therefore, it is sufficient to show that the $F^{\mathrm{O},m}$ correspondence has a closed graph. Let $\sigma = (\sigma^t)_{t\in\mathbb{N}}$ and $\mu = (\mu^t)_{t\in\mathbb{N}}$ be sequences of strategies and predictors such that for all t in $\mathbb{N}$, $\sigma^t \in F^{\mathrm{O},m}(\mu^t)$. Suppose that σ converges to σ^* and μ converges to μ^*. Let i be an agent, $x_i \in \mathcal{X}_i$, and $z_i \in \mathcal{Z}_i$. For $t \in \mathbb{N}$, the fact that $\sigma^t \in F^{\mathrm{O},m}(\mu^t)$ implies that $\sigma_i^t \in F_i^{\mathrm{O},m}(\mu_i^t)$. This translates to

$$\sigma_i^t(x_i, z_i) \in \underset{a_i\in\mathcal{A}_i}{\arg\max}\, \mathbb{E}_{s_i^+\sim\mu_i^t(z_i)} \Bigg[u_i\big(x_i, a_i, s_i^+\big) + \mathbb{E}\Big[U_{\mu_i^t}\big(x_i^+, z_i^+\big) \,\Big|\, x_i, z_i, a_i, s_i^+\Big]\Bigg],$$

where $U_{\mu_i^t}$ is the value function of the MDP induced by μ_i^t. The arguments used in the proof of the existence of an approximate equilibrium show that the right-hand side is a continuous function of μ_i^t. Therefore, in the limit,

$$\sigma_i^*(x_i, z_i) \in \underset{a_i\in\mathcal{A}_i}{\arg\max}\, \mathbb{E}_{s_i^+\sim\mu_i^*(z_i)} \Bigg[u_i\big(x_i, a_i, s_i^+\big) + \mathbb{E}\big[U_{\mu_i^*}\big(x_i^+, z_i^+\big) \,\big|\, x_i, z_i, a_i, s_i^+\big]\Bigg],$$

Which proves that $\sigma_i^* \in F_i^{\mathrm{O},m}(\mu_i^*)$. This fact is true for any agent and therefore $F^{\mathrm{O},m}$ has a closed graph.

Everything is now in place to apply Kakutani's fixed-point theorem and to conclude that an exact m EEE always exists. □

This is the second time Brouwer's theorem is used before using Kakutani's to prove the existence of an exact equilibrium. This two-step process is centered around the continuity required by Brouwer's theorem. This continuity guarantees that, when a sequence of strategies is optimal with respect to a sequence of parameters, if both sequences converge, then the optimality is still true in the limit. There is not much to gain by using directly Kakutani's theorem as the core of the proof is shared with Brouwer's.

The following section characterizes empirical-evidence equilibria in the setting of perfect-monitoring repeated games.

5.8 Exogenous Empirical-evidence Equilibria in Perfect-monitoring Repeated Games

Having defined a new equilibrium concept, we wanted to compare it with existing ones. We focused on repeated games as these are the most studied stochastic games. Along the way, we found a partial characterization of EEEs in perfect-monitoring repeated games.

Consider a two-player perfect-monitoring repeated game with utilities $u_1 \colon \mathcal{A} \to \mathbb{R}$, $u_2 \colon \mathcal{A} \to \mathbb{R}$ and discount factors δ_1, δ_2. In this game, the agents are using exogenous models m_1 and m_2. Since there is no state in a repeated game, it means that m_i only depends on z_i and s_i^+ but not a_i. The associated agent knowledge diagrams are depicted in Figure 38.

The following proposition shows that, in the present setting, optimality is achievable with a strategy using only the last value of the model state instead of the whole history.

Proposition 8. *Let i be an agent, $\mu_i \colon \mathcal{Z}_i \to \Delta(\mathcal{A}_{-i})$ be a predictor, and $\sigma_i \colon \mathcal{Z}_i \to \Delta(\mathcal{A}_i)$ be a strategy such that the following condition holds:*

$$\forall z_i \in \mathcal{Z}_i,\ \forall a_i' \in \mathcal{A}_i,\ \mathbb{E}[u_i(\sigma_i(z_i), \mu_i(z_i))] \geq \mathbb{E}[u_i(a_i', \mu_i(z_i))].$$

The strategy σ_i is optimal for (u_i, δ_i) with respect to m_i and μ_i.

Proof. The expected payoff at time t depends on the model state z_i^t and the action a_i^t. Since the action has no impact on the model state, myopic optimization is sufficient to guarantee global optimality. □

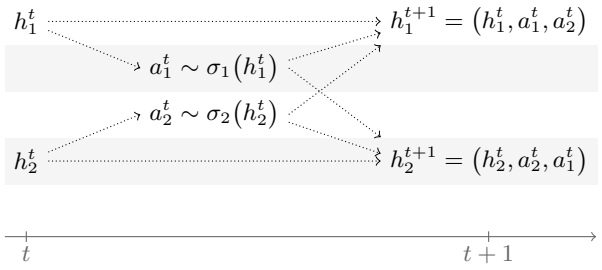
Theorem 8. *Let the pair (μ, σ) be an exogenous empirical-evidence equilibrium (xEEE) for the game (u, δ) with mockups m. Let π be the stationary distribution of the Markov chain over the model states z induced by σ and m.*

The pair (π, σ) is a correlated equilibrium for the one-shot game described by u.

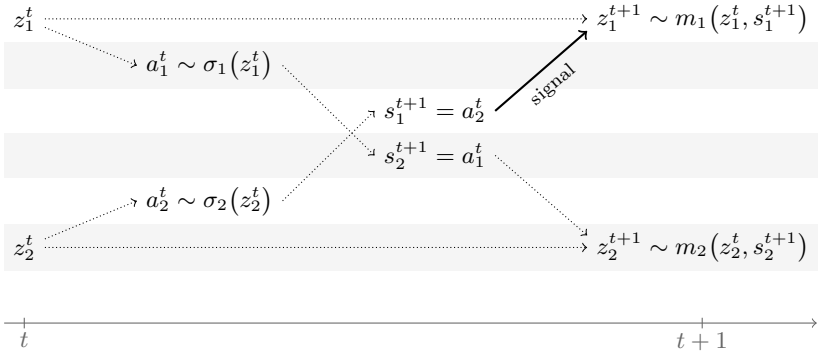
In particular, when all the agents use a memoryless model, the pair (π, σ) is a Nash equilibrium for u.

The careful choice of definitions in the previous sections makes the proof of this theorem straightforward. The key insight of the proof is to interpret the model state z_i as the type of agent i.

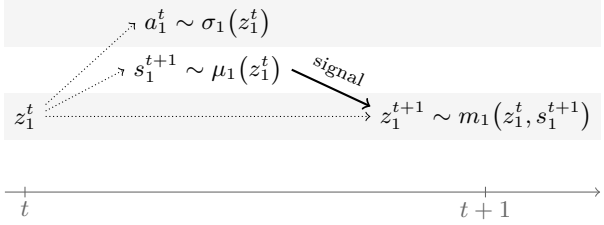
Proof. Fix an agent i. Pick a state $z_i \in \mathcal{Z}_i$ and a signal $s_i^+ = a_{-i} \in \mathcal{S}_i$. By definition of an xEEE, the predictor μ_i is consistent with m and σ. The



(a) *Classical repeated game setting*



(b) *System* $\mathbf{N}$ *in the exogenous empirical-evidence setting*



(c) *System* $\mathbf{M}_1$ *in the exogenous empirical-evidence setting*

Figure 38. *Agent knowledge in a two-player perfect-monitoring repeated game. The highlighted signal highlights how the consistency condition ties* $\mathbf{N}$ *and* $\mathbf{M}_1$ *together.*

consistency condition for state z_i and signal a_{-i} can be rewritten as follows:

$$\begin{aligned}
\mu_i(z_i)[a_{-i}] &= \mathbb{P}_\pi[a_{-i} \mid z_i] \\
&= \sum_{z_{-i} \in \mathcal{Z}_{-i}} \mathbb{P}_\pi[a_{-i} \mid z_i, z_{-i}] \mathbb{P}_\pi[z_{-i} \mid z_i] \\
&= \sum_{z_{-i} \in \mathcal{Z}_{-i}} \mathbb{P}_\pi[a_{-i} \mid z_{-i}] \mathbb{P}_\pi[z_{-i} \mid z_i] \\
&= \sum_{z_{-i} \in \mathcal{Z}_{-i}} \sigma_{-i}(z_{-i})[a_{-i}] \mathbb{P}_\pi[z_{-i} \mid z_i] \\
&= \mathbb{E}_\pi[\sigma_{-i}(z_{-i})[a_{-i}] \mid z_i] \\
&= \mathbb{E}_{Z \sim \pi}[\sigma_{-i}(Z_{-i})[a_{-i}] \mid Z_i = z_i],
\end{aligned}$$

where $\sigma_{-i}(z_{-i})[a_{-i}]$ denotes $\prod_{j \in -i} \sigma_j(z_j)[a_j]$. This equality holds for any a_{-i}, therefore, $\mu_i(z_i) = \mathbb{E}_{Z \sim \pi}[\sigma_{-i}(Z_{-i}) \mid Z_i = z_i]$.

Pick an action $a_i' \in \mathcal{A}_i$. By definition of an xEEE, the strategy σ_i is optimal with respect to μ_i. Substituting the expression for $\mu_i(z_i)$ in the optimality condition of Proposition 8 gives the following inequality:

$$\mathbb{E}_{Z \sim \pi}[u_i(\sigma_i(z_i), \sigma_{-i}(z_{-i})) \mid Z_i = z_i] \geq \mathbb{E}_{Z \sim \pi}[u_i(a_i', \sigma_{-i}(z_{-i})) \mid Z_i = z_i].$$

Interpreting z_i as the type of agent i in Proposition 1 guarantees that the pair (π, σ) is a correlated equilibrium for u. □

The following section illustrates this result.

5.8.1 An Example

In the hawk-dove game, two agents compete for a prize of value 6. The actions of each agent are to be aggressive or passive. In a biological analogy, the aggressive action is called hawk and the passive action is called dove. If only one agent is aggressive, this agent gets the prize to itself. If both agents are aggressive, a fight ensues and both agents are hurt. Finally, if both agents are passive, they split the prize equally. This story is encoded in the following normal-form game:

	h	d
H	$-1, -1$	$6, 0$
D	$0, 6$	$3, 3$

.

The actions of agent 1 are represented by the uppercase letters H and D. Those of agent 2 by their lowercase counterparts h and d. Recall that the actions of one agent are the signals of the other. Using uppercase and lowercase helps in avoiding confusion.

The analysis of the best-response correspondences yields Figure 39. The hawk-dove game has two pure Nash equilibria (H, d) and (D, h), and one mixed Nash equilibrium where the agents play $\frac{3}{4}\mathrm{H} + \frac{1}{4}\mathrm{D}$ and $\frac{3}{4}\mathrm{h} + \frac{1}{4}\mathrm{d}$ respectively.

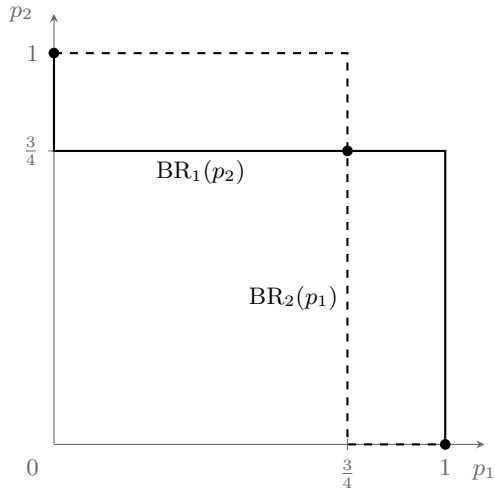


Figure 39. *Best responses and Nash equilibria for hawk-dove. Agent* 1 *plays* H *with probability* p_1*. Agent* 2 *plays* h *with probability* p_2*. The solid line is agent* 1*'s best response. The dashed line is agent* 2*'s best response. The filled circles indicate the Nash equilibria.*

Let us construct an xEEE implementing a correlated-equilibrium distribution. The correlated-equilibrium distribution chosen is the average of the two pure Nash equilibria $\alpha = \frac{1}{2}(\mathrm{H}, \mathrm{d}) + \frac{1}{2}(\mathrm{D}, \mathrm{h})$. The set of correlated-equilibrium distributions is a non-empty convex set containing all the Nash equilibria. Therefore, α is a correlated-equilibrium distribution, even though it is not a Nash equilibrium. Given the symmetric nature of the game, we chose to implement a symmetric equilibrium, meaning that the strategies of the two agents are identical. As a consequence, their predictors are also identical. Both agents use the previously-mentioned depth-2 model.

Let us describe what the solution looks like from agent 1's perspective. Agent 1's state is $z1 = \left(a_2^-, a_2\right)$, where a_2 is the latest observed action of agent 2 and a_2^- the one before that. If agent 1 sees that agent 2 alternates its actions, it supposes that agent 2 acts according to the plan and that this alternation will continue. If agent 2 uses the same action twice in a row, agent 1 is unsure about agent 2's behavior. According to these predictions, agent 1 builds optimal or approximately optimal strategies.

Let us now fill in the details. We provide three variations associated with eventual consistency, consistency, and a new concept called approximate

consistency.

Eventual Consistency

The first variation is closest to the story previously described. The agents use the following predictors associated with their depth-2 models:

$$\begin{aligned}
\mu_1(\mathrm{d},\mathrm{h}) &= \mathrm{d}, & & & \mu_2(\mathrm{D},\mathrm{H}) &= \mathrm{D},\\
\mu_1(\mathrm{h},\mathrm{d}) &= \mathrm{h}, & & & \mu_2(\mathrm{H},\mathrm{D}) &= \mathrm{H},\\
\mu_1(\mathrm{h},\mathrm{h}) &= \frac{3}{4}\mathrm{h}+\frac{1}{4}\mathrm{d}, & \text{and} & & \mu_2(\mathrm{H},\mathrm{H}) &= \frac{3}{4}\mathrm{H}+\frac{1}{4}\mathrm{D},\\
\mu_1(\mathrm{d},\mathrm{d}) &= \frac{3}{4}\mathrm{h}+\frac{1}{4}\mathrm{d}, & & & \mu_2(\mathrm{D},\mathrm{D}) &= \frac{3}{4}\mathrm{H}+\frac{1}{4}\mathrm{D}.
\end{aligned}$$

A pair of associated optimal strategies is

$$\begin{aligned}
\sigma_1(\mathrm{d},\mathrm{h}) &= \mathrm{H}, & & & \sigma_2(\mathrm{D},\mathrm{H}) &= \mathrm{h},\\
\sigma_1(\mathrm{h},\mathrm{d}) &= \mathrm{D}, & & & \sigma_2(\mathrm{H},\mathrm{D}) &= \mathrm{d},\\
\sigma_1(\mathrm{h},\mathrm{h}) &= \frac{1}{2}\mathrm{H}+\frac{1}{2}\mathrm{D}, & \text{and} & & \sigma_2(\mathrm{H},\mathrm{H}) &= \frac{1}{2}\mathrm{h}+\frac{1}{2}\mathrm{d},\\
\sigma_1(\mathrm{d},\mathrm{d}) &= \frac{1}{2}\mathrm{H}+\frac{1}{2}\mathrm{D}, & & & \sigma_2(\mathrm{D},\mathrm{D}) &= \frac{1}{2}\mathrm{h}+\frac{1}{2}\mathrm{d}.
\end{aligned}$$

These strategies induce a Markov chain over the state space $\mathcal{Z}_1 \times \mathcal{Z}_2 = \mathcal{A}_2^2 \times \mathcal{A}_1^2$. By computing the transition matrix, one verifies that this Markov chain is unichain and periodic with period two. Its communication class is $\{(\mathrm{h},\mathrm{d},\mathrm{D},\mathrm{H}),(\mathrm{d},\mathrm{h},\mathrm{H},\mathrm{D})\}$. In the limit, the chain alternates between these two states and the following relations hold:

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[z_1^t = (\mathrm{d},\mathrm{d})\big] &= 0, & & & \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[z_2^t = (\mathrm{D},\mathrm{D})\big] &= 0,\\
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[z_1^t = (\mathrm{h},\mathrm{h})\big] &= 0, & \text{and} & & \lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[z_2^t = (\mathrm{H},\mathrm{H})\big] &= 0.
\end{aligned} \tag{15}$$

In the limit, 14 out of the 16 states do not appear. In particular, any state in which an agent used the same action twice in a row is transient. Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s_1^{t+1} = \mathrm{d} \,\big|\, z_1^t = (\mathrm{d},\mathrm{h})\big] &= 1,\\
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s_1^{t+1} = \mathrm{h} \,\big|\, z_1^t = (\mathrm{h},\mathrm{d})\big] &= 1,\\
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s_2^{t+1} = \mathrm{D} \,\big|\, z_2^t = (\mathrm{D},\mathrm{H})\big] &= 1,\\
\lim_{t\to\infty} \mathbb{P}_{\mathbf{N},\sigma}\big[s_2^{t+1} = \mathrm{H} \,\big|\, z_2^t = (\mathrm{H},\mathrm{D})\big] &= 1.
\end{aligned} \tag{16}$$

Equation (15) guarantees that states (h,h) and (d,d) vanish. Therefore, the values of $\mu_1(\mathrm{h},\mathrm{h})$ and $\mu_1(\mathrm{d},\mathrm{d})$ are arbitrary. The values of $\mu_1(\mathrm{d},\mathrm{h})$

and $\mu_1(\mathrm{h},\mathrm{d})$ match the values observed in (16). Therefore, the predictors are eventually consistent.

Let us now look at the optimality condition. Eventual consistency allows for the predictors to take arbitrary values on transient states (h,h) and (d,d). Therefore, we chose values helping with the optimality condition. The values chosen correspond to the mixed Nash equilibrium in which the agents are indifferent between their two actions. Agent 1 responds optimally in each of the four states z_1.

Therefore, we have constructed an exact xEEE with the notion of eventual consistency which implements the desired correlated-equilibrium distribution.

Consistency

Let us now go from eventual consistency to consistency. The agents use strategies

$$\begin{aligned}
\sigma_1(\mathrm{d},\mathrm{h}) &= 0.999\,\mathrm{H} + 0.001\,\mathrm{D}, & & \sigma_2(\mathrm{D},\mathrm{H}) = 0.999\,\mathrm{h} + 0.001\,\mathrm{d},\\
\sigma_1(\mathrm{h},\mathrm{d}) &= 0.999\,\mathrm{D} + 0.001\,\mathrm{H}, & & \sigma_2(\mathrm{H},\mathrm{D}) = 0.999\,\mathrm{d} + 0.001\,\mathrm{h},\\
\sigma_1(\mathrm{h},\mathrm{h}) &= \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}, & \text{and} \quad & \sigma_2(\mathrm{H},\mathrm{H}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d},\\
\sigma_1(\mathrm{d},\mathrm{d}) &= \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}, & & \sigma_2(\mathrm{D},\mathrm{D}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d}.
\end{aligned}$$

The induced Markov chain over $\mathcal{Z}_1 \times \mathcal{Z}_2$ is irreducible and aperiodic. No state is transient. Therefore, predictors have to be defined for all states. The consistent predictors are

$$\begin{aligned}
\mu_1(\mathrm{d},\mathrm{h}) &= 0.996\,\mathrm{d} + 0.004\,\mathrm{h}, & & \mu_2(\mathrm{D},\mathrm{H}) = 0.996\,\mathrm{D} + 0.004\,\mathrm{H},\\
\mu_1(\mathrm{h},\mathrm{d}) &= 0.996\,\mathrm{h} + 0.004\,\mathrm{d}, & \text{and} \quad & \mu_2(\mathrm{H},\mathrm{D}) = 0.996\,\mathrm{H} + 0.004\,\mathrm{D},\\
\mu_1(\mathrm{h},\mathrm{h}) &= 0.5\,\mathrm{h} + 0.5\,\mathrm{d}, & & \mu_2(\mathrm{H},\mathrm{H}) = 0.5\,\mathrm{H} + 0.5\,\mathrm{D},\\
\mu_1(\mathrm{d},\mathrm{d}) &= 0.5\,\mathrm{h} + 0.5\,\mathrm{d}, & & \mu_2(\mathrm{D},\mathrm{D}) = 0.5\,\mathrm{H} + 0.5\,\mathrm{D}.
\end{aligned}$$

The probabilities reported as 0.5 are not exactly $\frac{1}{2}$. These probabilities are biased towards the symbol just observed twice with a bias of the order of 10^{-12}.

In this setting, consistency is immediate, by definition of the predictors. Optimality is slightly trickier. Recall that in an xEEE for a perfect-monitoring repeated game, the optimality condition translates to myopic optimality. Therefore, σ_1's optimality with respect to μ_1 and m_1 is equivalent to the following condition. For all $z_1 \in \mathcal{Z}_1$, agent 1's mixed action $\sigma_1(z_1)$ is a best response to agent 2's mixed action $\mu_1(z_1)$. This is not the case for the states (d,d) and (h,h). As seen in Figure 39, agent 1's sole best-response to $0.5\,\mathrm{h} + 0.5\,\mathrm{d}$ is H. However, σ_1 is approximately optimal with respect

to μ_1 for discount factors δ_1 close enough to one. Most of the time is spent in states (d,h) and (h,d) for which σ_1 is optimal. By taking δ_1 close enough to one, the effect of acting non-optimally in the other two states becomes negligible.

Proposition 8 proves that myopic optimality is sufficient to get optimality. This example illustrates that approximate optimality does not require approximate myopic optimality. Strategies can perform poorly on vanishing states.

The resulting equilibrium is an approximate xEEE. The associated distribution over actions

$$\begin{array}{ll} 0.004\,(\mathrm{H},\mathrm{h}) & 0.496\,(\mathrm{H},\mathrm{d}) \\ 0.496\,(\mathrm{D},\mathrm{h}) & 0.004\,(\mathrm{D},\mathrm{d})\text{'} \end{array}$$

is an approximation of the desired correlated-equilibrium distribution.

Approximate Consistency

To conclude this example, let us analyze what happens when approximately consistent predictors are used. The agents use the same smoothed strategies

$$\begin{array}{lll} \sigma_1(\mathrm{d},\mathrm{h}) = 0.999\,\mathrm{H} + 0.001\,\mathrm{D}, & & \sigma_2(\mathrm{D},\mathrm{H}) = 0.999\,\mathrm{h} + 0.001\,\mathrm{d}, \\ \sigma_1(\mathrm{h},\mathrm{d}) = 0.999\,\mathrm{D} + 0.001\,\mathrm{H}, & & \sigma_2(\mathrm{H},\mathrm{D}) = 0.999\,\mathrm{d} + 0.001\,\mathrm{h}, \\ \sigma_1(\mathrm{h},\mathrm{h}) = \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}, & \text{and} & \sigma_2(\mathrm{H},\mathrm{H}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d}, \\ \sigma_1(\mathrm{d},\mathrm{d}) = \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}, & & \sigma_2(\mathrm{D},\mathrm{D}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d}, \end{array}$$

and these predictors

$$\begin{array}{lll} \mu_1(\mathrm{d},\mathrm{h}) = 0.999\,\mathrm{d} + 0.001\,\mathrm{h}, & & \mu_2(\mathrm{D},\mathrm{H}) = 0.999\,\mathrm{D} + 0.001\,\mathrm{H}, \\ \mu_1(\mathrm{h},\mathrm{d}) = 0.999\,\mathrm{h} + 0.001\,\mathrm{d}, & & \mu_2(\mathrm{H},\mathrm{D}) = 0.999\,\mathrm{H} + 0.001\,\mathrm{D}, \\ \mu_1(\mathrm{h},\mathrm{h}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d}, & \text{and} & \mu_2(\mathrm{H},\mathrm{H}) = \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}, \\ \mu_1(\mathrm{d},\mathrm{d}) = \frac{1}{2}\mathrm{h} + \frac{1}{2}\mathrm{d}, & & \mu_2(\mathrm{D},\mathrm{D}) = \frac{1}{2}\mathrm{H} + \frac{1}{2}\mathrm{D}. \end{array}$$

These predictors are necessarily inconsistent. However, they are approximately consistent, meaning close to the consistent ones. Since the strategies are approximately optimal for the consistent predictors, Lemma 1 guarantees they are also approximately optimal for the approximately consistent predictors. Therefore, the pair (σ,μ) forms an approximate EEE in the sense of Definition 18.

This example explains why we did not formally define the notion of approximate consistency. Lemma 1 guarantees that we can trade approximate consistency for approximate optimality. For the sake of clarity, it is easier

to only have approximation in one place and we chose to have approximate optimality.

5.8.2 Average of Nash Equilibria

The example of the previous section easily extends to general finite games and yields a large set of correlated-equilibrium distributions. Most of the work for the proof has already been done in the example.

Theorem 9. *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\left(a^l\right)_{l=1}^k$ be k, non-necessarily distinct, pure Nash equilibria of u. Denote by $\alpha \in \Delta(\mathcal{A})$ the average of these Nash equilibria*

$$\alpha = \frac{1}{k}\sum_{l=1}^{k} a^l,$$

which is a correlated-equilibrium distribution.

For large enough discount factors $(\delta_i)_{i\in\mathcal{I}}$, α is implementable by an approximate xEEE, in which each agent uses a depth-k eventually consistent model.

Proof. Let $i \in \mathcal{I}$ be an agent. Define its depth-k predictor as follows:

$$\begin{aligned}
\mu_i\left(a^1_{-i}, a^2_{-i}, \cdots, a^{k-1}_{-i}, a^k_{-i}\right) &= a^1_{-i},\\
\mu_i\left(a^2_{-i}, a^3_{-i}, \cdots, a^k_{-i}, a^1_{-i}\right) &= a^2_{-i},\\
&\vdots\\
\mu_i\left(a^k_{-i}, a^1_{-i}, \cdots, a^{k-2}_{-i}, a^{k-1}_{-i}\right) &= a^k_{-i},
\end{aligned}$$

and for all the other states z_i,

$$\mu_i(z_i) = \frac{1}{k}\sum_{l=1}^{k} a^k_{-i}.$$

As opposed to the example, a mixed Nash equilibrium does not always exist. This reduced flexibility in defining the predictor on vanishing states, explains why only approximate xEEEs are guaranteed.

Define agent i's strategy as follows:

$$\begin{aligned}
\sigma_i\left(a^1_{-i}, a^2_{-i}, \cdots, a^{k-1}_{-i}, a^k_{-i}\right) &= a^1_i,\\
\sigma_i\left(a^2_{-i}, a^3_{-i}, \cdots, a^k_{-i}, a^1_{-i}\right) &= a^2_i,\\
&\vdots\\
\sigma_i\left(a^k_{-i}, a^1_{-i}, \cdots, a^{k-2}_{-i}, a^{k-1}_{-i}\right) &= a^k_i,
\end{aligned}$$

and for all the other states z_i,

$$\sigma_i(z_i) = \frac{1}{k}\sum_{l=1}^{k} a_i^k.$$

For a vanishing state z_i, the definition of $\mu_i(z_i)$ could have been anything. The definition of $\sigma_i(z_i)$ requires that each action appearing in one of the k pure Nash equilibria appear with positive probability.

The induced Markov chain is unichain and periodic with period k. Its communication class has k states corresponding to each of the k Nash equilibria. In the limit, the chain cycles through these k states in the order imposed by the labeling of the equilibria. The eventual consistency of the predictors is proven as in the example.

As previously mentioned, it is not always possible to guarantee optimality of σ_i with respect to μ_i. However, σ_i is optimal for all the states visited in the limit. The lack of optimality is only for vanishing states. Therefore, a large enough discount factor guarantees approximate optimality. □

Corollary 1. *Let $u\colon \mathcal{A} \to \mathbb{R}^{|\mathcal{I}|}$ describe a one-shot game. Let $\left(a^l\right)_{l=1}^{k}$ be k, non-necessarily distinct, pure Nash equilibria of u. Denote by $\alpha \in \Delta(\mathcal{A})$ the average of these Nash equilibria*

$$\alpha = \frac{1}{k}\sum_{l=1}^{k} a^l,$$

which is a correlated-equilibrium distribution.

For large enough discount factors $(\delta_i)_{i\in\mathcal{I}}$, α can be approximated by a correlated-equilibrium distribution induced by an approximate xEEE, in which each agent uses a depth-k consistent model.

Proof. Let $i \in \mathcal{I}$ be an agent and $\varepsilon > 0$. Denote by ω_i the uniform distribution over $\mathcal{A}_i$. Define agent i's strategy as follows:

$$\begin{aligned}
\sigma_i\left(a_{-i}^1, a_{-i}^2, \cdots, a_{-i}^{k-1}, a_{-i}^k\right) &= (1-\varepsilon)a_i^1 + \varepsilon\omega_i,\\
\sigma_i\left(a_{-i}^2, a_{-i}^3, \cdots, a_{-i}^k, a_{-i}^1\right) &= (1-\varepsilon)a_i^2 + \varepsilon\omega_i,\\
&\vdots\\
\sigma_i\left(a_{-i}^k, a_{-i}^1, \cdots, a_{-i}^{k-2}, a_{-i}^{k-1}\right) &= (1-\varepsilon)a_i^k + \varepsilon\omega_i,
\end{aligned}$$

and for all the other states z_i,

$$\sigma_i(z_i) = (1-\varepsilon)\frac{1}{k}\sum_{l=1}^{k} a_i^k + \varepsilon\omega_i.$$

The joint strategy $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ induces an irreducible aperiodic Markov chain. Define $\mu = (\mu_i)_{i\in\mathcal{I}}$ as the associated consistent predictors.

Using the same proofs as in the example, one shows that, for ε small enough, the pair (σ, μ) forms an approximate xEEE. Furthermore, this construction induces an approximation of the correlated-equilibrium distribution α. □

5.9 Learning Empirical-evidence Equilibria

5.9.1 A Learning Rule

The fixed points of $G^{m,\varepsilon}$ 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:

$$\mu^{t+1} = \mu^t + \alpha^t\big(G^{m,\varepsilon}\big(\mu^t\big) - \mu^t\big), \tag{17}$$

where α^t is a step size. The long-run behavior of (17) is related to properties of the following differential equation:

$$\dot{\mu} = G^{m,\varepsilon}(\mu) - \mu. \tag{18}$$

In particular, Benaïm showed that the limit set of (17) is a connected set internally chain recurrent for the flow induced by $G^{m,\varepsilon} - \mathrm{Id}$, where Id is the identity function [46]. The following, easily verified, proposition tells us that, if (17) converges, it might yield an (ε, m) EEE.

Proposition 9. *The fixed points of $G^{m,\varepsilon}$ are connected sets internally chain recurrent for the flow induced by $G^{m,\varepsilon} - \mathrm{Id}$.*

Note, however, that these fixed points might not be the only connected sets internally chain recurrent for this flow.

The existence of a Lyapunov function for (18) guarantees that there is a unique connected sets internally chain recurrent for the flow induced by $G^{m,\varepsilon} - \mathrm{Id}$ which is the equilibrium point. Therefore, if there exists such a Lyapunov function for the continuous system, we can conclude that (17) converges to an EEE.

5.9.2 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, $\mathcal{A} = \{\mathrm{Sell}, \mathrm{Hold}, \mathrm{Buy}\}$. The assets can be traded at a low price or at a high price, $\mathcal{S} = \{\mathrm{Low}, \mathrm{High}\}$. 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 (17) were simulated. The first one used (17) directly with a fixed step size of $\alpha^t = 0.1$. The stationary distribution π_σ was computed at each time step to obtain the true value of $G^{m,\varepsilon}(\mu^t)$. The algorithm is presented in Algorithm 3. The results of the simulations using the theoretical predictor are presented in Figure 40. 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.

In the second simulated version of (17), the stationary distribution is 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 algorithm is presented in Algorithm 4. The results of the simulation using the empirical predictors are presented in Figure 41. 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.

5.9.3 Effect of the Finite Observation Window

In the empirical simulation, approximate predictors are used. The function $F^{\mathrm{M},m}$ computes predictors consistent with given strategies through the stationary distribution of a Markov chain. The function $F_l^{\mathrm{M},m}$ is obtained when approximate predictors are instead computed from l rounds of play. The following proposition shows that important properties of $F^{\mathrm{M},m}$ are recovered by $F_l^{\mathrm{M},m}$, as l goes to infinity. A proof sketch for this proposition is given to highlight useful tools in this context. Given a real problem, the proposition can be adapted, using the appropriate textbook material on stochastic approximation and Lyapunov stability of perturbed systems.

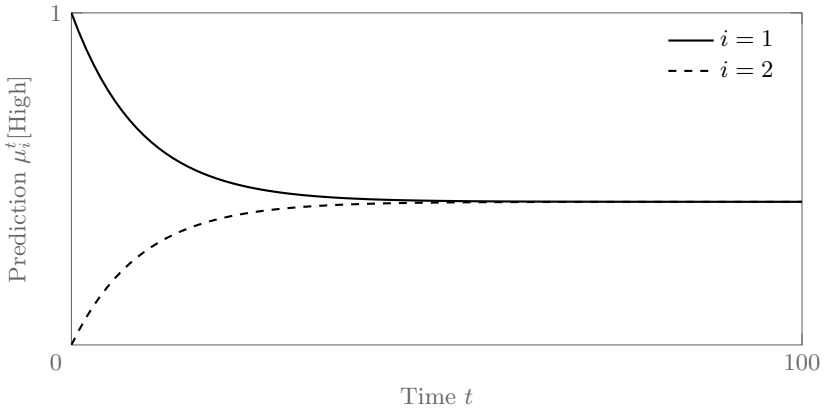


Figure 40. *Simulation results of two agents learning a depth-0 model of the price for the market example with the theoretical predictor.*

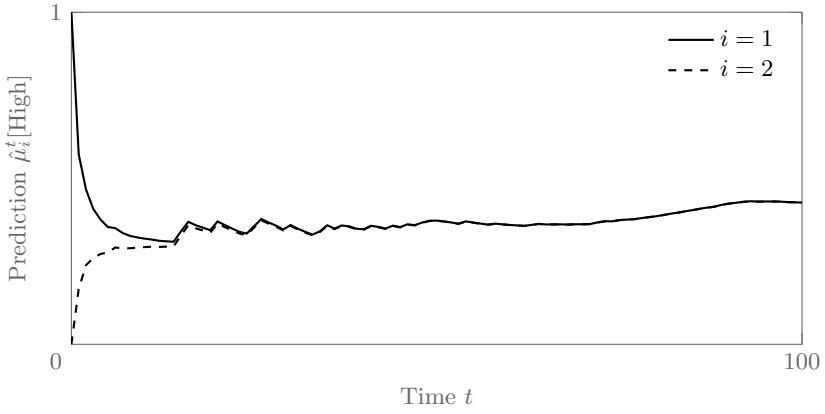


Figure 41. *Simulation results of two agents learning a depth-0 model of the price for the market example with an empirical predictor.*

Algorithm 3 Learning with Theoretical Predictor

procedure THEORETICAL PREDICTOR LEARNING(ε)
  $\mu_1 \leftarrow 1$
  $\mu_2 \leftarrow 0$
  for $t \in [1, 100]$ **do**
    $\sigma_1 \leftarrow$ an optimal strategy for μ_1
    $\sigma_2 \leftarrow$ an optimal strategy for μ_2
    $T_\sigma \leftarrow$ the transition matrix for the Markov chain induced by σ_1 and σ_2
    $\pi_\sigma \leftarrow$ the eigenvector associated with the eigenvalue 1 for T_σ
    $\hat{\mu} \leftarrow \mathbb{P}_{\pi_\sigma}[\text{High}]$
    $\mu_1 \leftarrow \mu_1 + 0.1(\hat{\mu} - \mu_1)$
    $\mu_2 \leftarrow \mu_2 + 0.1(\hat{\mu} - \mu_2)$
  end for
end procedure

Algorithm 4 Learning with Empirical Predictor

procedure EMPIRICAL PREDICTOR LEARNING(ε)
  $\mu_1 \leftarrow 1$
  $\mu_2 \leftarrow 0$
  for $t \in [1, 100]$ **do**
    $\sigma_1 \leftarrow$ an optimal strategy for μ_1
    $\sigma_2 \leftarrow$ an optimal strategy for μ_2
    $h \leftarrow 0$
    for $\tau \in [1, 100]$ **do**
      agent 1 places an order according to σ_1
      agent 2 places an order according to σ_2
      if the observed price is High **then**
        $h \leftarrow h + 1$
      end if
    end for
    $\hat{\mu} \leftarrow \frac{h}{100}$
    $\mu_1 \leftarrow \mu_1 + \left(\frac{1}{t}\right)^{\frac{3}{4}}(\hat{\mu} - \mu_1)$
    $\mu_2 \leftarrow \mu_2 + \left(\frac{1}{t}\right)^{\frac{3}{4}}(\hat{\mu} - \mu_2)$
  end for
end procedure

Assumption 4. *When computing $F_l^{\mathrm{M},m}$, at the beginning of the l rounds of play, the state of Nature w, the states of the agents x, the state of the models z, and the seed to the pseudorandom number generators are reset.*

Proposition 10. *Under Assumption 4, if there exists a Lyapunov function for the following continuous-time system:*

$$\dot{\mu} = G^{m,\varepsilon}(\mu) - \mu, \tag{19}$$

then for l large enough, the following discrete-time system, using approximate predictors, converges to an EEE:

$$\hat{\mu}^{t+1} = \hat{\mu}^t + \alpha^t\big(G_l^{m,\varepsilon}\big(\hat{\mu}^t\big) - \hat{\mu}^t\big), \tag{20}$$

where $G_l^{m,\varepsilon} = F_l^{\mathrm{M},m} \circ F^{\mathrm{O},m,\varepsilon}$ and $(\alpha^t)_{t\in\mathbb{N}}$ is a non summable but square summable sequence of positive numbers.

The proof starts with the use of dynamical system

$$\dot{\hat{\mu}} = G_l^{m,\varepsilon}(\hat{\mu}) - \hat{\mu}, \tag{21}$$

and Lyapunov stability of perturbed systems such as [47, Lemma 9.1 or Lemma 9.2]. For a large enough l, dependent on the chosen lemma, (21) constitutes a perturbed version of (19). Therefore, for l large enough, a Lyapunov function for (19) is also a Lyapunov function for (21). The existence of a Lyapunov function for (21) implies that the only connected set internally chain recurrent for the flow induced by $G_l^{m,\varepsilon} - \mathrm{Id}$ is the singleton containing the equilibrium point. Assumption 4 allows the application of deterministic stochastic-approximation results. In particular, [46, Th. 1.2] guarantees that the limit set of the sequence $(\hat{\mu}^t)_{t\in\mathbb{N}}$, solution to (20), is a connected set internally chain recurrent for the flow induced by $G_l^{m,\varepsilon}$. Therefore, $(\hat{\mu}^t)_{t\in\mathbb{N}}$ converges to an EEE.

6 Conclusion

We developed the framework of EEEs in stochastic games. 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 previously exposed, 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 this dissertation, we laid down the foundations of a general consistency framework. In this framework, EEEs have emerged as a solution concept. We proved the existence of EEEs for a general setting. We provided a characterization of EEEs in perfect-monitoring repeated games. Finally, we explored the learning of EEEs with a particular interest for the finite observation window case. Some other interesting open questions are listed below.

6.1 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 [42]. Since a depth-k consistent model is depth-$k-1$

consistent as well, a larger k is synonymous with better understanding of the environment. They proved that agents with a larger k did not always receive a larger payoff. This question has to be addressed for a more general setting.

As k increases, the agent gets a more accurate prediction of the strings of signals. This raises the question to know what happens in the limit.

6.2 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.

6.3 Learning

EEEs were informally defined as fixed points of a simple iterative process. The existence of fixed points has been established. 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 5.9. 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 [26]. The adaptive learning rule replaced a matrix inversion step by a simpler maximization one.

6.4 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 [48]. 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?

6.5 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?

Bibliography

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

[2] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67–96, Mar. 1974.

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

[4] S. Mac Lane, *Categories for the Working Mathematician*. New York: Springer, Sep. 1998.

[5] D. I. Spivak, *Category Theory for the Sciences*. Cambridge, MA: MIT Press, 2014.

[6] "The Haskell Programming Language," Feb. 2014. [Online]. Available: http://www.haskell.org

[7] Y. Shoam and K. E. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge, MA: Cambridge University Press, Dec. 2008.

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

[9] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

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

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

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

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

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

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

[16] J. C. Ely and J. Välimäki, "A robust folk theorem for the prisoner's dilemma," *Journal of Economic Theory*, vol. 102, no. 1, pp. 84–105, Jan. 2002.

[17] J. C. Ely, J. Hörner, and W. Olszewski, "Belief-free equilibria in repeated games," *Econometrica*, vol. 73, no. 2, pp. 377–415, Mar. 2005.

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

[19] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences of the USA*, vol. 39, no. 10, pp. 1095–1100, Oct. 1953.

[20] S. Hart and A. Mas-Colell, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, no. 1, pp. 26–54, May 2000.

[21] ——, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.

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

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

[24] 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.

[25] 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.

[26] S. Hart and A. Mas-Colell, *Economic Essays*. New York: Springer, 2001, ch. A reinforcement procedure leading to correlated equilibrium, pp. 181–200.

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

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

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

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

[31] 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.

[32] 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.

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

[34] 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.

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

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

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

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

[39] 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.

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

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

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

[47] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 2002.

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

Vita

Nicolas Dudebout is a PhD candidate in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. His research is in control theory with an emphasis on game-theoretic learning. In 2008, he received a Masters in Electrical and Computer Engineering from Georgia Tech.

Born and raised in France, Nicolas attended *classes préparatoires* at Lycée Condorcet in Paris. Upon completion in 2005, Nicolas was admitted into Supélec in Paris to pursue a Bachelor of Science in Electrical and Computer Engineering. He graduated first in his class in 2007.

When he is not in the lab, Nicolas can be found contributing to open-source projects, playing tennis, cooking or spending time with his wife, Laura, and infant daughter, Claire.

Artwork by Ben J. Adams