Learning to Play Network Games:

Does Rationality Yield Nash Equilibrium?

by

Amy Rachel Greenwald

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
Graduate School of Arts and Science
New York University
May 1999

Advisor: Bhubaneswar Mishra

In loving memory
of
my Grandpa Ralph

Abstract

This thesis investigates the design and implementation of automated intelligent agents suitable for controlling and optimizing the utilization of telecommunication networks. Since many issues that arise in networking are inherently resource allocation problems, the mathematical underpinnings of this research are found in economics, particularly, non-cooperative game theory. Specific to this approach is the application of the theory of repeated games, which emphasizes learning to play equilibrium strategies, in contexts that model telecommunication network environments. This methodology is herein referred to as computational game theory.

The main theoretical result derived in this thesis is that rational learning does not converge to Nash equilibrium in network games without pure strategy Nash equilibria. This result has important implications for the networking community, where it is often assumed both that players are rational and that the network operating point is a Nash equilibrium. In addition, it is observed via simulations that low-rationality learning, where agents trade-off between exploration and exploitation, typically converges to mixed strategy Nash equilibria. In the presence of extreme asynchrony, however, even if players exhibit low-rationality learning, Nash equilibrium is nonetheless not an appropriate solution in games that model network interactions.

This thesis is also concerned with the reverse migration of ideas, specifically the application of computational learning in the economic world of electronic commerce. In particular, the dynamics of interactions among *shopbots*, who seek to minimize costs for buyers, and *pricebots*, who aim to maximize profits for sellers, are researched. As in network control games, rational learning does not converge to Nash equilibrium; certain low-rationality learning algorithms, however, do in fact converge to mixed strategy Nash equilibria in shopbot economics.

Acknowledgments

This thesis would not have been possible were it not for the unyielding patience of Dean Foster. Dean is an individual whose interdisciplinary expertise serves as an inspiration to me in every aspect of my research, from game theory to computer simulation. Dean's influences will undoubtedly guide me for many years beyond my graduate study. Thank you, Dean, for the career guidance, and of course, for the sushi! Next, I would like to thank Rohit Parikh. First of all, thank you Rohit for feeding me many an Indian meal during stressful times of research and writing that left little time for cooking of my own. Even more importantly, thank you for being an exquisite proofreader of my writings, not only correcting typographical and grammatical errors (any remaining errors surely are my own), but in addition, introducing new ideas for tangential problems of study along the way. I hope we can continue to pursue joint research in years to come. Many thanks as well to Bud Mishra, who believed in me, and demonstrated as much by allowing me to lead the interdisciplinary life of my liking, in spite of words of warning from others that this might hinder my search for an academic position in computer science. In addition, I'd like to thank Scott Shenker, for giving me my first big break, an internship at Xerox PARC, which allowed me to get my feet off the ground. Scott, I thank you for showing me how to actually get things done, in the unstructured world of graduate student life. Last, but by no means least, I thank Jeff Kephart, the most recent influence on my research, my career, and consequently my life. During the past year, Jeff has served as my mentor, guiding my research and teaching me a great many things. Whenever I felt anxious or insecure during my job interviews, I suppressed those feelings by remembering the words of encouragement and support I constantly receive from Jeff. Thank you, Jeff, for always making time for me in your already overbooked schedule.

In addition to my academic influences, this thesis would not have been possible without the undying love and support of my immediate family. Many thanks, Mom, for abandoning your dream that I become a Wall Street tycoon, and supporting my career as a scientist. I know now what you have said all along: "I can be anything I *want* to be." And thanks Dad, for providing the role model of a scientist and inspiring the scientist in me ever since I was young girl with interests in math and computing. Michele and Carolyn, thanks for being my sisters and my best friends, right by my side through all the ups and downs of life in New York City. What a year!

In addition, I'd like to thank Ron Even and Peter Wyckoff, for being supportive friends and proofreaders, Fabian Monrose and Toto Paxia, for installing Linux on my various machines, thereby enabling the writing of this thesis, and Arash Baratloo for introducing me to the Ph.D. program in computer science at NYU.

Amy Greenwald New York City, 1999

Contents

Chapter 0

Introduction

In recent years, there has been a dramatic expansion of the global communications infrastructure, in terms of the bandwidth of telecommunication links, the degree of connectivity, and the intelligence embedded in network switches. This growth has been accompanied by the potential for creating of a new range of network capabilities, such as software agents that collaboratively mine and warehouse information on the Internet, multi-media data transfer using shared bandwidth and buffer space, and secure financial infrastructures supporting electronic commerce.

The challenges associated with these developments, however, are many, primarily because any realistic resource allocation scheme on this massive a scale cannot rely on the cooperative behavior of network entities or on complete knowledge of network properties. Moreover, given the increasingly dynamic nature of large-scale computer networks — links are constantly coming up and going down, and servers often crash or need to be rebooted — adaptive control is essential, in order to ensure robustness and maintain efficiency in the face of unexpected environmental changes.

This thesis investigates the design and analysis of automated intelligent agents suitable for control and optimization in telecommunication networks. Since many networking issues are inherently problems of resource allocation, the mathematical underpinnings of this research are found in economics, particularly, non-cooperative game theory. Specific to this approach is the application of the theory of repeated games, which emphasizes learning to play equilibrium strategies, in contexts that model telecommunication network environments.

0.1 Economics in Computer Science

Since its inception, the model of computation proposed by von Neumann consisting of a solitary sequential processor has served as a foundation for the theory of computing. The explosive growth of the telecommunications industry, however, exemplified by remarkable public interest in the Internet, electronic commerce (e-commerce), and multi-media, demands techniques for resource allocation and process coordination that go beyond standard computational methodology. It is argued by Schneider [97] that computer scientists must refine their intuition in order to successfully approach the problems that arise in decentralized computing:

> Distributed systems are hard to design and understand because we lack intuition for them... In any event, distributed systems are being built. We must develop an intuition, so that we can design distributed systems that perform as we intend... and so that we can understand existing distributed systems well enough for modification as needs change.

Unlike the field of computer science, theories on the interaction of complex agents in distributed environments form an integral part of mathematical economics. In particular, economics provides elegant models in which to describe optimal ways of sharing resources and coordinating behavior in multi-agent systems, where decisions are often based on local, delayed, and conflicting information. In situations where traditional computational paradigms fail to yield viable solutions, economics provides a robust framework in which to study dynamic behavior among intelligent agents in informationally and computationally decentralized settings.

An expanding body of literature on systems which apply economic ideas to control and optimization problems in computer science expounds the merit of this point of view. For example, SPAWN is a computational economy designed at Xerox PARC to manage and coordinate distributed tasks on multiple processors [112]. Similarly, WALRAS is an asynchronous distributed system developed jointly at Michigan and MIT which operates via a market pricing mechanism [22]. Finally, CAFÉ (Complex Adaptive Financial Environment) [34] is a market simulator implemented at NYU in which it is possible to model ATM networks in an economic environment.

0.2 Game Theory in Computer Science

Game theory, a field whose conception is also attributed to von Neumann, *et al.* [111], is the branch of economics which is of particular interest in this thesis. Game theory is far reaching in its applications to the social sciences: *e.g.*, politics and psychology; similarly, this work explores an application of game theory in the field of computer science. In particular, game theory is proposed as a model of distributed computation. The idea of modeling distributed computation as games was suggested previously by such prominent game theorists as Aumann [5] and Rubinstein [90]. Specific to this approach is the application of *computational game theory* (*i.e.*, repeated game theory together with computational learning theory) to the study of network interactions.

Network games are interactions in which (i) players are automated agents, either mobile or residing on host machines, acting on behalf of human users (ii) strategies are requests for shared resources, often given by specifying transmission rates and routes, and (iii) utilities are determined by the collective strategic behavior of all agents, measured in terms of network properties such as delay, loss, and jitter. For example, the interactions of Internet users over shared bandwidth and buffer space yield one class of network games. Today, most Internet flow control decisions are specified by the TCP protocol, which states that upon detecting congestion, machines should halve their rate of transmission [62]. Consequently, the efficient operation of the Internet crucially depends upon cooperation on the part of its users. In contrast, game-theoretic reasoning is based on the assumption of non-cooperative behavior. Given the scale of today's Internet and its increasing rate of growth,[1] non-cooperative game theory is suited to the study of network interactions [59, 71, 99, 100, 118]. Classical game theory, however, relies on the assumption that players have complete knowledge of the underlying structure of games as well as common knowledge of rationality. This thesis distinguishes itself from other literature on network games by building on the premise that *learning is essential in network games*, since complete and common knowledge of network properties is not generally obtainable.

[1] Today the Internet boasts 180 million users [88] with 900 million users expected by the year 2004 [1]. In contrast, around the time of the specification of the TCP protocol, the names and addresses of all the registered users of the Internet were listed, approximately 24 users per page, in the 680 page Digital Data Network Directory [28, 105].

As it is not appropriate to assume guaranteed cooperation on the part of agents in network interactions, this thesis focuses on the study of networks via game theory; moreover, since it is also not appropriate to assume unlimited access to knowledge in network environments, this thesis investigates learning in network games. The subject of this research is the fundamental question of computational game theory[2] in the case of network games: *i.e.*, *What is the outcome of learning in network games?* Naturally, the answer to this question is highly dependent upon the particular assumptions that are made about the learning behavior of agents, which is closely related to their degree of rationality. In this regard, the key findings of this thesis are as follows:

1. Rational learning does *not* give rise to Nash equilibrium in network games; on the contrary, together rationality and learning yield paradoxical outcomes.

2. Low-rationality learning, where agents continually trade-off between exploration and exploitation, *does* converge to Nash equilibrium in network games.

3. Responsive learning, namely low-rationality learning based on finite histories, does *not* lead to Nash equilibrium in network contexts, where play is assumed to be highly asynchronous.

In short, Nash equilibrium is *not* the outcome of either rational or responsive learning in games that model network environments. These results invalidate the prevailing assumption in the literature on network games which claims that Nash equilibrium describes the operating point of networks (see, for example, [100]). The following subsections elaborate on these findings in terms of two sample network games, namely the Santa Fe bar problem and shopbot economics.

0.2.1 Learning

The *Santa Fe bar problem* (SFBP) was introduced by Brian Arthur [2] in the study of inductive learning and bounded rationality. This problem and its natural extensions also serve as abstractions of network games, since they model network flow control and routing problems. Here is the scenario:

[2] The fundamental question of computational game theory in general is as follows: *What is the outcome of computational learning in repeated games?*

> N (say, 100) people decide independently each week whether to go to a bar that offers entertainment on a certain night ... Space is limited, and the evening is enjoyable if things are not too crowded – especially, if fewer than 60 percent of the the possible 100 are present ... a person or agent *goes* (deems it worth going) if he expects fewer than 60 to show up or *stays home* if he expects more than 60 to go.
>
> Choices are unaffected by previous visits; there is no collusion or prior communication among the agents; and the only information available is the number who came in past weeks.[3]

Arthur first analyzed SFBP assuming the inhabitants of Santa Fe to be *rational* (*i.e.*, optimizing). In particular, he argued intuitively that learning and rationality are incompatible in SFBP. Consider, for example, Cournot best-reply dynamics [24], which is a learning mechanism by which agents play a rational strategy in response to the belief that the other agents' actions at the next time step will mimic their most recent actions. If attendance at the bar at time t is less than or equal to capacity, then the Cournot best-reply at time $t+1$ is to go to the bar; but then, attendance at the bar at time $t+1$ is greater than capacity, which implies that the Cournot best-reply at time $t+2$ is not to go to the bar, and so on.[4] Over time, the following is the result of learning via Cournot best-reply dynamics in SFBP: (i) agents' strategies do not converge: *i.e.*, attendance at the bar does not converge, and (ii) agents beliefs about other agents' strategies never reflect the others' actual behavior. In particular, play does not converge to Nash equilibrium. In this thesis, it is shown that no rational learning algorithm, of which Cournot best-reply dynamics and Bayesian updating are examples, converges to Nash equilibrium in SFBP. On the other hand, it is observed via simulations, that low-rationality, non-Bayesian learning, where agents trade-off between exploration and exploitation, typically does yield convergence to Nash equilibrium in SFBP.

[3] The problem was inspired by the El Farol bar in Santa Fe which offers live music on Thursday nights.

[4] Schelling [96] refers to phenomena of this kind as self-negating prophecies.

0.2.2 Network Games

This thesis is concerned with computational learning in *network games*. The context of network games differs from the traditional game-theoretic context in four important ways, which are testimony to the fact that learning is essential in network games. The four essential properties of network contexts, namely limited information, dynamic structure, automation, and asynchrony, are described below:

1. Agents have extremely limited information pertaining to the characteristics of the shared resource; in other words, they do not know the underlying structure of the game. Moreover, agents are not explicitly aware of the existence of other agents, as there is no way of directly observing the presence of others, and they are therefore incapable of accurately modeling their opponents.

2. The structure of the game, in terms of agents, strategies, and utilities, are all subject to change over time. Shared resources like network links and servers periodically crash, and often experience other unpredictable changes in their capabilities, such as upgrades or route changes. Moreover, users of network resources come and go frequently.

3. Play is often carried out by computer algorithms, rather than by human users. For instance, congestion control algorithms (*e.g.*, TCP), which are embedded in operating systems, manage the sharing of network links. Similarly, automated algorithms are being designed that control the retry behavior for submission of queries to shared databases.

4. Games are played in an asynchronous fashion, without any notion of definable rounds of play, since the rates at which agents adapt their strategies vary widely. Due to the geographic dispersion of the Internet, for example, communication delays to shared resources may differ by several orders of magnitude; moreover, processor speeds tend to vary substantially. Agents closer to shared resources, and those who are privy to faster processors or smarter algorithms, have the potential to learn more rapidly and more effectively.

The following questions are investigated in this thesis: (i) What sort of collective behavior emerges via low-rationality, responsive learning among a set of automated agents who interact repeatedly in network contexts? (ii) Is the asymptotic play of network games characterized by traditional game-theoretic solution concepts such as Nash equilibrium? These questions are researched empirically, by simulating a set of sample responsive learning algorithms, and observing the strategies that are played in the long-run. The findings reported in this thesis suggest that the asymptotic play of network games is rather different from that of standard game-theoretic contexts. In particular, Nash equilibrium is not generally the outcome of responsive learning in asynchronous settings of limited information.

0.2.3 Information Economics

Computational game theory is useful not only in the study of typical problems that arise in network optimization and control, but moreover, it is applicable to economic interactions which transpire via networks, such as those that comprise the burgeoning world of e-commerce. In addition to applying economic principles to network design, this thesis also concerns the reverse migration of ideas, namely the implementation of agent technology in the domain of information economics.

Shopbots, agents that search the Internet for advertised goods and services on behalf of consumers, herald a future in which autonomous agents will be an essential component of nearly every facet of e-commerce. Moreover, shopbots deliver on one of the great promises of e-commerce: radical reductions in the costs of obtaining and distributing information. In the framework of computational game theory, this thesis proposes and analyzes an economic model of shopbots, and simulates an electronic marketplace inhabited by shopbots and *pricebots*, the latter being automated, price-setting agents that seek to maximize profits for sellers, just as shopbots seek to minimize costs for buyers. Analysis reveals that like the Santa Fe bar problem, rational learning in shopbot economics leads to instabilities that manifest themselves as price wars among pricebots. In contrast, low-rationality learning yields behaviors ranging from tacit collusion to exponential cycling, with only sophisticated learning algorithms converging to mixed strategy Nash equilibria.

0.3 Thesis Overview

Following this introductory chapter, this thesis continues with an introduction to the basic formalisms of game theory, providing a mathematical description of normal form games and the canonical game-theoretic solution concept, namely Nash equilibrium. The discussion proceeds with the definitions of strategies that are rationalizable, dominated, and overwhelmed, and the corresponding solution concepts that arise via iterative elimination processes. The notion of information games is then introduced, in order to make explicit the knowledge and beliefs of players, and to define two further solution concepts, namely correlated and Bayesian Nash equilibria. Lastly, the relationships between the various of notions of equilibria that are presented are described in the framework of information games.

Chapter 2 begins by reviewing a suite of optimality criteria which characterize degrees of low-rationality. In this chapter, optimality is described in terms of the fundamental property known as *no regret*. Intuitively, a sequence of plays is optimal if a player feels no regret for playing the given strategy sequence rather than playing any other possible sequence of strategies. The types of optimality which are described herein, listed in order from weakest to strongest, are as follows: no model-based regret, no external regret, no internal regret, and no clairvoyant regret. Note that the material pertaining to optimality in this chapter, while based on existing ideas in the literature on machine learning, statistics, and stochastic control, is reinterpreted and presented from a unique perspective in a unified game-theoretic framework.

The second half of Chapter 2 presents sample learning algorithms which satisfy the various optimality criteria mentioned above and the responsive learning criterion defined by Shenker and Friedman [43]. The responsive learning algorithms that are presented include responsive learning automata [42] and an algorithm that updates additively [30]; the mixing method [36] and a multiplicative updating algorithm [39] satisfy no external regret; no internal regret holds for two algorithms that converge to correlated equilibrium [37, 57]. Some of these algorithms were initially proposed for settings quite different than network contexts, where responsiveness is not of interest and the information level is significantly higher; these algorithms are redesigned here for use in network contexts. There are no known clairvoyant algorithms.

Chapter 3 presents a formalization of the intuitive argument given by Arthur that rationality precludes learning in the Santa Fe bar problem. In particular, it is argued that rational learning (*e.g.*, Cournot best-reply dynamics and Bayesian updating) yields unstable behavior in SFBP because rationality and predictivity, two conditions sufficient for convergence to Nash equilibrium, are inherently incompatible. This result has important implications for the networking games community, where it is often assumed both that players are rational and that the network operating point is a Nash equilibrium. On the other hand, it is observed via simulations, that low-rationality, no regret learning, which is non-Bayesian in nature, typically converges to mixed strategy Nash equilibria in SFBP.

Chapter 4 describes simulations of responsive learning agents in repeated games in network contexts, which are characterized by limited information and asynchronous play. Several key questions are addressed, including: (i) What are the appropriate game-theoretic solution concepts in network contexts? (ii) Is the asymptotic play of network games characterized by traditional game-theoretic solution concepts such as Nash equilibrium? (iii) To what extent does the asymptotic play depends on three factors, namely, asynchronous play, limited available information, and the degree of responsiveness of learning. The main conclusion of this chapter is that responsive learning in network contexts does not in general give rise to traditional game-theoretic solution concepts, such as Nash equilibrium; moreover, this result depends on the interplay of all three factors considered simultaneously.

The following chapter is concerned with the reverse migration of ideas, specifically the application of computational learning theory in the economic world of electronic commerce. As shopbots are increasingly being viewed as an essential component of information economics, this chapter proposes of a model of shopbot economics which is intended to capture some of the essence of shopbots, and attempts to shed light on their potential impact on the electronic marketplace. Analysis of this model yields a negative result on the existence of pure strategy Nash equilibria, which creates rational learning instabilities similar to those observed in SFBP. No internal regret learning algorithms, however, converge to mixed strategy Nash equilibria in shopbot economics; weaker no regret learning algorithms lead to exponential cycles reminiscent of the behavior of fictitious play in the Shapley game.

List of Figures

List of Tables

Chapter 1

Equilibrium Concepts

As a subfield of economics, game theory provides a framework in which to model the interaction of intelligent agents, or *players*, with different and often conflicting interests, who make decisions among possible *strategies*, while aiming to maximize individual *payoffs*. In contrast to the extreme cases of monopoly, where a single entity usurps all market power, and perfect competition, where it is assumed that individual actions have negligible effects on the market, *the payoffs of a game are jointly determined by the strategies of all players.*

In accordance with economics, a fundamental assumption of game theory is that players are *rational*. Rationality implies that players act so as to maximize their payoffs. Presently, we develop the classical theory of non-cooperative, non-zero-sum games in settings of complete information based on the assumption of rationality. In non-cooperative games, there is a possibility of negotiation prior to play, but there are no coalitions or binding agreements. In non-zero-sum games, there is in general a blend of contentious and cooperative behavior – this is in contrast to zero-sum games, in which the interests of players are diametrically opposed. In games of complete information, all information relevant to the game is available. While the canonical game-theoretic solution concept under these strict conditions is Nash equilibrium, this chapter emphasizes alternative forms of equilibria which generalize that of Nash and are later shown to arise as a result of various learning processes in repeated games.

1.1 One-Shot Games

The most well-known game-theoretic scenario is the paradoxical situation known as the *Prisoners' Dilemma*, which was popularized by Axelrod [7] in his popular science book. The following is one (uncommon) variant of the story.[1]

A crime has been committed for which two prisoners are held incommunicado. The district attorney questions the prisoners. If both prisoners confess, they will be punished, but not terribly severely, as the D.A. will reward them for their honesty. (Associate payoff 4 with this outcome.) If only one prisoner confesses, the confessor will be severely punished for carrying out the crime singlehandedly (payoff 0), while the other prisoner will be let off scot free (payoff 5). Lastly, if neither prisoner confesses, the D.A. threatens to convict both prisoners, although under a slightly less severe sentence than a sole confessor receives (payoff 1).

The Prisoners' Dilemma is a two player, strategic (or normal) form game. Such games are easily described by payoff matrices, where the strategies of player 1 and player 2 serve as row and column labels, respectively, and the corresponding payoffs are listed as pairs in matrix cells such that the first (second) number is the payoff to player 1 (2). The payoff matrix which describes the Prisoners' Dilemma is depicted in Figure 1.1, with C denoting "cooperate" or "confess", and D denoting "defect" or "don't cooperate."

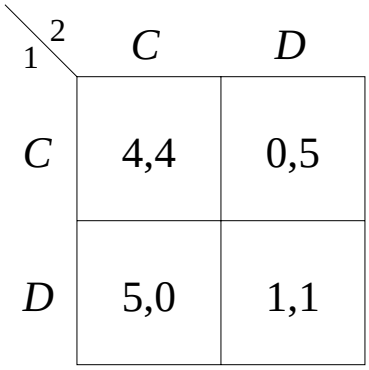


Figure 1.1: The Prisoners' Dilemma

[1] The original anecdote due to A.W. Tucker appears in Rapoport [87]; the latter author is the two-time winner of the Prisoners' Dilemma computer tournament organized by Axelrod.

This game is known as the Prisoners' Dilemma because the rational outcome is (D, D), which yields suboptimal payoffs of $(1, 1)$. The reasoning is as follows. If player 1 plays C, then player 2 is better off playing D, since D yields a payoff of 5, whereas C yields only 4; but if player 1 plays D, then player 2 is again better off playing D, since D yields a payoff of 1, whereas C yields 0. Hence, regardless of the strategy of player 1, a rational player 2 plays D. By a symmetric argument, a rational player 1 also plays D. Thus, the outcome of the game is (D, D).

A second well-known example of a two-player game is a game called *Matching Pennies*. In this game, each of the two players flips a coin, and the payoffs are determined as follows (see Figure 1.2). Let player 1 be the *matcher*, and let player 2 be the *mismatcher*. If the coins come up matching (*i.e.*, both heads or both tails), then player 2 pays player 1 the sum of \$1. If the coins do not match (*i.e.*, one head and one tail), then player 1 pays player 2 the sum of \$1. This is an example of a zero-sum game where the interests of the players are diametrically opposed; this class of games is so-called because the payoffs in the matrix sum to zero.

1 \ 2	*H*	*T*
H	1,-1	-1,1
T	-1,1	1,-1

Figure 1.2: Matching Pennies

Another popular two-player game is called the *Battle of the Sexes*. A man and a woman would like to spend an evening out together; however, the man prefers to go to a football game (strategy F), while the woman prefers to go to the ballet (strategy B). Both the man and the woman prefer to be together, even at the event that is not to their liking, rather than go out alone. The payoffs of this coordination game are shown in Figure 1.3; the woman is player 1 and the man is player 2.

1 \ 2	B	F
B	2,1	0,0
F	0,0	1,2

Figure 1.3: Battle of the Sexes

The final example of a game that is discussed herein is an ecological game which was studied by Maynard Smith [103] in his analysis of the theory of evolution in terms of games (see Figure 1.4). The game is played between animals of the same size who live in the wilderness and encounter one another in their search for prey. During an encounter between two animals, each animal has a choice between behaving as a hawk: *i.e.*, fighting for the prey; or as a dove: *i.e.*, sharing the prey peacefully. If both animals decide to play like hawks, then each animal has an equal chance of winning the value v of the prey or of losing the fight at cost c, where $0 < v < c$; thus, the expected payoff to both players is $(v - c)/2$. Alternatively, if both animals act as doves, then the prey is shared with equal payoffs $v/2$. Finally, if one animal behaves like a hawk and the other behaves like a dove, then the hawk gets a payoff worth the full value of the prey and the other gets nothing. In this game, the animals prefer to choose opposing strategies: if one animal plays hawk (dove), then it is in the best interest of the other to play dove (hawk), and inversely, if one animal plays dove, then it is in the best interest of the other to play hawk.

This section described several popular examples of one-shot – in contrast with repeated – strategic form games and their probable outcomes, or Nash equilibria. While Nash equilibrium is generally accepted as the appropriate solution concept in the deductive analysis of strategic form games, the Nash equilibria in the stated examples are somewhat peculiar. In particular, in the Prisoners' Dilemma, the Nash equilibrium payoffs are sub-optimal. Moreover, in the game of Matching Pennies,

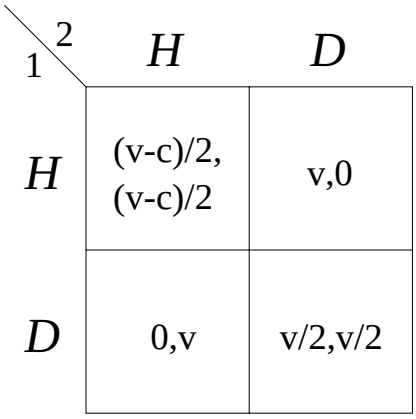


Figure 1.4: Hawks and Doves

there is no pure strategy Nash equilibrium; the unique Nash equilibrium is probabilistic. Finally, in the Battle of the Sexes, and in the game of Hawks and Doves, the Nash equilibrium is not unique. In view of these quirky outcomes, this thesis considers alternative forms of equilibria which arise as a result of learning in the repeated play of strategic form games.

1.2 Nash Equilibrium

This section develops the formal theory of finite games in strategic (or normal) form. Let $\mathcal{I} = \{1, \ldots, I\}$ be a set of *players*, where $I \in \mathbb{N}$ is the number of players. The (finite) set of *pure strategies* available to player $i \in \mathcal{I}$ is denoted by S_i, and the set of pure strategy profiles is the Cartesian product $S = \prod_i S_i$. By convention, write $s_i \in S_i$ and $s = (s_1, \ldots, s_I) \in S$. In addition, let $S_{-i} = \prod_{j \neq i} S_j$ with element $s_{-i} \in S_{-i}$, and write $s = (s_i, s_{-i}) \in S$. The payoff (or reward) function $r_i : S \to \mathbb{R}$ for the i^{th} player is a real-valued function on S; in this way, the payoffs to player i depend on the strategic choices of all players. This description is summarized in the following definition.

Definition 1.2.1 A *strategic form game* Γ is a tuple

$$\Gamma = (\mathcal{I}, (S_i, r_i)_{i \in \mathcal{I}})$$

where

- $\mathcal{I} = \{1, \ldots, I\}$ is a finite set of players ($i \in \mathcal{I}$)
- S_i is a finite strategy set ($s_i \in S_i$)
- $r_i : S \to \mathbb{R}$ is a payoff function

Example 1.2.2 The Prisoners' Dilemma consists of a set of players $\mathcal{I} = \{1, 2\}$, with strategy sets $S_1 = S_2 = \{C, D\}$, and payoffs as follows:

$$r_1(C,C) = r_2(C,C) = 4 \qquad r_1(C,D) = r_2(D,C) = 0$$
$$r_1(D,D) = r_2(D,D) = 1 \qquad r_1(D,C) = r_2(C,D) = 5 \qquad \square$$

Nash Equilibrium

A Nash equilibrium is a strategy profile from which none of the players has any incentive to deviate. In particular, no player can achieve strictly greater payoffs by choosing any strategy other than the one prescribed by the profile, given that all other players choose their prescribed strategies. In this sense, a Nash equilibrium specifies optimal strategic choices for all players.

In the Prisoners' Dilemma, (D, D) is a Nash equilibrium: given that player 1 plays D, the best response of player 2 is to play D; given that player 2 plays D, the best response of player 1 is to play D. The Battle of the Sexes has two pure strategy Nash equilibria, namely (B, B), and (F, F), by the following reasoning. If the woman plays B, then the best response of the man is B; if the man plays B, then the best response of the woman is B. Analogously, if the woman plays F, the best response of the man is F; if the man plays F, the best response of the woman is F.

In the game of Matching Pennies, there is no pure strategy Nash equilibrium. If player 1 plays H, then the best response of player 2 is T; but if player 2 plays T, the best response of player 1 is not H, but T. Moreover, if player 1 plays T, then the best response of player 2 is H; but if player 2 plays H, then the best response of player 1 is not T, but H. This game, however, does have a mixed strategy Nash equilibrium. A mixed strategy is a randomization over a set of pure strategies. In particular, the probabilistic strategy profile in which both players choose H with probability $\frac{1}{2}$ and T with probability $\frac{1}{2}$ is a mixed strategy Nash equilibrium.

Formally, a mixed strategy set for player i is the set of probability distributions over the pure strategy set S_i, which is computed via the simplex operator Δ: *i.e.*,

$$\Delta(S_i) = \{q_i : S_i \to [0,1] \mid \sum_{s_i \in S_i} q_i(s_i) = 1\}$$

For convenience, let $Q_i \equiv \Delta(S_i)$. The usual notational conventions extend to mixed strategies: *e.g.*, $Q = \prod_i Q_i$ and $q = (q_i, q_{-i}) \in Q$. In the context of mixed strategies, the expected payoffs to player i from strategy profile q is given by:

$$\mathbb{E}[r_i(q)] = \sum_{s \in S} r_i(s) \prod_{j=1}^{I} q_j(s_j)$$

As usual, the payoffs to player i depend on the mixed strategies of all players.

An implication of the assumption of rationality is that a rational player always plays an optimizing strategy, or a *best response* to the strategies of the other players.

Definition 1.2.3 A strategy $q_i^* \in Q_i$ is a *best response* for player i to opposing strategy $q_{-i} \in Q_{-i}$ iff $\forall q_i \in Q_i$,

$$\mathbb{E}[r_i(q_i^*, q_{-i})] \geq \mathbb{E}[r_i(q_i, q_{-i})]$$

Definition 1.2.4 The best response set for player i to strategy profile q_{-i} is:

$$\text{BR}_i(q_{-i}) = \{q_i^* \in Q_i \mid \forall q_i \in Q_i,\ \mathbb{E}[r_i(q_i^*, q_{-i})] \geq \mathbb{E}[r_i(q_i, q_{-i})]\}$$

The set $\text{BR}_i(q_{-i})$ is often abbreviated $\text{BR}_i(q)$. Let $\text{BR}(q) = \prod_i \text{BR}_i(q)$.

A Nash equilibrium is a strategy profile in which all players choose strategies that are best responses to the strategic choices of the other players. Nash equilibrium is characterized in terms of best response sets.

Definition 1.2.5 A Nash equilibrium is a strategy profile q^* *s.t.* $q^* \in \text{BR}(q^*)$.

It is apparent from this definition that a Nash equilibrium is a fixed point of the best response relation. The proof of existence of Nash equilibrium utilizes a fundamental result in topology, namely Kakutani's fixed point theorem[2] [64], which is a generalization of Brouwer's fixed point theorem [18].

[2] Every continuous correspondence on a non-empty, convex, bounded, and closed subset of a finite-dimensional Euclidean space into itself has a fixed point.

Theorem 1.2.6 (Nash, 1951) *All finite strategic form games have mixed strategy Nash Equilibria.*

This section formally defined the canonical solution concept for strategic form games, namely Nash equilibrium. The following sections describe equilibria which generalize that of Nash. The first class of such equilibrium concepts arise via the iterative deletion of sub-optimal strategies, where various definitions of sub-optimality yield various equilibria. Initially, the case of strictly pure strategy equilibria is studied. Later, the framework is extended to so-called information games, which are strategic form games equipped with probability spaces (*i.e.*, knowledge and/or belief systems) that offer one possible justification for mixed strategies. In this latter framework, the notions of correlated and Bayesian-Nash equilibria are defined.

1.3 Iterative Solutions

The solution concepts described in this section arise as fixed points of monotonic operators. Consider an elimination operator $E_i : 2^S \to 2^{S_i}$, which is defined for $i \in \mathcal{I}$. Let $E(T) = \prod_{i \in \mathcal{I}} E_i(T)$, for $T \subseteq S$. Now define $E^m(T)$ inductively as follows:

$$\begin{aligned} E^0(T) &= T \\ E^{m+1}(T) &= E(E^m(T)) \end{aligned}$$

If the elimination operator E is monotonic with respect to set inclusion, then the (greatest) fixed point $E^\infty(T) = \cap_{m=0}^{\infty} E^m(T)$ exists. The iterative solution concepts defined below, namely R^∞, D^∞, and O^∞, arise according to specific choices of the monotonic elimination operator E.

In classical game theory, two iterative solution concepts prevail, namely the set of rationalizable strategies (R^∞) and the set of undominated strategies (D^∞). In what follows, the intuition and the theory underlying these ideas is developed as a prerequisite for understanding the non-traditional solution concept known as O^∞. This section restricts attention to pure strategies; the following section extends the discussion to mixed strategy equilibria in the context of information games.

R^∞

The notion of rationalizable strategies was introduced independently by Pearce [85] and Bernheim [11]. In general, a (possibly mixed) strategy is rationalizable if it is a best response to some choice of (possibly mixed and possibly correlated) strategies by the other players. In this section, attention is restricted to strategies which are themselves pure, and moreover, rationalizable, assuming opposing strategies are pure as well. In other words, a pure strategy is rationalizable if it is a best response, as compared to all possible mixed strategies, to some choice of pure strategies by the opponents. The set R^∞ is the fixed point that arises as a result of the iterative deletion of strategies that are not rationalizable.

Definition 1.3.1 Pure strategy $s_i \in S_i$ is **not** a *rationalizable* strategy for player i iff $\forall s_{-i} \in S_{-i}$, $\exists q_i^* \in Q_i$ *s.t.* $\mathbb{E}[r_i(q_i^*, s_{-i})] > r_i(s_i, s_{-i})$.

1 \ 2	*L*	*M*	*R*
T	3	1 (2)	0
B	0	1 (2)	3

Figure 1.5: Rationalizable Strategies

Example 1.3.2 The game in Figure 1.5 is abbreviated; it depicts only the payoffs for player 2. In fact, this figure depicts two separate games, one in which strategy M yields payoffs of 1, and a second in which strategy M generates payoffs of 2. In this example, strategy M is *not* a rationalizable strategy for player 2, when M yields payoffs of 1, because there does not exist a strategy for player 1 to which M is a best response. In particular, the best response to strategy T is L, while the best response to strategy B is R.

In the case where the payoffs achieved by strategy M are equal to 2, however, M is in fact a best response to the mixed strategy $(\frac{1}{2}, \frac{1}{2})$ of player 2. In general, the set of pure strategy best responses to opposing mixed strategies is larger than the set of pure strategy best responses to opposing pure strategies. A notion of rationalizability which allows for opposing mixed, as well as pure, strategies is developed in the next section. □

Presently, it is shown that the set of pure rationalizable strategies which arises via comparison only with pure strategies is equivalent to that which arises via comparison with both pure and mixed strategies. The following lemma restates the definition of rationalizable in an equivalent form.

Lemma 1.3.3 *Pure strategy $s_i \in S_i$ is* **not** *a* rationalizable *strategy for player i iff* $\forall s_{-i} \in S_{-i}$, $\exists s_i^* \in S_i$, s.t. $r_i(s_i^*, s_{-i}) > r_i(s_i, s_{-i})$.

Proof 1.3.4

$$
\begin{aligned}
& s_i \text{ is not rationalizable} \\
\text{iff} \quad & \forall s_{-i} \in S_{-i},\ \exists q_i^* \in Q_i,\ \mathbb{E}[r_i(q_i^*, s_{-i})] > r_i(s_i, s_{-i}) \\
\text{iff} \quad & \forall s_{-i} \in S_{-i},\ \exists q_i^* \in Q_i,\ \textstyle\sum_{s_i \in S_i} q_i^*(s_i) r_i(s_i, s_{-i}) > r_i(s_i, s_{-i}) \\
\Rightarrow \quad & \forall s_{-i} \in S_{-i},\ \exists s_i^* \in S_i,\ r_i(s_i^*, s_{-i}) > r_i(s_i, s_{-i})
\end{aligned}
$$

The converse holds trivially since every pure strategy s_i^* can be expressed as mixed strategy q_i^* with $q_i^*(s_i^*) = 1$. □

Corollary 1.3.5 *Pure strategy $s_i^* \in S_i$* **is** *a* rationalizable *strategy for player i iff* $\exists s_{-i} \in S_{-i}$ s.t. $\forall s_i \in S_i$, $r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})$.

The operator R_i eliminates the strategies that are not rationalizable for player i, and returns the set of rationalizable strategies. Given a set $T \subseteq S$,

$$R_i(T) = \{s_i^* \in S_i \mid \exists s_{-i} \in T_{-i}, \forall s_i \in T_i,\ r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})\}$$

where T_i denotes the projection of T onto S_i. As usual, $R(T) = \prod_{i \in \mathcal{I}} R_i(T)$.

Definition 1.3.6 The set of rationalizable strategies R^∞ is $E^\infty(S)$ with $E = R$.

Theorem 1.3.7 *Given a strategic form game Γ, the set of rationalizable strategies of the game contains the set of pure strategy Nash equilibria.*

Proof 1.3.8 Must show that the set of pure strategy Nash equilibria is contained in $R^m(S)$ for all $m \in \mathbb{N}$. If $m = 0$, then $R^m(S) = S$; since this is the set of all strategy profiles of the game, it surely contains the set of pure strategy Nash equilibria. Assume $m > 0$. If the set of pure strategy Nash equilibria is empty, then the theorem holds trivially. Otherwise, assume there exists pure strategy Nash equilibrium $s^* = (s_1^*, \ldots, s_I^*) \in R^m(S)$. All components s_i^* satisfy the following:

$$\begin{aligned} & \exists s_{-i}^* \in R_{-i}^m(S) \text{ s.t. } s_i^* \in \text{BR}_i(s_{-i}^*) \\ \text{iff} \quad & \exists s_{-i}^* \in R_{-i}^m(S) \text{ s.t. } \forall s_i \in S_i,\ r_i(s_i^*, s_{-i}^*) \geq r_i(s_i, s_{-i}^*) \\ \text{iff} \quad & s_i^* \in R_i^{m+1}(S) \end{aligned}$$

Thus, s^* is contained in $R^{m+1}(S)$, and the theorem holds. □

Remark 1.3.9 Given a strategic form game Γ, the set of rationalizable strategies need not equal the set of pure strategy Nash equilibria.

While the set of probability distributions over the rationalizable strategies does contain the set of Nash equilibria, these two notions do not coincide. In the game of matching pennies, for example, the set of rationalizable strategies is the entire space of strategies. If the matcher believes that the mismatcher is playing strategy H, then strategy H is rationalizable for the matcher. If the mismatcher believes that the matcher is playing strategy H – presumably because the mismatcher believes that the matcher believes that the mismatcher is playing strategy H – then the mismatcher rationalizes playing strategy T. By extending this line of reasoning, strategy T is rationalizable for the matcher, and strategy H is rationalizable for the mismatcher. Therefore, the set of rationalizable strategies includes all possible pure strategies. The unique Nash equilibrium, however, is the mixed strategy $(\frac{1}{2}H, \frac{1}{2}T)$ for both players.

This section discussed the iterative deletion of pure rationalizable strategies. Next, the iterative deletion of pure dominated strategies is studied. As it is often difficult to justify the use of mixed strategies, it is of interest to define these concepts in the realm of pure strategies. It is important to note, however, that while these two notions differ in terms of pure strategy equilibria, they agree on mixtures.

D^∞

This section describes the set of pure strategies that arises via the iterative deletion of dominated strategies. A strategy is dominated if there exists another strategic choice which yields greater payoffs against all possible opposing strategies, both pure and mixed. Unlike rationalizable strategies, the set of dominated strategies does not depend on whether opponents play pure or mixed strategies. The set D^∞ is the fixed point that arises as a result of the iterative deletion of dominated strategies.

Definition 1.3.10 Strategy $s_i \in S_i$ is a *dominated* strategy for player i iff $\exists s_i^* \in S_i$ *s.t.* $\forall q_{-i} \in Q_{-i}$, $\mathbb{E}[r_i(s_i^*, q_{-i})] > \mathbb{E}[r_i(s_i, q_{-i})]$.

Example 1.3.11 In the Prisoners' Dilemma, strategy C is dominated by strategy D for player 1, since regardless of which strategy player 2 employs, player 1 is better off choosing strategy D. In particular, if player 2 plays C, then D yields a payoff of 5 for player 1, whereas C yields only 4; but if player 2 plays D, then D yields a payoff of 1 for player 1, whereas C yields 0. By a symmetric argument, strategy C is also dominated by strategy D for player 2. Thus, by the iterative deletion of dominated strategies, $D^\infty = \{(D, D)\}$. □

The following lemma shows that it suffices to define dominated strategies with respect to pure opposing strategies.

Lemma 1.3.12 *Strategy $s_i \in S_i$ is a* dominated *strategy for player i iff* $\exists s_i^* \in S_i$ s.t. $\forall s_{-i} \in S_{-i}$, $r_i(s_i^*, s_{-i}) > r_i(s_i, s_{-i})$.

Proof 1.3.13

$$\begin{aligned}
& \exists s_i^* \in S_i,\ \forall s_{-i} \in S_{-i},\ r_i(s_i^*, s_{-i}) > r_i(s_i, s_{-i}) \\
\Rightarrow\ & \exists s_i^* \in S_i,\ \forall q_{-i} \in Q_{-i},\ \textstyle\sum_{s_{-i} \in S_{-i}} q_{-i}(s_{-i}) r_i(s_i^*, s_{-i}) > \sum_{s_{-i} \in S_{-i}} q_{-i}(s_{-i}) r_i(s_i, s_{-i}) \\
\text{iff}\ & \exists s_i^* \in S_i,\ \forall q_{-i} \in Q_{-i},\ \mathbb{E}[r_i(s_i^*, q_{-i})] > \mathbb{E}[r_i(s_i, q_{-i})] \\
\text{iff}\ & s_i \text{ is dominated}
\end{aligned}$$

The contrapositive holds trivially since every pure strategy s_{-i} can be expressed as mixed strategy q_{-i} with $q_{-i}(s_{-i}) = 1$. □

The operator D_i eliminates the dominated strategies, returning the set of strategies that are not dominated for player i.

$$D_i(T) = \{s_i^* \in S_i \mid \forall s_i \in T_i, \exists s_{-i} \in T_{-i},\ r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})\}$$

As usual, $D(T) = \prod_{i \in \mathcal{I}} D_i(T)$.

Definition 1.3.14 The serially undominated strategy set D^∞ is $E^\infty(S)$ with $E = D$.

Strategies which are dominated are not rationalizable. In particular, if there exists a strategy that performs better than a given strategy against all possible opposing strategies, then the given strategy is not a best response to any choice of opposing strategies. It follows that rationalizable strategies are not dominated.

Lemma 1.3.15 *Given $T \subseteq S$, $R_i(T) \subseteq D_i(T)$, for all $i \in \mathcal{I}$.*

Proof 1.3.16

$$\begin{aligned} & R_i(T) \\ = \ & \{s_i^* \in S_i \mid \exists s_{-i} \in T_{-i}, \forall s_i \in T_i,\ r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})\} \\ \subseteq \ & \{s_i^* \in S_i \mid \forall s_i \in T_i, \exists s_{-i} \in T_{-i},\ r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})\} \\ = \ & D_i(T) \end{aligned}$$

□

It follows immediately from the above lemma that $R(T) \subseteq D(T)$, for all $T \subseteq S$. The following theorem states that $R^\infty \subseteq D^\infty$.

Theorem 1.3.17 $R^\infty \subseteq D^\infty$.

Proof 1.3.18 Must show that

$$\forall k, \forall T \subseteq S,\ R^k(T) \subseteq D^k(T)$$

The proof is by induction on k. The base case is trivial: $R^0(T) \subseteq D^0(T)$, since $T \subseteq T$. Now assume $R^k(T) \subseteq D^k(T)$. It follows via monotonicity of the operators that $R(R^k(T)) \subseteq R(D^k(T))$. Now by the lemma, $R(D^k(T)) \subseteq D(D^k(T))$. Finally, $R^{k+1}(T) = R(R^k(T)) \subseteq R(D^k(T)) \subseteq D(D^k(T)) = D^{k+1}(T)$. □

Remark 1.3.19 There exist games for which $R^\infty \subset D^\infty$.

Consider once again Figure 1.5, where strategy M yields payoff 1 for player 1. In this example, strategy M is *not* a rationalizable strategy; thus, $R^\infty = \{T, B\}$. Strategy M is also *not* a dominated strategy, however, and therefore can not be deleted in the iterative deletion of dominated strategies; thus, $D^\infty = \{T, M, B\}$.

Remark 1.3.20 [Pearce, 1984] In two player, two strategy games, the notions of non-rationalizable and dominated strategies coincide.

O^∞

This section discusses a solution concept due to Shenker and Friedman [43] known as the set of unoverwhelmed strategies. Recall that the definition of dominated strategy compares the vector of payoffs from one strategy with the vectors of payoffs of the other strategies, *payoff by payoff*. In contrast, the definition of overwhelmed strategy compares the entire set of payoffs yielded by one strategy with the entire set of payoffs yielded by the others. As the natural counterpart of Lemma 1.3.12 holds in the case of overwhelmed strategies, this discussion proceeds in terms of pure opposing strategies. The set O^∞ is the fixed point that arises as a result of the iterative deletion of overwhelmed strategies.

Definition 1.3.21 For player $i \in \mathcal{I}$, the strategy $s_i \in S_i$ is an *overwhelmed* strategy iff $\exists s_i^* \in S_i \ s.t. \ \forall s_{-i}, t_{-i} \in S_{-i}, \ r_i(s_i^*, s_{-i}) > r_i(s_i, t_{-i})$.

Example 1.3.22 An example of overwhelmed strategies is presented in Figure 1.6; only player 1's payoffs are depicted. In this game, strategy T overwhelms strategy B, since the set of payoffs $\{4, 3\}$ is everywhere greater than the set of payoffs $\{2, 1\}$. □

The operator O_i eliminates the overwhelmed strategies, and returns the set of strategies that are not overwhelmed for player i.

$$O_i(T) = \{s_i^* \in S_i \mid \forall s_i \in T_i, \exists s_{-i}, t_{-i} \in T_{-i}, \ r_i(s_i^*, s_{-i}) > r_i(s_i, t_{-i})\}$$

As usual, $O(T) = \prod_{i \in \mathcal{I}} O_i(T)$.

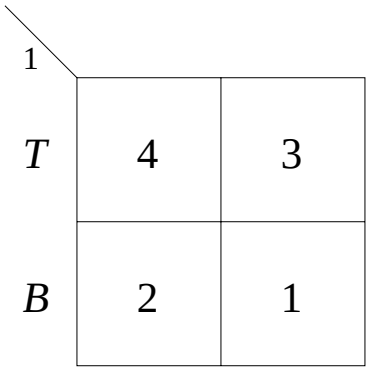


Figure 1.6: Unoverwhelmed Strategies

Definition 1.3.23 The set of serially unoverwhelmed strategies O^∞ is the set $E^\infty(S)$ with $E = O$.

Overwhelmed strategies are also dominated strategies, by letting $t_{-i} = s_{-i}$ in Definition 1.3.21. Thus, strategies which are not dominated are not overwhelmed.

Lemma 1.3.24 *Given $T \subseteq S$, $D_i(T) \subseteq O_i(T)$, for all $i \in \mathcal{I}$.*

Proof 1.3.25

$$\begin{aligned} & D_i(T) \\ = \; & \{s_i^* \in S_i \mid \forall s_i \in T_i,\, \exists s_{-i} \in T_{-i},\, r_i(s_i^*, s_{-i}) \geq r_i(s_i, s_{-i})\} \\ \subseteq \; & \{s_i^* \in S_i \mid \forall s_i \in T_i,\, \exists s_{-i}, t_{-i} \in T_{-i},\, r_i(s_i^*, s_{-i}) \geq r_i(s_i, t_{-i})\} \\ = \; & O_i(T) \quad \square \end{aligned}$$

Theorem 1.3.26 $D^\infty \subseteq O^\infty$.

Proof 1.3.27 The proof is analogous to that of Theorem 1.3.17. $\square$

Remark 1.3.28 There exist games for which $D^\infty \subset O^\infty$.

Consider the game depicted in Figure 1.7, which is a slight variant of that of Figure 1.6. In this example, T does not overwhelm B, but T dominates B. Thus, $D^\infty \subset O^\infty$.

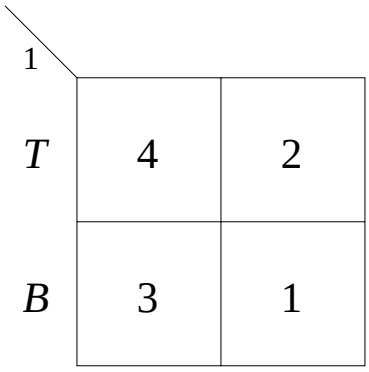


Figure 1.7: Undominated vs. Unoverwhelmed

1 \ 2	A_2	B_2	C_2	D_2	E_2
A_1	3,3	2,1	1,2	2,2	5,2
B_1	2,1	0,3	3,0	2,2	0,0
C_1	1,2	3,0	0,3	2,2	1,1
D_1	2,2	2,2	2,2	0,0	1,1
E_1	2,5	1,1	0,0	1,1	4,4

Figure 1.8: Differentiating Equilibria

Remark 1.3.29 There exist games for which $PNE \subset R^\infty \subset D^\infty \subset O^\infty$, where PNE is the set of pure strategy Nash equilibria.

Figure 1.8 depicts a game in which none of the solution concepts defined thus far coincide. First of all, no strategies are overwhelmed; thus, the set of unoverwhelmed strategies $O^\infty = \{A_1, B_1, C_1, D_1, E_1\} \times \{A_2, B_2, C_2, D_2, E_2\}$. The set of undominated strategies D_1^∞ (D_2^∞), however, does not contain strategy E_1 (E_2), since strategy A_1 dominates E_1 (similarly, strategy A_2 dominates E_2); thus, $D^\infty = \{A_1, B_1, C_1, D_1\} \times \{A_2, B_2, C_2, D_2\}$. The set of rationalizable strategies R_1^∞ (R_2^∞) in addition does not contain strategy D_1 (D_2), since D_1 (D_2) is not a best response to any choice of strategies for player 2 (1); thus, $R^\infty = \{A_1, B_1, C_1\} \times \{A_2, B_2, C_2\}$. Finally, the unique pure strategy Nash equilibrium in this game is given by (A_1, A_2). Therefore, $PNE \subset R^\infty \subset D^\infty \subset O^\infty$.

This concludes the discussion of iterative solution concepts in terms of solely pure strategies. The next section extends the formulation of strategic form games to structures referred to as information games, in attempt to provide justification for the notion of mixed strategies. In this framework, the connection between iterative solutions and other equilibrium concepts, including correlated and Bayesian-Nash equilibria, is explored.

1.4 Information Games

This section redevelops several of the equilibrium notions previously introduced in terms of information games. Information games are strategic form games equipped with an information structure in which to describe the knowledge and beliefs held by individual players. Such games provide a uniform framework in which to relate the heretofore unrelated solution concepts of correlated equilibrium and rationalizable strategies. In addition, an extension of information games known as Bayesian games is introduced in this section, where payoffs depend on an exogenously determined state of the world. Finally, Bayesian-Nash equilibrium and von Neumann-Morgenstern equilibrium, which arise in the context of Bayesian games, are defined.

1.4.1 Mixed Strategy Equilibria

An information game is a strategic form game in which players maintain a database of knowledge and beliefs about the possible outcomes of the game. This database is stored as a belief system, where beliefs are represented by probabilities, and knowledge is understood as belief with probability 1. The assumption of probabilistic beliefs leads to randomizations over the choice of pure strategies; hence, the notion of mixed strategies. In other words, mixed strategies arise as probability distributions over the possible states of the world, as is described by belief systems.

Definition 1.4.1 A *belief system* $\mathcal{B} = (\Omega, (\mathcal{P}_i, p_i)_{i \in \mathcal{I}})$ is a probability space, where

- Ω is a finite set of possible states of the world ($\omega \in \Omega$)
- $\mathcal{P}_i \subseteq 2^\Omega$ is an information partition ($P_i \in \mathcal{P}_i$)[3]
- $p_i : \Omega \to \mathbb{R}$ is a probability measure

An element of information partition $\mathcal{P}_i$ at state ω is called an information set for player i, and is denoted by $P_i(\omega)$. Intuitively, $P_i(\omega)$ is an equivalence class consisting of those states that player i cannot distinguish from ω. The function p_i induces a conditional probability on Ω which is measurable with respect to the knowledge described by information partition $\mathcal{P}_i$ *s.t.* $p_i(\omega|\omega_1) = p_i(\omega|\omega_2)$ whenever $\omega_1 \in P_i(\omega_2)$; formally, $p_i[A|\mathcal{P}_i](\omega_1) = p_i[A|\mathcal{P}_i](\omega_2)$, whenever $\omega_1 \in P_i(\omega_2)$, for all $A \subseteq \Omega$.

Example 1.4.2 Consider the state of knowledge today of two players about the price of IBM stock tomorrow. Assume the possible states of the world are up and down: *i.e.*, $\Omega = \{U, D\}$. If neither player knows the state of the world that will obtain tomorrow, then the information partition of each player is the trivial partition, namely $\{\Omega\}$. The players' beliefs, however, need not agree. For example, player 1 may attribute equal probabilities to both up and down: *i.e.*, $p_1(U) = p_1(D) = \frac{1}{2}$; while player 2 may attribute probability $\frac{2}{3}$ to up and $\frac{1}{3}$ to down: *i.e.*, $p_2(U) = \frac{2}{3}$ and $p_2(D) = \frac{1}{3}$. These probabilities induce conditional probabilities as follows:

$$p_1[\{U\}|\mathcal{P}_1](U) = p_1[\{U\}|\mathcal{P}_1](D) = p_1[\{D\}|\mathcal{P}_1](U) = p_1[\{D\}|\mathcal{P}_1](D) = \tfrac{1}{2}$$

$$p_2[\{U\}|\mathcal{P}_2](U) = p_2[\{U\}|\mathcal{P}_2](D) = \tfrac{2}{3} \quad \text{and} \quad p_2[\{D\}|\mathcal{P}_2](U) = p_2[\{D\}|\mathcal{P}_2](D) = \tfrac{1}{3} \ \square$$

[3] Technically, $\mathcal{P}_i$ is a σ-field.

Example 1.4.3 In the Battle of the Sexes viewed as an information game, the set of states of the world consists of all possible outcomes of the strategic form game: *i.e.*, $\Omega = \{(B,B),(B,F),(F,B),(F,F)\}$. The woman's knowledge of the world is given by information partition $\mathcal{P}_W = \{\{(B,B),(B,F)\},\{(F,B),(F,F)\}\}$; in other words, she is conscious of her own strategic play but is uncertain of the man's decisions. In contrast, the man's knowledge of the world is described by information partition $\mathcal{P}_M = \{\{(B,B),(F,B)\},\{(B,F),(F,F)\}\}$. Some sample probabilities are given by $p_W(B,B) = p_W(F,F) = \frac{1}{2}$ and $p_M(B,B) = p_M(B,F) = p_M(F,B) = p_M(F,F) = \frac{1}{4}$. These probabilities induce conditional probabilities for the woman, for example, as follows:

$$\begin{array}{rcl} p_W[\{(B,B),(F,B)\}|\mathcal{P}_W](B,B) & = & p_W[\{(B,B),(F,B)\}|\mathcal{P}_W](B,F) = 1 \\ p_W[\{(B,B),(F,B)\}|\mathcal{P}_W](F,B) & = & p_W[\{(B,B),(F,B)\}|\mathcal{P}_W](F,F) = 0 \\ p_W[\{(B,F),(F,F)\}|\mathcal{P}_W](B,B) & = & p_W[\{(B,F),(F,F)\}|\mathcal{P}_W](B,F) = 0 \\ p_W[\{(B,F),(F,F)\}|\mathcal{P}_W](F,B) & = & p_W[\{(B,F),(F,F)\}|\mathcal{P}_W](F,F) = 1 \end{array}$$

□

The above examples consider exogenously and endogenously determined possible states of the world, respectively. In particular, in Example 1.4.2, the state of the world is assumed to be independent of the players' decisions, while in Example 1.4.3, the state of the world is assumed to be fully determined by the players' strategic decisions. Until otherwise noted, this section proceeds from the point of view that the state of the world is determined endogenously.

Definition 1.4.4 An *information game* $\Gamma_{\mathcal{B}}$ is a strategic form game Γ together with a belief system $\mathcal{B}$ and, for all players $i \in \mathcal{I}$, an *adapted strategy* $A_i : \Omega \to S_i$: *i.e.*,

$$\Gamma_{\mathcal{B}} = (\mathcal{I}, \mathcal{B}, (S_i, r_i, A_i)_{i\in\mathcal{I}})$$

By definition, an adapted strategy A_i is measurable: *i.e.*, $A_i(\omega_1) = A_i(\omega_2)$ whenever whenever $\omega_1 \in P_i(\omega_2)$. This condition implies that identical strategies are played at indistinguishable states of the world: *i.e.*, strategic decisions depend only on the information partition $\mathcal{P}_i$. A vector of adapted strategies $(A_i)_{i\in I}$ is called an adapted strategy profile.

Rationality

In information games, the payoff functions are random variables, since the state of the world ultimately dictates payoffs. Given adapted strategy profile A, the expected payoffs for player i as predicted by player j are based on player j's beliefs, which are described by probability measure p_j:

$$\mathbb{E}_j[r_i(A)] = \sum_{\omega \in \Omega} p_j(\omega)\, r_i(A(\omega))$$

Definition 1.4.5 Given an information game $\Gamma_{\mathcal{B}}$, the adapted strategy A_i^* is *rational* for player i iff for all adapted strategies A_i,

$$\mathbb{E}_i[r_i(A_i^*, A_{-i})] \geq \mathbb{E}_i[r_i(A_i, A_{-i})]$$

An adapted strategy A_i^* is rational for player i which maximizes i's expectation of i's payoffs, given i's beliefs. A player is rational who plays rational adapted strategies.

Common Prior Assumption

The *common prior assumption* implies that people have different probabilities about the possibility of events occurring because people have access to different information. In other words, in the absence of differences in information, people ascribe the same probabilities to events. This is in contrast to utility functions, for example, where people have different utilities simply because they have different preferences (*e.g.*, some like coffee, but some like tea). The common prior assumption holds whenever $p_i = p_j$, for all $i, j \in \mathcal{I}$.

Correlated Equilibrium

This section introduces a generalization of Nash equilibrium due to Aumann [4] known as correlated equilibrium, which allows for possible dependencies in strategic choices. A daily example of a correlated equilibrium is a traffic light; red (green) traffic signal suggests that cars should stop (go), and in fact, these suggestions are best responses to the simultaneous suggestions for the strategies of others.

Definition 1.4.6 Given information game $\Gamma_{\mathcal{B}}$, a *correlated equilibrium* is an adapted strategy profile *s.t.*:

- All players are rational.
- The common prior assumption holds.

Example 1.4.7 Consider the Battle of the Sexes viewed as an information game (see Example 1.4.3). Assume the beliefs held by the woman are as given in the previous example, namely $p_W(B,B) = p_W(F,F) = \frac{1}{2}$, and similarly, for the man, $p_M(B,B) = p_M(F,F) = \frac{1}{2}$. The conditional probabilities for the woman are as given in Example 1.4.3; for the man, they are as follows:

$$\begin{array}{rclcrcl}
p_M[\{(B,B),(B,F)\}|\mathcal{P}_M](B,B) & = & p_M[\{(B,B),(B,F)\}|\mathcal{P}_M](F,B) & = & 1 \\
p_M[\{(B,B),(B,F)\}|\mathcal{P}_M](B,F) & = & p_M[\{(B,B),(B,F)\}|\mathcal{P}_M](F,F) & = & 0 \\
p_M[\{(F,B),(F,F)\}|\mathcal{P}_M](B,B) & = & p_M[\{(F,B),(F,F)\}|\mathcal{P}_M](F,B) & = & 0 \\
p_M[\{(F,B),(F,F)\}|\mathcal{P}_M](B,F) & = & p_M[\{(F,B),(F,F)\}|\mathcal{P}_M](F,F) & = & 1
\end{array}$$

The woman is rational if her adapted strategy prescribes that she is to play B on $\{(B,B),(B,F)\}$ and F on $\{(F,B),(F,F)\}$; similarly, the man is rational if his adapted strategy prescribes that he is to play B on $\{(B,B),(F,B)\}$ and F on $\{(B,F),(F,F)\}$. This is a correlated equilibrium in which the players abide by the joint probability distribution $\frac{1}{2}(B,B), \frac{1}{2}(F,F)$. □

Battle of the Sexes has three Nash equilibria, two of which are pure strategy equilibria, namely (B,B) and (F,F), and the mixed strategy $(\frac{2}{3},\frac{1}{3})$ for the woman and $(\frac{1}{3},\frac{2}{3})$ for the man, which yields equal expected payoffs of $(\frac{2}{3},\frac{2}{3})$ to both. The correlated equilibrium described in Example 1.4.7, however, yields expected payoffs of $(\frac{3}{2},\frac{3}{2})$. In general, it is possible to achieve correlated equilibrium payoffs as any convex combination of Nash equilibria. Moreover, it is also possible to achieve payoffs via correlated equilibrium outside the convex hull of Nash equilibrium payoffs. In the game depicted in Figure 1.9, for example, the Nash equilibrium achieved via mixed strategies $(\frac{1}{5},\frac{4}{5})$ and $(\frac{4}{5},\frac{1}{5})$ for players 1 and 2, respectively, yields expected payoffs of 4 for both players. In contrast, the correlated equilibrium strategies presented in Figure 1.9 generate expected payoffs of $4\frac{1}{4}$.

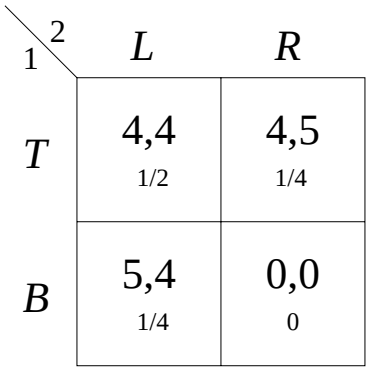


Figure 1.9: Correlated Equilibrium

Subjective Correlated Equilibrium

The notion of correlated equilibrium described in the previous section is sometimes referred to as objective correlated equilibrium, since, as noticed by Aumann [4], it is also possible to define a notion of subjective correlated equilibrium in which the common prior assumption does not hold.

Definition 1.4.8 Given information game $\Gamma_{\mathcal{B}}$, a *subjective correlated equilibrium* is an adapted strategy profile *s.t.*:

- All players are rational.

Theorem 1.4.9 *The set of objective correlated equilibria (*CE*) is contained within the set of subjective correlated equilibria (*SE*).*

Proof 1.4.10 The proof follows immediately from the definitions of objective and subjective correlated equilibria, since objective correlated equilibria form a special case of the class of subjective correlated equilibria. □

Example 1.4.11 Consider once again the Battle of the Sexes, as in Example 1.4.7, where the woman and the man play rational adapted strategies. Assume, however, that $p_W(B,B) = 1$ and $p_M(F,F) = 1$. The subjective correlated equilibrium outcome in this case is (B,F). □

Tying It All Together

This section brings together the ideas underlying the iterative solution concepts of the previous section with the notions of correlated equilibria. Initially, it is necessary to extend the iterative solution concepts to the case of mixed strategies. This discussion proceeds in terms of the set of serially undominated strategies, since in this case, it is possible to postpone the consideration of mixtures until after all pure strategies have been eliminated, without altering the final solution. Define the following:

$$D_i^{*\infty}(D^\infty) = \{q_i^* \in Q_i \mid \forall q_i \in \Delta(D_i^\infty), \exists s_{-i} \in D_i^\infty, \mathbb{E}[r_i(q_i^*, s_{-i}) \geq \mathbb{E}[r_i(q_i, s_{-i})]\}$$

As usual, $D^{*\infty}(T) = \prod_{i \in \mathcal{I}} D_i^{*\infty}(T)$. Let $D^{*\infty} \equiv D^{*\infty}(D^\infty)$. Of course, $D^\infty \subseteq D^{*\infty}$ since every pure strategy s_i can be described by mixed strategy $\sigma_i(s_i) = 1$.

It is also possible to define a mixed strategy notion of rationalizable strategies in which a mixed strategy is rationalizable if it is a best response to some choice of (possibly correlated) strategies by the other players. It is known, however, that the solutions $R^{*\infty} = D^{*\infty}$, as long as opponents' strategies are allowed to be correlated (see, for example, Fudenberg and Tirole [47]). In particular, in the space of probability distributions, the order of quantifiers is no longer of relevance. The proof of this fact is a variant of the minimax theorem, which in turn is usually proven using the separating hyperplane theorem. The following theorem relates the various generalizations of Nash equilibrium that have been defined thus far.

Theorem 1.4.12 CE $\subseteq D^{*\infty} = R^{*\infty} \subseteq$ SE.

Proof 1.4.13 The rationality constraint in the definition of correlated equilibrium ensures that no players will play strategies which are not rationalizable: CE $\subseteq R^{*\infty}$. The final relationship follows since $R^{*\infty} \subseteq R^{*1}$, which is equivalent to SE. □

1.4.2 Nash Equilibrium Revisited

Nash equilibrium is a special case of correlated equilibrium in which one player's beliefs about the state of the world is independent of any other's. Independent beliefs leads to independent randomizations over the choice of pure strategies.

Definition 1.4.14 An *independent belief system* $\mathcal{B} = (\Omega, (\mathcal{P}_i, p_i)_{i\in\mathcal{I}})$ is a special case of a belief system where $p_i(P_j \cap P_k) = p_i(P_j)p_i(P_k)$, for all $P_j \in \mathcal{P}_j$, $P_k \in \mathcal{P}_k$, $i, j, k \in \mathcal{I}$.

Nash Equilibrium

This section redefines Nash equilibrium in the context of information games.

Definition 1.4.15 Given an information game $\Gamma_{\mathcal{B}}$, a *Nash equilibrium* is an adapted strategy profile *s.t.*:

- All players are rational.
- The common prior assumption holds.
- $\mathcal{B}$ is an independent belief system.

Example 1.4.16 Consider once again the Battle of the Sexes, as in Example 1.4.7, where $p_W(B,B) = p_M(B,B) = \frac{1}{2}$ and $p_W(F,F) = p_M(F,F) = \frac{1}{2}$. Recall that this is a correlated equilibrium. This is not, however, a Nash equilibrium, as the independence property fails. In particular,

$$p_W[\{(B,B),(B,F)\}]p_W[\{(B,B),(F,B)\}] = \left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right) = \tfrac{1}{4} \neq \tfrac{1}{2} = p_W(B,B)$$
$$p_W[\{(F,B),(F,F)\}]\,p_W[\{(B,F),(F,F)\}] = \left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right) = \tfrac{1}{4} \neq \tfrac{1}{2} = p_W(F,F)$$

and similarly for the man. On the other hand, both $p_W(B,B) = p_M(B,B) = 1$ and $p_W(F,F) = p_M(F,F) = 1$ satisfy the independence property, and are therefore pure strategy Nash equilibria. Finally, the following probabilities form a mixed strategy Nash equilibrium:

$$p_W(B,B) = p_M(B,B) = \tfrac{2}{9} \qquad p_W(B,F) = p_M(B,F) = \tfrac{4}{9}$$
$$p_W(F,B) = p_M(F,B) = \tfrac{1}{9} \qquad p_W(F,F) = p_M(F,F) = \tfrac{2}{9}$$

since

$$p_W[\{(B,B),(B,F)\}]p_W[\{(B,B),(F,B)\}] = \left(\tfrac{2}{3}\right)\left(\tfrac{1}{3}\right) = \tfrac{2}{9} = p_W(B,B)$$
$$p_W[\{(B,B),(B,F)\}]p_W[\{(B,F),(F,F)\}] = \left(\tfrac{2}{3}\right)\left(\tfrac{2}{3}\right) = \tfrac{4}{9} = p_W(B,F)$$
$$p_W[\{(F,B),(F,F)\}]\,p_W[\{(B,B),(F,B)\}] = \left(\tfrac{1}{3}\right)\left(\tfrac{1}{3}\right) = \tfrac{1}{9} = p_W(F,B)$$
$$p_W[\{(F,B),(F,F)\}]\,p_W[\{(B,F),(F,F)\}] = \left(\tfrac{1}{3}\right)\left(\tfrac{2}{3}\right) = \tfrac{2}{9} = p_W(F,F)$$

and similarly for the man. □

Theorem 1.4.17 *The set of Nash equilibria is a subset of that of correlated equilibria.*

The notion of Nash equilibrium that is defined in this section for information games implies the usual definition of Nash equilibrium in (informationless) strategic form games. Given a Nash equilibrium for an information game, the Nash equilibrium of the corresponding strategic form game is derived as follows: player i plays strategy $A_i(\omega)$ with probability $p_i[P_i(\omega)]$, where $p_i[P_i(\omega)] = \sum_{\omega' \in P_i(\omega)} p_i(\omega')$, for all $\omega \in \Omega$. It follows immediately from the fact that all adapted strategies are rational at Nash equilibrium that the induced probabilities in the strategic form game constitute a Nash equilibrium as well.

Subjective Nash Equilibria

Just as it is of interest to define a subjective notion of correlated equilibrium, it is similarly possible to define a subjective notion of Nash equilibrium. This definition proceeds as follows:

Definition 1.4.18 Given an information game $\Gamma_{\mathcal{B}}$, a *subjective Nash equilibrium* is an adapted strategy profile *s.t.*:

- All players are rational.
- $\mathcal{B}$ is an independent belief system.

Theorem 1.4.19 *The set of Nash equilibria is a subset of the set of subjective Nash equilibria.*

1.4.3 Bayesian-Nash Equilibrium

In contrast with the games considered thus far which have not accounted for exogenous influences, this section considers games in which the state of the world is determined exogenously, so-called Bayesian games. The term Bayesian game is also sometimes used to refer to games of incomplete information in which there are still no exogenous effects, but the players are unaware of one another's payoffs. In this latter case, players condition on the other players' probable payoffs, just as in the Bayesian games considered here, where players condition on the probable state of the world.

Definition 1.4.20 A *Bayesian game* $\Gamma_{\mathcal{B}}(\omega)$ is an information game $\Gamma_{\mathcal{B}}$ in which the payoff functions depend on the state of the world: *i.e.*, $r_i : S \times \Omega \to \mathbb{R}$.

Example 1.4.21 A well-known example of a Bayesian game is the so-called envelope paradox. A father offers each of his two sons an envelope with either $\$10^m$ or $\$10^n$, where $|m - n| = 1$, for $0 \leq m, n \leq 6$. Each brother can accept his envelope, or chose to engage in a bet in which he pays the father \$1 for the right to swap envelopes with his brother, provided that his brother has also chosen to engage in this bet. Otherwise, he simply loses \$1. □

1 \ 2	B	NB
B	10^n-1, 10^m-1	10^m-1, 10^n
NB	10^m, 10^n-1	10^m, 10^n

Figure 1.10: The Envelope Paradox

The envelope paradox can be modeled as an information game in which the two brothers have private knowledge regarding the payoffs, but the state of the world is unknown. In particular, let (m, n) denote the state of the world, where m (n) is the exponent of the payoff to the first (second) brother. The pure strategy sets of the brothers are BET and NO BET. The payoff matrix in Figure 1.10 depicts the outcomes of the envelope game in terms of the unknown state of the world. The envelope paradox is so-called because so long as $m, n \neq 6$, it is in the best interest of both brothers to accept the bet. Since the probablity that the second brother receives 10 times as much money as the first brother is $\frac{1}{2}$, the expected value of the second brother's lot, given the information that the first brother has about his own lot, is approximately 5 times greater than the first brother's lot. Thus, it is in the first brother's best interest to bet. The reasoning is analogous for the second brother. This paradox is resolved via the concept of Bayesian-Nash equilibrium.

Bayesian-Nash Equilibrium

Bayesian-Nash equilibrium is an extension of correlated equilibrium to the class of Bayesian games.[4]

Definition 1.4.22 Given Bayesian game $\Gamma_{\mathcal{B}}(\omega)$, a *Bayesian-Nash equilibrium* is an adapted strategy profile *s.t.*:

- All players are rational.
- The common prior assumption holds.

The strategy profile in which both players do not bet is the unique Bayesian-Nash equilibrium in the Bayesian information game which describes the envelope paradox. Suppose the state of the world is $(2,3)$ and moreover, assume the brothers have common prior beliefs about the state of the world *s.t.* $p_1(2,3) = p_2(2,3) = 1$. Given this belief structure, the second brother surely should not bet, since he has already received the greater lot. In addition, given that the second brother should not bet, the first brother also should not bet, since he would incur a loss of \$1 otherwise. These strategic choices form the unique Bayesian-Nash equilibrium. In essence, the envelope paradox is resolved whenever the common prior assumption is satisfied; since the only sets in the information partitions of the brothers which are assigned positive probabilities are, respectively, $\{(m-1,m),(m,m+1)\}$ and $\{(m,m+1),(m+1,m+2)\}$, the common prior assumption is only satisfied when $p_1(m,m+1) = p_2(m,m+1) = 1$.

Summary and Conclusions

This concludes the discussion of information games. Information games generalize strategic form games via explicit use of information structures. While it is by all means possible to define equilibria solely in terms of strategic form games, the framework of information games clarifies the relationships among the numerous solution concepts, and moreover, information games motivate the play of mixed strategies.

[4] Nash equilibrium was defined in the 50's [83], Bayesian-Nash equilibrium was defined in the 60's [56], but correlated equilibrium was not defined until the 70's [4]. Hence, although Bayesian-Nash equilibrium is in effect correlated, it is not so-called, since upon its invention, correlated equilibrium did not exist.

1.5 Discussion: Repeated Games

This chapter presented an overview of one-shot games and several relevant equilibrium concepts. In order to determine the domain of applicability of the various solutions, this thesis is concerned with the dynamics of learning over repeated instances of one-shot games. Repeated games form a special case of the class of infinite-horizon, extensive form games, or game trees. Unlike finite game trees, however, where there exists solutions via bottom-up algorithms like dynamic programming and minimax, in the case of infinitely repeated games, this would lead to an infinite regress. Instead, the learning dynamics studied in this thesis fall into the category of reinforcement learning algorithms, which experiment among possible strategies and as the name suggests, reinforce those that yield relatively high payoffs. The following chapter discusses a suite of optimality criteria and corresponding learning algorithms for which repeated play converges to various generalizations of Nash equilibrium.

Chapter 2

Optimality and Learning

The ultimate goal of learning in repeated games is to induce optimal behavior. A learning algorithm which achieves this objective is said to be *clairvoyant*. More specifically, a clairvoyant algorithm generates a sequence of plays that consists only of best responses, where a strategy is called a best response if the payoffs achieved by that strategy are at least as great as the payoffs achieved by any other strategy. It follows that clairvoyant learning algorithms converge to Nash equilibrium play. Unfortunately, no clairvoyant algorithms are known to exist. As a result, this thesis investigates relaxed notions of optimality for which there do in fact exist provably effective learning algorithms for repeated games. Since the optimality concepts that are studied herein are weaker than clairvoyance, they give rise to learning algorithms which converge to solution concepts that generalize that of Nash equilibrium.

The first half of this chapter presents a suite of optimality criteria which appear in the literature on game theory, machine learning, and stochastic control. In this thesis, optimality is described in terms of *no regret*. Intuitively, a sequence of plays is optimal if there is no regret for playing the given strategy sequence rather than playing any other possible sequence of strategies. The types of optimality which are described herein, listed in order from weakest to strongest, are as follows: no model-based regret, no external regret, no internal regret, and no clairvoyant regret. The second half of this chapter presents examples of learning algorithms which satisfy the aforementioned optimality criteria.

2.1 Optimality Criteria

Given a sequence of T decisions, clairvoyant algorithms generate best response play at all times. It follows that the expected cumulative payoffs achieved via a clairvoyant algorithm after T decisions are at least as great as those that could be achieved by any other possible sequences of actions. When this rigid criterion is relaxed, it gives rise to optimality criteria such as no model-based regret, no external regret, and no internal regret.

No model-based regret is an optimality criterion suggested by stochastic control theorists, such as Narendra and Thathachar [82]. This type of regret is so-called because it is dependent on the model of the environment that is under consideration, which is described by a probability distribution. An algorithm is said to exhibit *no model-based regret* if the difference between the expected cumulative payoffs that are achieved by the algorithm and those that could be achieved by any fixed alternative strategy is insignificant, with respect to *expectations* over the opposing sequence of plays as determined by the environment.

Whereas stochastic control theory considers expectations of the environment, in other words average-case performance, machine learning is focused on performance in the worst case. In particular, machine learning researchers consider the difference between the expected payoffs that are achieved by a given algorithm, as compared to the payoffs that could be achieved by any other fixed sequence of decisions, with respect to *actual* opposing strategy sequences. If the difference between sums is negligible, then the algorithm is said to exhibit *no external regret*. Early no external regret algorithms appeared in Blackwell [14], Hannan [53], and Banos [10].

Game theorists Foster and Vohra [37] consider an alternative measure of worst-case performance. If the difference between the cumulative payoffs that are achieved by a sequence of strategies generated by a given algorithm, in comparison with the cumulative payoffs that could be achieved by a remapped sequence of strategies, is insignificant, then the algorithm is said to exhibit *no internal regret*.[1] Note that no internal regret implies no external regret implies no model-based regret.

[1] A sequence is remapped if there is a mapping f of the strategy space into itself *s.t.* for each occurrence of strategy s_i in the original sequence, the mapped strategy $f(s_i)$ appears in the remapped sequence.

2.1.1 No Model-based Regret

An algorithm is said to exhibit no model-based regret if and only if the difference between the expected cumulative payoffs that are achieved by the algorithm and those that could be achieved by any fixed alternative strategy is insignificant, *with respect to expectations over the opposing sequence of plays*. In other words, no model-based regret is no regret in the average case, given a prespecified model of the environment. (This and other types of average-case behavior are analyzed in a survey paper by Narendra and Thathachar [82].[2]) This type of regret is so-called because it depends on the model of the opponents, who taken collectively act as the environment. As the model of the environment is taken as given, model-based regret is applicable in situations in which it is assumed that the strategic decisions taken by individuals do not impact the environment. This assumption is reminiscent of common assumptions in the economic theory of perfect competition.

Regret is a feeling of remorse over something that has happened, particularly as a result of one's own actions. In game-theoretic notation, the regret felt by player i whenever strategy s_i is played is formulated as the difference between the payoffs obtained by utilizing strategy s_i and the payoffs that could have been achieved had some other strategy, say $\bar{s}_i$, been played instead. In particular, if the probability distribution q^t_{-i} serves as a model of the environment at time t, the expected regret felt by player i at time t is the difference between the expected payoff of strategy $\bar{s}_i$ and strategy s_i:

$$\mathbb{E}[\mathrm{R}^t_{s_i \to \bar{s}_i}] \;=\; \mathbb{E}[r_i(\bar{s}_i, q^t_{-i}) - r_i(s_i, q^t_{-i})] \tag{2.1}$$

The cumulative expected regret through time T that is felt by player i from strategy s_i towards strategy $\bar{s}_i$ is the summation over the instantaneous values of expected regret, whenever strategy s_i is played rather than strategy $\bar{s}_i$:

$$\mathbb{E}[\mathrm{R}^T_{s_i \to \bar{s}_i}] \;=\; \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} \mathbb{E}[\mathrm{R}^t_{s_i \to \bar{s}_i}] \tag{2.2}$$

[2] Narendra and Thathachar refer to no model-based regret as ϵ-optimality, but this denomination is ambiguous in this chapter, which contains a review of optimality criteria. Consequently, this property has been renamed for the purposes of this exposition.

where $\mathbf{1}_{\{a=b\}}$ is the indicator function.[3] Finally, the model-based regret felt by player i towards strategy $\bar{s}_i$ is the summation over all strategies $s_i \in S_i$ of the strategic values of model-based regret:

$$\mathrm{MR}^T_{S_i \to \bar{s}_i} = \sum_{s_i \in S_i} \mathbb{E}[\mathrm{R}^T_{s_i \to \bar{s}_i}] \tag{2.3}$$

Definition 2.1.1 Given a sequence of plays $\{s^t\}$ of length T, the sequence of plays $\{s^t_i\}$ for player i is said to exhibit *no model-based regret* iff $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathrm{MR}^T_{S_i \to \bar{s}_i} < \epsilon T$$

In words, a sequence of plays exhibits no model-based regret if the difference between the cumulative payoffs that are achieved by the given sequence and those that could be achieved by any fixed alternative strategy is insignificant, with respect to expectations over the opposing sequence of plays. Thus, it suffices to compare the payoffs of the given sequence with the payoffs that could have been obtained by the best possible fixed strategy, for if the given sequence achieves payoffs that are comparable with the best alternative, then the given sequence achieves payoffs that are comparable with all the alternatives. The no model-based regret condition can be restated in terms of an optimal fixed strategy.

Lemma 2.1.2 *Given a model of the environment expressed as a sequence of weights* $\{q^t_{-i}\}$ *of length* T. *Consider a sequence of plays* $\{s^{*t}_i\}$ *constructed* s.t. $\forall t, s^{*t}_i = s^*_i$, *where* s^*_i *is optimal in the following sense:*

$$s^*_i = \arg\max_{\bar{s}_i \in S_i} \sum_{s_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s^t_i = s_i\}} \mathbb{E}[r_i(\bar{s}_i, q^t_{-i})]$$

The sequence of plays $\{s^t_i\}$ *for player* i *exhibits no model-based regret iff* $\forall \epsilon > 0$,

$$\sum_{t=1}^{T} \mathbb{E}[r_i(s^{*t}_i, q^t_{-i})] - \sum_{t=1}^{T} \mathbb{E}[r_i(s^t_i, q^t_{-i})] < \epsilon T$$

[3] The indicator function is defined as follows:

$$\mathbf{1}_{\{a=b\}} = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

For convenience, in the description of learning algorithms, $\mathbf{1}_{s^t_i = s_i}$ is written $\mathbf{1}^t_{s_i}$.

Proof 2.1.3 The lemma follows directly from the construction of the sequence of plays $\{s_i^{*t}\}$ and the definition of no model-based regret. For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i,\ \mathrm{MR}^T_{S_i \to \bar{s}_i} < \epsilon T \\
\text{iff}\quad & \forall \bar{s}_i \in S_i,\ \textstyle\sum_{s_i \in S_i} \sum_{t=1}^T \mathbf{1}_{\{s_i^t = s_i\}} \mathbb{E}[r_i(\bar{s}_i, q_{-i}^t) - r_i(s_i, q_{-i}^t)] < \epsilon T \\
\text{iff}\quad & \max_{\bar{s}_i \in S_i} \textstyle\sum_{s_i \in S_i} \sum_{t=1}^T \mathbf{1}_{\{s_i^t = s_i\}} \mathbb{E}[r_i(\bar{s}_i, q_{-i}^t) - r_i(s_i, q_{-i}^t)] < \epsilon T \\
\text{iff}\quad & \textstyle\sum_{s_i \in S_i} \sum_{t=1}^T \mathbf{1}_{\{s_i^t = s_i\}} \mathbb{E}[r_i(s_i^*, q_{-i}^t) - r_i(s_i, q_{-i}^t)] < \epsilon T \\
\text{iff}\quad & \textstyle\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[r_i(s_i^t, q_{-i}^t)] < \epsilon T
\end{aligned}$$

□

The optimality criteria discussed in this chapter are useful not only to describe the properties of sequences of plays, but in addition, they are applicable to sequences of weights which are often generated via learning algorithms. A sequence of weights exhibits no model-based regret if the difference between the cumulative payoffs that are achieved by the given sequence and those that could be achieved by any fixed alternative strategy is insignificant, with respect to expectations over the opposing sequence of weights generated by the environment.

Definition 2.1.4 Given opposing sequence of weights $\{q_{-i}^t\}$ of length T employed by the environment, an algorithm is said to exhibit *no model-based regret* iff it gives rise to a sequence of weights $\{q_i^t\}$ *s.t.* $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathbb{E}[\mathrm{MR}^T_{S_i \to \bar{s}_i}] \ < \ \epsilon T$$

Theorem 2.1.5 *Let $\{s_i^{*t}\}$ be as defined in Lemma 2.1.2. An algorithm that gives rise to a sequence of weights $\{q_i^t\}$ of length T exhibits no model-based regret against opposing sequence of weights $\{q_{-i}^t\}$ iff $\forall \epsilon > 0$,*

$$\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[r_i(q_i^t, q_{-i}^t)] < \epsilon T$$

Proof 2.1.6 The theorem follows via Lemma 2.1.2 and the definition of no model-based regret.

For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i, \mathbb{E}[\mathrm{MR}^T_{S_i \to \bar{s}_i}] < \epsilon T \\
\text{iff} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[r_i(s_i^t, q_{-i}^t)]\right] < \epsilon T \\
\text{iff} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[\textstyle\sum_{s_i \in S_i} \mathbf{1}_{\{s_i^t = s_i\}} r_i(s_i, q_{-i}^t)]\right] < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[\sum_{s_i \in S_i} q_i^t(s_i) r_i(s_i, q_{-i}^t)] < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[r_i(q_i^t, q_{-i}^t)] < \epsilon T
\end{aligned}$$

□

The property of no model-based regret is satisfied by a suite of additive updating algorithms, which are described in the second half of this chapter. The following sections discuss a series of refinements of the no model-based optimality criterion. While no model-based regret is an average-case performance measure, the remaining optimality criteria are measures of performance in the worst-case; more specifically, performance is measured over all possible opposing sequences of plays, rather than with respect to a specific model.

2.1.2 No External Regret

Recall that an algorithm exhibits no model-based regret if it yields no regret in the average case, with respect to a given model of the environment. The no external regret property is a strengthening of no model-based regret which measures performance in the worst case. In particular, no external regret is satisfied by algorithms that exhibit no regret with respect to all adversarial strategy sequences. Early no external regret algorithms were described in Blackwell [14], Hannan [53], Banos [10], and Megiddo [77]; Recent no regret algorithms appeared in Cover [26] and Auer, Cesa-Bianchi, Freund, and Schapire [3].

The regret felt by player i at time t is formulated as the difference between the payoffs obtained by utilizing strategy s_i and the payoffs that could have been achieved had strategy $\bar{s}_i$ been played instead, given the opposing strategy that is *actually* employed by the environment at time t:

$$\mathrm{R}^t_{s_i \to \bar{s}_i} \quad = \quad r_i(\bar{s}_i, s_{-i}^t) - r_i(s_i, s_{-i}^t) \tag{2.4}$$

Given sequence of plays $\{s^t\}$, the cumulative regret felt by player i from strategy s_i towards strategy $\bar{s}_i$ is computed as the summation over the length of the sequence of the instantaneous regret felt whenever strategy s_i is played rather than $\bar{s}_i$:

$$\mathrm{R}^T_{s_i \to \bar{s}_i} = \sum_{t=1}^{T} \mathbf{1}_{s_i^t = s_i} \mathrm{R}^t_{s_i \to \bar{s}_i} \tag{2.5}$$

Finally, the external regret felt by player i towards strategy $\bar{s}_i$ arising via the sequence $\{s^t\}$ is the summation over all strategies $s_i \in S_i$ of the individual strategic regrets:

$$\mathrm{ER}^T_{S_i \to \bar{s}_i} = \sum_{s_i \in S_i} \mathrm{R}^T_{s_i \to \bar{s}_i} \tag{2.6}$$

Definition 2.1.7 Given a sequence of plays $\{s^t\}$ of length T, the sequence of plays $\{s_i^t\}$ for player i is said to exhibit *no external regret* iff $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathrm{ER}^T_{S_i \to \bar{s}_i} < \epsilon T$$

Lemma 2.1.8 *Given a sequence of plays $\{s^t\}$ of length T. Consider a sequence of plays $\{s_i^{*t}\}$ constructed* s.t. $\forall t, s_i^{*t} = s_i^*$, *where s_i^* is optimal in the following sense:*

$$s_i^* = \arg\max_{\bar{s}_i \in S_i} \sum_{s_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} r_i(\bar{s}_i, s_{-i}^t)$$

The sequence of plays $\{s_i^t\}$ for player i exhibits no external regret iff $\forall \epsilon > 0$,

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} r_i(s_i^t, s_{-i}^t) < \epsilon T$$

Proof 2.1.9 The lemma follows directly from the construction of the sequence of plays $\{s_i^{*t}\}$ and the definition of no external regret. For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i,\ \mathrm{ER}^T_{S_i \to \bar{s}_i} < \epsilon T \\
\text{iff}\ & \forall \bar{s}_i \in S_i,\ \textstyle\sum_{s_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s_{-i}^t) - r_i(s_i, s_{-i}^t)] < \epsilon T \\
\text{iff}\ & \max_{\bar{s}_i \in S_i} \textstyle\sum_{s_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s_{-i}^t) - r_i(s_i, s_{-i}^t)] < \epsilon T \\
\text{iff}\ & \textstyle\sum_{s_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(s_i^*, s_{-i}^t) - r_i(s_i, s_{-i}^t)] < \epsilon T \\
\text{iff}\ & \textstyle\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} r_i(s_i^t, s_{-i}^t) < \epsilon T
\end{aligned}$$

□

Definition 2.1.10 Given any opposing sequence of plays $\{s_{-i}^t\}$ of length T that is employed by the environment, an algorithm is said to exhibit *no external regret* iff it gives rise to a sequence of weights $\{q_i^t\}$ *s.t.* $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathbb{E}[\text{ER}_{S_i \to \bar{s}_i}^T] \;<\; \epsilon T$$

Theorem 2.1.11 *Let $\{s_i^{*t}\}$ be as defined in Lemma 2.1.8. An algorithm that gives rise to a sequence of weights $\{q_i^t\}$ of length T exhibits no external regret against opposing sequence of plays $\{s_{-i}^t\}$ iff $\forall \epsilon > 0$,*

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} \mathbb{E}[r_i(q_i^t, s_{-i}^t)] < \epsilon T$$

Proof 2.1.12 The theorem follows via Lemma 2.1.8 and the definition of no external regret. For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i, \mathbb{E}[\text{ER}_{S_i \to \bar{s}_i}^T] < \epsilon T \\
\text{iff} \quad & \mathbb{E}[\textstyle\sum_{t=1}^T r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^T r_i(s_i^t, s_{-i}^t)] < \epsilon T \\
\text{iff} \quad & \mathbb{E}[\textstyle\sum_{t=1}^T r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^T \sum_{s_i \in S_i} \mathbf{1}_{\{s_i^t = s_i\}} r_i(s_i, s_{-i}^t)] < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^T \sum_{s_i \in S_i} q_i^t(s_i) r_i(s_i, s_{-i}^t)] < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^T \mathbb{E}[r_i(q_i^t, s_{-i}^t)] < \epsilon T
\end{aligned}$$

□

Theorem 2.1.13 *No external regret implies no model-based regret.*

Proof 2.1.14 The result follows via Theorems 2.1.5 and 2.1.11. In particular, an algorithm satisfies no external regret iff for arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i, \; \mathbb{E}[\text{ER}_{S_i \to \bar{s}_i}^T] < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^T \mathbb{E}[r_i(q_i^t, s_{-i}^t)] < \epsilon T \\
\Rightarrow \quad & \textstyle\sum_{t=1}^T \mathbb{E}[r_i(s_i^{*t}, q_{-i}^t)] - \sum_{t=1}^T \mathbb{E}[r_i(q_i^t, q_{-i}^t)] < \epsilon T \\
\text{iff} \quad & \forall \bar{s}_i \in S_i, \; \mathbb{E}[\text{MR}_{S_i \to \bar{s}_i}^T] < \epsilon T
\end{aligned}$$

Therefore, the algorithm satisfies no model-based regret. Note that this result follows directly from the fact that $\text{MR}_{s_i \to \bar{s}_i}^T = \mathbb{E}[\text{ER}_{s_i \to \bar{s}_i}^T]$. □

	a	*b*	*c*
A	2,2	1,1	1,0
B	1,1	2,2	1,0
C	0,1	0,1	0,0

Figure 2.1: No External Regret

Remark 2.1.15 No external regret does not ensure convergence inside the set O^∞.

Figure 2.1 depicts a game which permits a sequence of plays that exhibits no external regret and employs strategies outside of R^∞, D^∞, and O^∞. In particular, a sequence of plays which consists of $(A, a), (B, b)$, and (C, c), each appearing $\frac{1}{3}$ of the time, exhibits no external regret. If player 1 were to consider playing strategy A (B) everywhere, while this would increase the payoffs obtained whenever strategy C is played, it would equivalently decrease the payoffs obtained whenever strategy B (A) is played. Thus, the given sequence of plays yields no external regret for player 1, and similarly for player 2. However, $R^\infty = D^\infty = O^\infty = \{A, B\} \times \{a, b\}$; in particular, strategies C and c are eliminated. It follows that no external regret learning need not converge inside the set R^∞, D^∞, or O^∞.

Recall that no external regret implies no model-based regret. It follows that like no external regret, no model-based regret does not ensure convergence inside the set of unoverwhelmed strategies. This type of situation cannot arise in sequences of plays that satisfy no internal regret, since learning via no internal regret converges to correlated equilibrium, which is contained by the set of undominated strategies. No internal regret is described in the following section.

2.1.3 No Internal Regret

Recall that an algorithm is said to exhibit no external regret if the difference between the cumulative payoffs that are achieved by the algorithm and those that could be achieved by any fixed alternative is insignificant. In particular, the no external regret criterion considers the substitution of *all* the algorithmic decisions by one unique strategy. But that strategy, while it may be preferable to *some* of the algorithmic choices, need not be preferable everywhere. The no internal regret criterion is a refinement of the no external regret criterion in which the only substitutions that are considered are those which are preferable: *i.e.*, regret is positive, when one strategy is considered in place of another. This alternative measure of worst-case performance is due to Foster and Vohra [37].

The mathematical formulation of the no internal regret criterion arises out of a slight modification of the no external regret condition (see Equation 2.5). Let

$$\mathrm{IR}^T_{s_i \to \bar{s}_i} = (\mathrm{R}^T_{s_i \to \bar{s}_i})^+ \tag{2.7}$$

where $X^+ = max\{X, 0\}$, and let

$$\mathrm{IR}^T_{S_i \to \bar{s}_i} = \sum_{s_i \in S_i} \mathrm{IR}^T_{s_i \to \bar{s}_i} \tag{2.8}$$

Definition 2.1.16 Given a sequence of plays $\{s^t\}$ of length T, the sequence of plays $\{s_i^t\}$ for player i is said to exhibit *no internal regret* iff $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathrm{IR}^T_{S_i \to \bar{s}_i} < \epsilon T$$

The no internal regret optimality criterion compares one sequence of plays to a second in which a given strategy is everywhere replaced by the same strategy if the replacement strategy yields cumulative payoffs greater than those achieved by the original. Now consider the following substitution condition: in a sequence of plays, all occurrences of strategy a_i are replaced by a_i^*, all occurrences of b_i replaced by b_i^*, and so on, where x_i^* is a strategy which achieves maximal cumulative payoffs whenever strategy x_i is played. This latter criterion is equivalent to the original definition of no internal regret, as the following lemma demonstrates.

Lemma 2.1.17 *Given a sequence of plays $\{s^t\}$ of length T. Corresponding to every $s_i \in S_i$ that appears in the sequence $\{s_i^t\}$, there exists s_i^*, which is optimal in the following sense:*

$$s_i^* = \arg\max_{\bar{s}_i \in S_i} \sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} r_i(\bar{s}_i, s_{-i}^t)$$

Consider the comparative sequence of plays $\{s_i^{*t}\}$ which is defined s.t. *$\forall t$, $s_i^{*t} = s_i^*$ whenever $s_i^t = s_i$. The sequence of plays $\{s_i^t\}$ for player i satisfies no internal regret iff $\forall \epsilon > 0$,*

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} r_i(s_i^t, s_{-i}^t) < \epsilon T$$

Proof 2.1.18 The lemma follows directly from the construction of the sequence of plays $\{s_i^{*t}\}$ and the definition of no internal regret. For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i,\ \mathrm{IR}^T_{S_i \to \bar{s}_i} < \epsilon T \\
\text{iff}\ & \forall \bar{s}_i \in S_i,\ \textstyle\sum_{s_i \in S_i} \left(\sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s_{-i}^t) - r_i(s_i, s_{-i}^t)] \right)^+ < \epsilon T \\
\text{iff}\ & \max_{\bar{s}_i \in S_i} \textstyle\sum_{s_i \in S_i} \left(\sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s_{-i}^t) - r_i(s_i, s_{-i}^t)] \right)^+ < \epsilon T \\
\text{iff}\ & \textstyle\sum_{s_i \in S_i} \left(\sum_{t=1}^{T} \mathbf{1}_{\{s_i^t = s_i\}} [r_i(s_i^*, s_{-i}^t) - r_i(s_i, s_{-i}^t)] \right) < \epsilon T \\
\text{iff}\ & \textstyle\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} r_i(s_i^t, s_{-i}^t) < \epsilon T
\end{aligned}$$

□

Example 2.1.19 Recall the game considered in Figure 2.1. While the sequence of plays which consists of $(A, a), (B, b)$, and (C, c), each appearing $\frac{1}{3}$ of the time, exhibits no external regret, this sequence does not exhibit no internal regret. In particular, whenever (C, c) is played, there is internal regret, since greater payoffs would be achieved by a sequence which substitutes either $A(a)$ or $B(b)$ everywhere $C(c)$ appears. In contrast, the sequence of plays $(A, b), (B, b)$ repeated *ad infinitum* exhibits no internal regret for player 2, since player 2, would not achieve greater payoffs by substituting either strategy a or c everywhere for strategy b. On the other hand, the sequence of plays for player 1 does not satisfy the property of no internal regret, since player 1 would achieve greater payoffs by playing strategy B everywhere in place of strategy A. □

Definition 2.1.20 Given any opposing sequence of plays $\{s_{-i}^t\}$ of length T that is employed by the environment, an algorithm is said to exhibit *no internal regret* iff it gives rise to a sequence of weights $\{q_i^t\}$ *s.t.* $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathbb{E}[\text{IR}_{S_i \to \bar{s}_i}^T] \quad < \quad \epsilon T$$

Theorem 2.1.21 *Let $\{s_i^{*t}\}$ be as defined in Theorem 2.1.17. An algorithm that gives rise to sequence of weights $\{q_i^t\}$ of length T exhibits no internal regret against opposing sequence of plays $\{s_{-i}^t\}$ iff $\forall \epsilon > 0$,*

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s_{-i}^t) - \sum_{t=1}^{T} \mathbb{E}[r_i(q_i^t, s_{-i}^t)] < \epsilon T$$

Proof 2.1.22 The proof follows from Lemma 2.1.17, and is analogous to the proof of Theorem 2.1.11. □

Theorem 2.1.23 *No internal regret implies no external regret.*

Proof 2.1.24 Note that $\forall T, \forall \bar{s}_i \in S_i$, $\text{ER}_{S_i \to \bar{s}_i}^T \leq \text{IR}_{S_i \to \bar{s}_i}^T$. Now, since no internal regret implies that $\forall s_i \in S_i$, $\mathbb{E}[\text{IR}_{S_i \to \bar{s}_i}^T] < \epsilon T$, it follows that $\mathbb{E}[\text{ER}_{S_i \to \bar{s}_i}^T] < \epsilon T$, for arbitrary $\epsilon > 0$. Thus, no internal regret implies no external regret. □

Theorem 2.1.25 (Foster and Vohra, 1997) *Learning that achieves no internal regret converges to correlated equilibrium.*

2.1.4 No Clairvoyant Regret

This section discusses the strongest type of regret described in this chapter, namely no clairvoyant regret. Intuitively, a sequence of plays is said to exhibit no clairvoyant regret if it contains only best responses. The mathematical formulation of the no clairvoyant regret criterion arises out of a modification of the original definition of regret, in the spirit of the no internal regret condition (see Equation 2.4). Let

$$\text{CR}_{s_i \to \bar{s}_i}^t \quad = \quad (\text{R}_{s_i \to \bar{s}_i}^t)^+ \tag{2.9}$$

The expressions $\text{CR}_{s_i \to \bar{s}_i}^T$ and $\text{CR}_{S_i \to \bar{s}_i}^T$ are defined as usual.

Definition 2.1.26 Given a sequence of plays $\{s^t\}$ of length T, the sequence of plays $\{s_i^t\}$ for player i is said to exhibit *no clairvoyant regret* iff $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\text{CR}^T_{S_i \to \bar{s}_i} \quad < \quad \epsilon T$$

Lemma 2.1.27 *Given a sequence of plays $\{s^t\}$ of length T. Consider a sequence of plays $\{s_i^{*t}\}$ which is constructed* s.t. *$\forall t$, $s_i^{*t} = s_i^*$, where at time t, s_i^* is optimal in the following sense:*

$$s_i^* = \arg\max_{\bar{s}_i \in S_i} r_i(\bar{s}_i, s^t_{-i})$$

The sequence of plays $\{s_i^t\}$ exhibits no clairvoyant regret iff $\forall \epsilon > 0$,

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s^t_{-i}) - \sum_{t=1}^{T} r_i(s_i^t, s^t_{-i}) < \epsilon T$$

Proof 2.1.28 The lemma follows directly from the construction of the sequence of plays $\{s_i^{*t}\}$ and the definition of no clairvoyant regret. For arbitrary $\epsilon > 0$,

$$\begin{aligned}
& \forall \bar{s}_i \in S_i,\ \text{CR}^T_{S_i \to \bar{s}_i} < \epsilon T \\
\text{iff} \quad & \forall \bar{s}_i \in S_i,\ \textstyle\sum_{t=1}^T \sum_{s_i \in S_i} \left(\mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s^t_{-i}) - r_i(s_i, s^t_{-i})] \right)^+ < \epsilon T \\
\text{iff} \quad & \max_{\bar{s}_i \in S_i} \textstyle\sum_{t=1}^T \sum_{s_i \in S_i} \left(\mathbf{1}_{\{s_i^t = s_i\}} [r_i(\bar{s}_i, s^t_{-i}) - r_i(s_i, s^t_{-i})] \right)^+ < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T \sum_{s_i \in S_i} \left(\mathbf{1}_{\{s_i^t = s_i\}} [r_i(s_i^*, s^t_{-i}) - r_i(s_i, s^t_{-i})] \right) < \epsilon T \\
\text{iff} \quad & \textstyle\sum_{t=1}^T r_i(s_i^{*t}, s^t_{-i}) - \sum_{t=1}^T r_i(s_i^t, s^t_{-i}) < \epsilon T
\end{aligned}$$

□

Definition 2.1.29 Given any opposing sequence of plays $\{s^t_{-i}\}$ of length T that is employed by the environment, an algorithm is said to exhibit *no clairvoyant regret* iff it gives rise to a sequence of weights $\{q_i^t\}$ *s.t.* $\forall \epsilon > 0, \forall \bar{s}_i \in S_i$,

$$\mathbb{E}[\text{CR}^T_{S_i \to \bar{s}_i}] \quad < \quad \epsilon T$$

Theorem 2.1.30 *Let $\{s_i^{*t}\}$ be as defined in Theorem 2.1.27. An algorithm that gives rise to a sequence of weights $\{q_i^t\}$ of length T exhibits no clairvoyant regret against opposing sequence of plays $\{s^t_{-i}\}$ iff $\forall \epsilon > 0$,*

$$\sum_{t=1}^{T} r_i(s_i^{*t}, s^t_{-i}) - \sum_{t=1}^{T} \mathbb{E}[r_i(q_i^t, s^t_{-i})] < \epsilon T$$

Proof 2.1.31 The proof follows from Lemma 2.1.27, and is analogous to the proof of Theorem 2.1.11. □

Theorem 2.1.32 *No clairvoyant regret implies no internal regret.*

Proof 2.1.33 The proof is analogous to the proof of that no internal regret implies no external regret. □

Remark 2.1.34 No clairvoyant regret implies no internal regret, which implies no external regret, which implies no model-based regret.

Remark 2.1.35 If the players utilize clairvoyant learning algorithms that generate convergent sequences of weights, then these sequences converge to a Nash equilibrium. Otherwise, play moves through the space of Nash equilibria *ad infinitum.*

2.1.5 Discussion: Adaptive Learning

Milgrom and Roberts [78] defined a property of learning that applies to repeated scenarios in which individual players are rational, but they have no knowledge of one another's rationality. So-called *consistency with adaptive learning* is satisfied if the only strategies which are eventually played are those which are not dominated with respect to the recent history of play. Repeated play among individually rational agents that is consistent with adaptive learning eventually approaches collectively rational behavior; in particular, such play converges to D^∞.

2.1.6 Discussion: Responsive Learning

In network contexts, it has been argued [43, 50] that the key property which entails learning is not optimality, but rather is *responsiveness*. Responsiveness is the ability to respond to changes in the environment in bounded time. When the most basic of optimality criteria, namely no model-based regret, is combined with responsiveness and a certain monotonicity property, this leads to the class of so-called *reasonable* learning algorithms introduced in [43]. It is shown in [43] that the asymptotic play of a set of reasonable learners lies within O^∞.[4]

[4] It is conjectured, however, that O^∞ is not a precise solution concept, only an upper bound (see [43, 50]).

2.2 Learning Algorithms

The remainder of this chapter describes a suite of learning algorithms which satisfy the optimality and responsiveness criteria that are described in the first half of this chapter: two additive updating algorithms satisfy the reasonable learning criterion; learning via the mixing method, which also utilizes additive updating but scales the payoffs, achieves no external regret; learning via multiplicative updating also exhibits no external regret; finally, two algorithms are described which satisfy no internal regret. Note that all of these algorithms are non-Bayesian, in that players do not maintain a set of prior probabilistic beliefs with respect to which they optimize play. In contrast, these algorithms maintain a vector of weights over the set of pure strategies that serves as a mixed strategy probability distribution during play. The algorithms are distinguished by their manners of computing this vector of weights.

Most of these algorithms were initially proposed for use in settings quite different than networking, where responsiveness is not of interest and the available information level is significantly higher. In what follows, all the algorithms are extended for use in network contexts. The algorithms which are applicable in high-information settings are called *informed*, while those designed for low-information settings are called *naive*. Overall, the algorithms are considered in four varieties, depending on whether they are informed or naive, and whether or not they are responsive. The informed, non-responsive varieties of the algorithms are described in full detail, followed by the appropriate modifications for the naive and responsive variants.

2.2.1 Notation

Given a repeated strategic form game $\Gamma^t = (\mathcal{I}, (S_i, r_i)_{i \in \mathcal{I}})^t$. The algorithms that follow are presented from the point of view of player $i \in \mathcal{I}$, as if player i is playing a game against nature, where nature is taken to be a conglomeration of all the opponents of player i. From this perspective, let $r^t_{s_i}$ denote the payoffs achieved by player i at time t via strategy s_i; more specifically, $r^t_{s_i} = r(s_i, s^t_{-i})$, but since s^t_{-i} is often not known, it is therefore not denoted. Mixed strategy weights for player i at time t are given by the probability vector $w^t_i = (w^t_{s_i})_{s_i \in S_i}$.

2.2.2 Responsive Learning Automata

This section describes responsive learning automata, first introduced by Shenker and Friedman [42], which are responsive, as the name suggests, and applicable in naive settings. This algorithm is an extension of the usual sorts of learning automata (see Narendra and Thathachar [82] for a survey of the literature) which is (1) responsive, and (2) handles positive payoffs outside of the range $[0, 1]$. This discussion includes a further extension by which the algorithm is modified to handle negative payoffs.

Learning automata are applicable in naive settings. In particular, if strategy s_i is employed at time t, then the updating procedure depends only on $r^t_{s_i}$. The main idea of learning automata is to update weights by adding a factor that is based on the current payoff to the past weight of the current strategy, and subtracting some fraction of this factor from the past weights of the strategies which are not played. This is achieved by updating weights as follows, if strategy s_i is played at time t:

$$w^{t+1}_{s_i} = w^t_{s_i} + \gamma r^t_{s_i} \sum_{\bar{s}_i \neq s_i} w^t_{\bar{s}_i} \tag{2.10}$$

$$w^{t+1}_{\bar{s}_i} = w^t_{\bar{s}_i}(1 - \gamma r^t_{s_i}),\ \forall \bar{s}_i \neq s_i \tag{2.11}$$

where $0 < \gamma < 1$ is a parameter which controls the tradeoff between learning rapidly (when γ is close to 1) and accuracy (when γ is close to 0). The rightmost terms in these equations are normalization factors. Notice that $0 \leq w^t_{s_i} \leq 1$ requires that $0 \leq r^t_{s_i} \leq 1$.

To achieve responsiveness in learning automata, it suffices to assign some minimal probability, say $0 < \epsilon \leq 1$, to all strategies. Responsive learning automata are defined as follows: for parameter $0 < \gamma, \epsilon \leq 1$, if strategy s_i is employed at time t,

$$w^{t+1}_{s_i} = w^t_{s_i} + \gamma r^t_{s_i} \sum_{\bar{s}_i \neq s_i} w^t_{\bar{s}_i} a^t_{\bar{s}_i} \tag{2.12}$$

$$w^{t+1}_{\bar{s}_i} = w^t_{\bar{s}_i}(1 - \gamma r^t_{s_i} a^t_{\bar{s}_i}), \quad \forall \bar{s}_i \neq s_i \tag{2.13}$$

where

$$a^t_{\bar{s}_i} = \min\left\{1, \frac{w^t_{\bar{s}_i} - \epsilon}{\gamma r^t_{s_i} w^t_{\bar{s}_i}}\right\} \tag{2.14}$$

The term $a^t_{\bar{s}_i}$ is derived as follows:

$$\begin{aligned} & w^t_{\bar{s}_i}(1-\gamma r^t_{s_i}) \geq \epsilon \\ \Rightarrow \quad & \frac{w^t_{\bar{s}_i}-\epsilon}{\gamma r^t_{s_i} w^t_{\bar{s}_i}} \geq 1 \\ \Rightarrow \quad & a^t_{\bar{s}_i} = \min\left\{1, \frac{w^t_{\bar{s}_i}-\epsilon}{\gamma r^t_{s_i} w^t_{\bar{s}_i}}\right\} \end{aligned}$$

The term $a^t_{\bar{s}_i}$ ensures that the probabilities $w^t_{\bar{s}_i}$ are bounded below by ϵ. More specifically, if $w^t_{\bar{s}_i}(1-\gamma r^t_{s_i}) \geq \epsilon$, then $\frac{w^t_{\bar{s}_i}-\epsilon}{\gamma r^t_{s_i} w^t_{\bar{s}_i}} \geq 1$, and it suffices to update probabilities as usual: *i.e.*, $a^t_{\bar{s}_i} = 1$. If this condition is not satisfied, however, then $a^t_{\bar{s}_i}$ is defined such that $w^t_{\bar{s}_i} = \epsilon$. Note that if all the probabilities $w^t_{\bar{s}_i}$ are bounded below by ϵ, then the probability $w^t_{s_i}$ is bounded above by $1-\epsilon(\frac{|S_i|-1}{|S_i|})$. Notice also that the normalizing term $a^t_{\bar{s}_i}$ eliminates the requirement the $0 \leq r^t_{s_i} \leq 1$.

The definition of responsive learning automata given by Equations 2.12 – 2.13 is not suitable to handle negative payoffs. While $a^t_{\bar{s}_i}$ ensures that $w^t_{\bar{s}_i} \geq \epsilon$, in the case of negative payoffs, it is also necessary to ensure that $w^t_{\bar{s}_i} \leq 1-\epsilon(\frac{|S_i|-1}{|S_i|})$. This is accomplished via an additional term $b^t_{\bar{s}_i}$ which is defined as follows.

$$b^t_{\bar{s}_i} = \min\left\{1, \frac{w^t_{\bar{s}_i} + \epsilon(\frac{|S_i|-1}{|S_i|}) - 1}{\gamma r^t_{s_i} w^t_{\bar{s}_i}}\right\} \tag{2.15}$$

A further requirement in the case of negative payoffs is that $w^t_{s_i} \geq \epsilon$. The term $c^t_{s_i}$ ensures that this condition is satisfied:

$$c^t_{s_i} = \min\left\{1, \frac{\epsilon - w^t_{s_i}}{\gamma r^t_{s_i} \sum_{\bar{s}_i \neq s_i} a^t_{\bar{s}_i} b^t_{\bar{s}_i}}\right\} \tag{2.16}$$

The final version of responsive learning automata which is equipped to handle both positive and negative payoffs in the range $[x, y]$, for $x, y \in \mathbb{R}$, is as follows: for $0 < \gamma, \epsilon \leq 1$, if strategy i is played at time t,

$$w^{t+1}_{s_i} = w^t_{s_i} + \gamma r^t_{s_i} c^t_{s_i} \sum_{j \neq i \in N} w^t_{\bar{s}_i} a^t_{\bar{s}_i} b^t_{\bar{s}_i} \tag{2.17}$$

$$w^{t+1}_{\bar{s}_i} = w^t_{\bar{s}_i}(1 - \gamma r^t_{s_i} c^t_{s_i} a^t_{\bar{s}_i} b^t_{\bar{s}_i}), \quad \forall j \neq i \in N \tag{2.18}$$

where $a^t_{\bar{s}_i}, b^t_{\bar{s}_i}$, and $c^t_{s_i}$ are defined in Equations 2.14, 2.15, and 2.16, respectively.

It is shown in Shenker and Friedman [43] that responsive learning automata are reasonable learners. It follows that learning via this algorithm converges to O^∞. The following two sections describe alternative means of learning via additive updating, the first of which is also reasonable, due to Roth and Erev [30], and the second of which satisfies no external regret, due Foster and Vohra [37].

2.2.3 Additive Updating

A second example of a reasonable learning algorithm, which also happens to employ additive updating, is the responsive and naive algorithm of Roth and Erev [30]. In fact, this algorithm can be derived from the corresponding informed and non-responsive version. Additive updating in informed settings is based on a cumulative sum of weighted payoffs achieved by all strategies. This weighted sum is computed as if randomized strategies, or interior points, were attainable. In particular, define $\sigma^t_{s_i}$ as follows:

$$\sigma^t_{s_i} = \sum_{x=0}^{t} w^x_{s_i} r^x_{s_i} \tag{2.19}$$

Now the weight of strategy $s_i \in S_i$ at time $t+1$ is the ratio of the weighted payoffs achieved by strategy s_i to the sum of the weighted payoffs achieved by all strategies:

$$w^{t+1}_{s_i} = \frac{\sigma^t_{s_i}}{\sum_{\bar{s}_i \in S_i} \sigma^t_{\bar{s}_i}} \tag{2.20}$$

In naive settings, the only information pertaining to payoff functions that is ever available is the payoff of the strategy that is in fact employed at that time. Thus, the naive updating rule must utilize an estimate of weighted payoffs which depends only on this limited information. Such an estimate is given by $\hat{\sigma}^t_{s_i}$, which is defined as follows:[5]

$$\hat{\sigma}^t_{s_i} = \sum_{x=1}^{t} \mathbf{1}^x_{s_i} r^x_{s_i} \tag{2.21}$$

[5] Roth and Erev [30] use the following estimate of weighted payoffs: for $0 < \epsilon < 1$,

$$\hat{\sigma}^t_{s_i} = \sum_{x=1}^{t} \left[(1-\epsilon)\mathbf{1}^x_{s_i} r^x_{s_i} + \frac{\epsilon}{|S_i|-1} \sum_{\bar{s}_i \neq s_i} \mathbf{1}^x_{\bar{s}_i} r^x_{\bar{s}_i} \right]$$

Note that this estimator "kills two birds with one stone", since it both estimates payoffs and ensures that probabilities are non-zero.

where $\mathbf{1}^t_{s_i}$ is the indicator function: $\mathbf{1}^t_{s_i} = 1$ if strategy s_i is played at time t, and $\mathbf{1}^t_{s_i} = 1$, otherwise. Note that $\hat{\sigma}^t_{s_i}$ is an accurate and unbiased estimator of $\sigma^t_{s_i}$. In particular, the expected value of this estimator is the actual value of cumulative weighted payoffs:

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^t_{s_i}] &= \sum_{x=1}^{t} \mathbb{E}[\mathbf{1}^x_{s_i}] r^x_{s_i} \\ &= \sum_{x=1}^{t} w^x_{s_i} r^x_{s_i} \\ &= \sigma^t_{s_i}\end{aligned}$$

In naive settings, the updating rule given in Equation 2.20 is modified such that $\hat{\sigma}^t_{s_i}$ is used in place of $\sigma^t_{s_i}$. This update procedure yields a set of weights which must be adjusted for use in naive settings, in order to ensure that the space of possible payoffs be adequately explored. This is achieved by imposing an artificial lower bound on the probability with which strategies are played. In particular, let

$$\hat{w}^t_{s_i} = (1-\epsilon) w^t_{s_i} + \frac{\epsilon}{|S_i|} \tag{2.22}$$

Finally, this additive updating rule can be modified to achieve responsiveness in both informed and naive settings by exponentially smoothing the cumulative payoffs. This technique, which is inspired by Roth and Erev [30], updates according to $\tilde{\sigma}^t_{s_i}$ in Equation 2.20, rather than $\sigma^t_{s_i}$ or $\hat{\sigma}^t_{s_i}$. In informed and naive settings, respectively, for $0 < \gamma \leq 1$,

$$\tilde{\sigma}^{t+1}_{s_i} = (1-\gamma)\tilde{\sigma}^t_{s_i} + r^t_{s_i} \qquad \text{and} \qquad \tilde{\sigma}^{t+1}_{s_i} = (1-\gamma)\tilde{\sigma}^t_{s_i} + \mathbf{1}^t_{s_i} r^t_{s_i} \tag{2.23}$$

Note that the algorithm studied by Roth and Erev [30] is precisely the naive and responsive version of additive updating which is presented in this section. It has been observed (see [50]) that this algorithm is reasonable in the sense of Shenker and Friedman [43]; therefore, learning converges to O^∞. The following sections describe additive and multiplicative learning rules that satisfy the no external regret optimality criterion.

2.2.4 Additive Updating Revisited

This section presents an additive updating rule due to Foster and Vohra, known as the mixing method [36], which achieves no external regret. This algorithm is presented in its original formulation, as a non-responsive algorithm applicable in informed settings, as well as in its responsive and naive variations.

The mixing method updates weights based on the cumulative payoffs achieved by all strategies, including the payoffs that would have been obtained by strategies which were not played. The cumulative payoffs obtained at time t for strategy s_i (notation $\rho^t_{s_i}$) is computed as follows:

$$\rho^t_{s_i} = \sum_{x=1}^{t} r^x_{s_i} \tag{2.24}$$

Now consider the following update rule, in which the weight of strategy $s_i \in S_i$ at time $t+1$ is the ratio of the cumulative payoffs achieved by strategy s_i to the sum of the cumulative payoffs achieved by all strategies:

$$w^{t+1}_{s_i} = \frac{\rho^t_{s_i}}{\sum_{\bar{s}_i \in S_i} \rho^t_{\bar{s}_i}} \tag{2.25}$$

As it stands, the update rule presented in Equation 2.25 performs poorly; in particular, it does not exhibit no external regret. For example, consider a one-player game, where Player has two strategies, say A and B. Suppose the game is such that strategy A always yields payoffs of 1 and strategy B always yields payoffs of 2 for Player. In this game, fixed strategy B yields average payoffs of 2; however, the algorithm assigns weights $w^t_A = 1/3$ and $w^t_B = 2/3$, which yields expected payoffs of only 5/3.

The performance of this algorithm can be improved via a technique known as the *mixing method* introduced in Foster and Vohra [36]. In particular, by scaling the difference between payoffs at a rate of $\sqrt{t}$, the mixing method achieves no external regret. Consider the case of two strategies, say A and B. According to the additive updating rule given in Equation 2.25,

$$w^{t+1}_A = \frac{\rho^t_A}{\rho^t_A + \rho^t_B} \quad \text{and} \quad w^{t+1}_B = \frac{\rho^t_B}{\rho^t_A + \rho^t_B} \tag{2.26}$$

It follows that

$$w_A^{t+1} - w_B^{t+1} = \frac{\rho_A^t - \rho_B^t}{\rho_A^t + \rho_B^t} \tag{2.27}$$

The mixing method modifies this straightforward updating procedure by scaling the difference between weights as follows. For $\alpha > 0$,

$$w_A^{t+1} - w_B^{t+1} = \frac{\alpha(\rho_A^t - \rho_B^t)}{\rho_A^t + \rho_B^t} \tag{2.28}$$

It is shown in Foster and Vohra [36] that the optimal value of α in Equation 2.28 is $\sqrt{t}$, and moreover, this algorithm exhibits no external regret. If the number of strategies is greater than 2, then the generalized algorithm utilizes pairwise mixing of strategies via Equation 2.28, followed by further mixing of the mixtures.

The mixing method can be modified for use in naive settings by utilizing an estimate of cumulative payoffs that depends only on the payoffs obtained by the strategies that are actually employed and the approximate weights associated with those strategies. First of all, as in the case of additive updating (see Equation 2.22), it is necessary to assign minimal probabilities to all strategies in order that the space of payoff functions be adequately explored. Now let

$$\hat{r}_{s_i}^t = \mathbf{1}_{s_i}^t \frac{r_{s_i}^t}{\hat{w}_{s_i}^t} \tag{2.29}$$

In other words, $\hat{r}_{s_i}^t$ is equal to 0 if strategy s_i is not employed at time t; otherwise, $\hat{r}_{s_i}^t$ is the payoff achieved by strategy s_i at time t scaled by the likelihood of playing strategy s_i. Estimated cumulative payoffs (notation $\hat{\rho}_{s_i}^t$) are given by:

$$\hat{\rho}_{s_i}^t = \sum_{x=1}^{t} \hat{r}_{s_i}^x \tag{2.30}$$

Note that $\hat{\rho}_{s_i}^t$ is an accurate and unbiased estimator.

$$\begin{aligned} \mathbb{E}[\hat{\rho}_{s_i}^t] &= \sum_{x=1}^{t} \mathbb{E}[\mathbf{1}_{s_i}^t] \frac{r_{s_i}^t}{\hat{w}_{s_i}^t} \\ &= \sum_{x=1}^{t} r_{s_i}^x \\ &= \rho_{s_i}^t \end{aligned}$$

The naive variant of the mixing method uses $\hat{\rho}_{s_i}^t$ in place of $\rho_{s_i}^t$ in Equation 2.28.

Finally, the mixing method can be made responsive via exponential smoothing. In the responsive variant of this algorithm $\tilde{\rho}^t_{s_i}$ is substituted for either $\rho^t_{s_i}$ or $\hat{\rho}^t_{s_i}$, depending on whether the setting is informed or naive. In particular, for $0 < \gamma \leq 1$, in informed settings and naive settings, respectively,

$$\tilde{\rho}^{t+1}_{s_i} = (1-\gamma)\tilde{\rho}^t_{s_i} + r^{t+1}_{s_i} \qquad \text{and} \qquad \tilde{\rho}^{t+1}_{s_i} = (1-\gamma)\tilde{\rho}^t_{s_i} + \hat{r}^{t+1}_{s_i} \tag{2.31}$$

The naive and responsive variant of the mixing method is also a reasonable learning algorithm. In fact, it satisfies an even stronger property which could be defined in terms of the no external regret optimality criterion, rather than no model-based regret, together with responsiveness.

The following section describes learning via multiplicative updating which like the mixing method exhibits no external regret. The development of the variants of the multiplicative updating algorithm is analogous to the development of additive updating.

2.2.5 Multiplicative Updating

This section describes an algorithm due to Freund and Schapire [39] that achieves no external regret in informed settings via multiplicative updating. The multiplicative update rule utilizes the cumulative payoffs achieved by all strategies, including the surmised payoffs of those strategies which are not played. In particular, the weight assigned to strategy s_i at time $t+1$, for $\beta > 0$, is given by:

$$w^{t+1}_{s_i} \quad = \quad \frac{(1+\beta)^{\rho^t_{s_i}}}{\sum_{\bar{s}_i \in S_i}(1+\beta)^{\rho^t_{\bar{s}_i}}} \tag{2.32}$$

The multiplicative updating rule given in Equation 2.32 can be modified in a manner identical to the mixing method, using $\hat{r}^t_{s_i}$ and $\hat{w}^t_{s_i}$, to become applicable in naive settings, and using $\tilde{\rho}^t_{s_i}$ to achieve responsiveness. A naive variant of this multiplicative updating algorithm which achieves no external regret appears in Auer, Cesa-Bianchi, Freund, and Schapire [3]. Like the mixing method, the naive and responsive variant of multiplicative updating is a reasonable learning algorithm.

2.2.6 No Internal Regret Learning

This section describes an algorithm due to Foster and Vohra [37] which achieves no internal regret in informed environments, and a simple implementation due to Hart and Mas-Colell [57]. In addition, the appropriate naive and responsive modifications are presented. Learning via the following no internal regret algorithms converges to correlated equilibrium, and therefore converges inside the set D^{∞}.

Consider the case of a 2-strategy informed game, with strategies A and B. The components of the weight vector, namely w_A^{t+1} and w_B^{t+1}, are updated according to the following formulae, which reflect cumulative feelings of regret:

$$w_A^{t+1} = \frac{\mathrm{IR}_{B\to A}^t}{\mathrm{IR}_{A\to B}^t + \mathrm{IR}_{B\to A}^t} \quad \text{and} \quad w_B^{t+1} = \frac{\mathrm{IR}_{A\to B}^t}{\mathrm{IR}_{A\to B}^t + \mathrm{IR}_{B\to A}^t} \tag{2.33}$$

If the regret for having played strategy $\bar{s}_i$ rather than strategy s_i is significant, then the algorithm updates weights such that the probability of playing strategy s_i is increased. In general, if strategy s_i is played at time t,

$$w_{\bar{s}_i}^{t+1} = \frac{1}{\mu}\mathrm{IR}_{s_i\to\bar{s}_i}^t \quad \text{and} \quad w_{s_i}^{t+1} = 1 - \sum_{\bar{s}_i \neq s_i} w_{\bar{s}_i}^{t+1} \tag{2.34}$$

where μ is a normalizing term that is chosen *s.t.*:

$$\mu > (|S_i| - 1) \max_{\bar{s_i}\in S_i} \mathrm{IR}_{s_i\to\bar{s}_i}^t \tag{2.35}$$

This generalized algorithm is due to Hart and Mas-Colell [57].

As usual in naive settings, an estimate of internal regret is computed which is based only on the payoffs obtained by the strategies that are actually played and the approximate weights associated with those strategies, as in Equation 2.22. Recall from Equation 2.4, the instantaneous regret at time x for having played strategy s_i rather than playing strategy $\bar{s}_i$ is given by:

$$\mathrm{R}_{s_i\to\bar{s}_i}^x = r_{\bar{s}_i}^x - r_{s_i}^x \tag{2.36}$$

An estimated measure of expected regret $\hat{\mathrm{R}}_{s_i\to\bar{s}_i}^x$ is given by:

$$\hat{\mathrm{R}}_{s_i\to\bar{s}_i}^x = \hat{r}_{\bar{s}_i}^x - \hat{r}_{s_i}^x \tag{2.37}$$

where $\hat{r}^x_{s_i}$ and $\hat{r}^x_{\bar{s}_i}$ are defined as in Equation 2.29. The expected value of this estimated measure of regret is actual regret:

$$\begin{aligned}
\mathbb{E}[\hat{\mathrm{R}}^x_{s_i \to \bar{s}_i}] &= \mathbb{E}[\mathbf{1}^x_{\bar{s}_i} \frac{r^x_{\bar{s}_i}}{\hat{w}^x_{\bar{s}_i}} - \mathbf{1}^x_{s_i} \frac{r^x_{s_i}}{\hat{w}^x_{s_i}}] \\
&= \mathbb{E}[\mathbf{1}^x_{\bar{s}_i}] \frac{r^x_{\bar{s}_i}}{\hat{w}^x_{\bar{s}_i}} - \mathbb{E}[\mathbf{1}^x_{s_i}] \frac{r^x_{s_i}}{\hat{w}^x_{s_i}} \\
&= r^x_{\bar{s}_i} - r^x_{s_i} \\
&= \mathrm{R}^t_{s_i \to \bar{s}_i}
\end{aligned}$$

Now an estimate of cumulative regret is given by

$$\hat{\mathrm{R}}^t_{s_i \to \bar{s}_i} = \sum_{x=1}^{t} \mathbf{1}^x_{s_i} \hat{\mathrm{R}}^x_{s_i \to \bar{s}_i} \tag{2.38}$$

and an estimate of cumulative internal regret is given by

$$\mathrm{I}\hat{\mathrm{R}}^t_{s_i \to \bar{s}_i} = (\hat{\mathrm{R}}^t_{s_i \to \bar{s}_i})^+ \tag{2.39}$$

Finally, weights are updated as in Equation 2.34, with the estimate of cumulative internal regret $\mathrm{I}\hat{\mathrm{R}}^t_{s_i \to \bar{s}_i}$ used in place of $\mathrm{IR}^t_{s_i \to \bar{s}_i}$.

Like the additive and multiplicative updating algorithms, the no internal regret learning algorithm can be made responsive via exponential smoothing of regret. In the informed and naive cases respectively,

$$\tilde{\mathrm{R}}^{t+1}_{s_i \to \bar{s}_i} = (1-\gamma)\tilde{\mathrm{R}}^t_{s_i \to \bar{s}_i} + \mathbf{1}^{t+1}_{s_i} \mathrm{R}^{t+1}_{s_i \to \bar{s}_i} \tag{2.40}$$

where $\mathrm{R}^{t+1}_{s_i \to \bar{s}_i}$ denotes the instantaneous regret at time $t+1$ for playing strategy s_i rather than $\bar{s}_i$, and

$$\tilde{\mathrm{R}}^{t+1}_{s_i \to \bar{s}_i} = (1-\gamma)\tilde{\mathrm{R}}^t_{s_i \to \bar{s}_i} + \mathbf{1}^{t+1}_{s_i} \hat{\mathrm{R}}^{t+1}_{s_i \to \bar{s}_i} \tag{2.41}$$

where $\hat{\mathrm{R}}^{t+1}_{s_i \to \bar{s}_i}$ denotes the approximate instantaneous regret at time $t+1$ for playing strategy s_i rather than $\bar{s}_i$. Finally, it suffices to use

$$\mathrm{I}\tilde{\mathrm{R}}^t_{s_i \to \bar{s}_i} = (\tilde{\mathrm{R}}^t_{s_i \to \bar{s}_i})^+ \tag{2.42}$$

as a measure internal regret in the responsive case, where $\tilde{\mathrm{R}}^t_{s_i \to \bar{s}_i}$ is given by either Equation 2.40 or 2.41, depending on whether the setting is informed or naive.

2.2.7 Summary: Learning, Optimality, and Equilibria

This chapter unified a suite of learning algorithms from the game-theoretic, machine learning, and stochastic control literature that satisfy a series of related optimality criteria, which in turn converge to various generalizations of Nash equilibrium. The results of this survey are summarized in Table 2.1. Simulation experiments of these algorithms in practical settings are reported in the next several chapters.

Learning Algorithms	*Optimality Criteria*	*Equilibria*
Friedman and Shenker	Reasonable	O^∞
Erev and Roth	Reasonable	O^∞
Additive Updating	Adaptive Learning	D^∞
Foster and Vohra	Adaptive Learning	D^∞
Freund and Schapire	Adaptive Learning	D^∞
Hart and Mas-Colell	No Internal Regret	CE

Table 2.1: Learning, Optimality, and Equilibria

Chapter 3

Santa Fe Bar Problem

The *Santa Fe bar problem* (SFBP) was introduced by Brian Arthur [2], an economist at the Santa Fe Institute, in the study of bounded rationality and inductive learning. This problem and its natural extensions serve as abstractions of network flow control and routing problems. Here is the scenario:

> *N [(say, 100)] people decide independently each week whether to go to a bar that offers entertainment on a certain night ... Space is limited, and the evening is enjoyable if things are not too crowded – especially, if fewer than* 60 *[or, some fixed but perhaps unknown capacity c] percent of the possible* 100 *are present ... a person or agent goes (deems it worth going) if he expects fewer than* 60 *to show up or stays home if he expects more than* 60 *to go. Choices are unaffected by previous visits; there is no collusion or prior communication among the agents; and the only information available is the number who came in past weeks.*[1]

SFBP is a non-cooperative, repeated game. The players are the patrons of the bar. Their strategy sets consist of two strategies, namely go to the bar or stay at home. Finally, the payoffs of the game are determined by the total number of players that choose to go to the bar. In this chapter, it is shown that rational learning does not converge to Nash equilibrium in SFBP; however, it is also demonstrated that various forms of boundedly rational learning do in fact converge to Nash equilibrium.

[1] The problem was inspired by the El Farol bar in Santa Fe which offers live music on Thursday nights.

3.1 Introduction

SFBP can be viewed as an abstraction of the general problem of sharing resources of limited capacity, where the only interaction among agents occurs through the joint use of shared resources. This game-theoretic model is applicable to several real-world situations, ranging from fishermen fishing in common waters, to farmers polluting common water supplies, to various other versions of the tragedy of the commons [55]. Moreover, the set of applications is rapidly expanding, as game-like scenarios emerge in the telecommunications infrastructure, where network bandwidth and buffer space serve as shared resources (see, for example, [33] and [71]) as well as web sites and shared databases (see, for example, [43, 100]). In contrast to the recent trend of proposing solutions to network resource allocation problems based on pricing congestible resources [76], Arthur suggests bounded rationality and inductive learning as possible mechanisms for generating stable solutions to such problems. In this chapter, we present a theoretical formalization of an argument that perfect rationality and learning are inherently incompatible in SFBP. On the practical side, we demonstrate via simulations that computational learning in which agents exhibit low-rationality does indeed give rise to equilibrium behavior.

We motivate our theoretical argument with the following intuitive analysis of SFBP under the standard economic assumption of rationality. Define an *undercrowded* bar as one in which attendance is less than or equal to c, and define an *overcrowded* bar as one in which attendance is strictly greater than c. Let the utility of going to an undercrowded bar be $1/2$ and the utility of going to an overcrowded bar be $-1/2$; in addition, the utility of staying at home is 0, regardless of the state of the bar.[2] If a patron predicts that the bar will be undercrowded with probability p, then his rational best-reply is to go to the bar if $p > 1/2$ and to stay home if $p < 1/2$.[3] Now, if the patrons indeed learn to predict probability p accurately, then their predictions

[2] The definition of the utility of staying at home as 0, regardless of the state of the bar, can be replaced, without changing the argument, by: the utility of staying at home is $1/2$, whenever the bar is overcrowded, and the utility of staying at home is $-1/2$, whenever the bar is undercrowded.

[3] In the case where $p = 1/2$, the patrons are indifferent between attending the bar and staying home and may behave arbitrarily: *e.g.*, go to the bar with probability q. We show that in all but finitely many cases this condition is incompatible with learning.

eventually come to match the actual probability that the bar is undercrowded, as it is determined by their (possibly randomized) strategic best-replies. Herein lies a contradiction. If the patrons learn to predict that the bar will be undercrowded with probability $p < 1/2$, then, in fact the bar will be empty with probability 1; on the other hand, if the patrons learn to predict that the bar will be undercrowded with probability $p > 1/2$, then the bar will be full with probability 1.[4] We conclude that rational patrons cannot learn via repeated play to make accurate predictions. In particular, rationality precludes learning.

3.1.1 Logical Implications

This paradoxical outcome in SFBP is arrived at via a diagonalization process in the spirit of Russell's paradox [91].[5] Just as the truth of being in Russell's set depends on the fact of (not) being in the set, the value of going to the bar depends on the act of going (or not going) to the bar. For the sake of argument, consider a bar of capacity 1/2 in a world of a single patron.[6] If the patron does not go to the bar, then the bar is undercrowded, in which case her best-reply is to go to the bar. But now the bar is overcrowded, and so her best-reply is to stay at home. Thus, rationality dictates that this patron should go to the bar if and only if she should not go to the bar.

The aforementioned paradox similarly arises in the two-player game of matching pennies, where player 1 aims to *match* player 2, while player 2 aims to *mismatch* player 1. In fact, matching pennies can be viewed as a special case of SFBP in which there are two players and both positive and negative externalities:[7] if player 1 prefers to go to the bar only when player 2 attends as well, while player 2 prefers to go to the bar only when player 1 stays at home, then player 1 is the matcher while player 2 is the mismatcher. In matching pennies, if player 1 prefers *heads*, then player 2 prefers *tails*, but then player 1 prefers *tails*, at which point player 2 actually prefers *heads*, and finally, player 1 prefers *heads* once again. It follows that player 1 prefers *heads* iff player 1 prefers *tails*. Similarly, for player 2.

[4] Schelling [96] refers to phenomena of this kind as self-negating prophecies.

[5] Russell's set is defined as the set of all sets that are not elements of themselves: *i.e.*, $\mathcal{R} = \{X | X \notin X\}$. Note that $\mathcal{R} \in \mathcal{R}$ iff $\mathcal{R} \notin \mathcal{R}$.

[6] Similarly, one could consider a bar of capacity 1 and a married couple who act in unison.

[7] See Footnote 10 to understand SFBP in terms of solely negative externalities.

The logical conflict that arises in the game of matching pennies is closely related to the fact that this games has no pure strategy Nash equilibria [84]; similarly, SFBP has no symmetric pure strategy Nash equilibrium, except in degenerate cases. In order to resolve these paradoxes, game-theorists introduce mixed strategies. The unique Nash equilibrium in matching pennies is for both players to play each of *heads* and *tails* with probability 1/2; a mixed strategy Nash equilibrium in SFBP, is for all players to go to the bar with probability $p \approx c/N$ and to stay at home with probability $1 - p$.[8]

3.1.2 Game-Theoretic Implications

This chapter presents a negative result on convergence to Nash equilibrium in SFBP which formalizes the above diagonalization argument. Two sufficient conditions for convergence to Nash equilibrium are *rationality* and *predictivity*. By rationality, we mean that players play best-replies to their beliefs. Predictivity is one way in which to capture the notion of learning:[9] a player is said to be predictive if that player's beliefs eventually coincide with the truth about which he is predicting. If players learn to predict (*i.e.*, if beliefs indeed converge to opponents' actual strategies), then best-replies to beliefs constitute a Nash equilibrium. In what follows, we observe that SFBP has multiple mixed strategy Nash equilibria, and we argue that if the players employ predictive learning algorithms, then assuming rationality, play does not converge to one of these Nash equilibria. Conversely, *if play converges to Nash equilibrium, then play is either not rational or not learned.*

In a seminal work by Kalai and Lehrer [65], sufficient conditions are presented for predictivity in the form of the so-called Harsanyi hypothesis, or absolute continuity assumption. In particular, their paper suggests that convergence to Nash equilibrium is in fact possible. Our negative results complement the recent theorems reported in Nachbar [81] and Foster and Young [38], who argue that unless rather unusual conditions hold, any conditions that are sufficient for prediction are unlikely to hold. Nachbar shows that unless players' initial beliefs somehow magically coincide with Nash equilibrium, repeated play of strategic form games among Bayesian rational

[8] Technically, this symmetric Nash equilibrium is the solution p to the equation $\sum_{x=0}^{c} \binom{N}{x} p^x (1-p)^{N-x} = \sum_{x=c+1}^{N} \binom{N}{x} p^x (1-p)^{N-x}$, which is roughly c/N.

[9] Learning can also be understood in terms of various merging properties; see Kalai and Lehrer [66].

players does not generally converge to Nash equilibrium. Similarly, Foster and Young prove that in two-player games of incomplete information with unique mixed strategy Nash equilibria, rationality is not compatible with predictivity. Our theorems argue in a similar vein that unless certain strict regularity conditions are satisfied, no means of rational learning converges to Nash equilibrium in SFBP, an N player game with multiple mixed strategy Nash equilibria.

3.1.3 Computer Science Implications

SFBP and its natural extensions serve as abstractions of various congestion control problems that arise in networking. Many authors who capitalize on the potential for the theory of repeated games as a model of networking environments do so because of the difficulty to enforce cooperation in large-scale networks; instead, it is more realistic and more general to assume non-cooperative networks. This generality is modeled in repeated games by assuming that players are rational. Those same authors who study networking games assuming rationality, however, often also assume that the network operating point is a Nash equilibrium. One might hope to justify this assumption on the grounds that Nash equilibrium is the outcome of rational learning. It is the conclusion of this study, however, that Nash equilibrium is *not* the outcome of rational learning in games that model networking environments.

The second half of this chapter aims to resolve the paradoxes of the first half via simulation experiments in computational learning. In particular, it is shown that low-rationality learning yields equilibrium behavior. Similarly, in Arthur's original paper, he demonstrated via simulations that boundedly rational agents are capable of generating collective attendance centered around the capacity of the bar. In contrast to Arthur's approach, however, the learning algorithms considered in this study are simple, and are therefore feasible for use in network games. Moreover, we extend our study to a special case of the so-called New York City bar problem (NYCBP) in which there are exactly two bars, and observe similar convergence results. In summary, (highly) rational learning does not validate the assumption that Nash equilibrium describes the solution of network games; however, low-rationality learning indeed yields Nash equilibrium behavior.

This next section formalizes SFBP in terms of the theory of learning in repeated games. In Section 3.2.1, it is shown that best-reply dynamics, a learning algorithm for which Cournot proved convergence to pure strategy Nash equilibrium in models of duopoly [24], yields oscillatory behavior in SFBP. Sections 3.2.2 and 3.2.3 contain our main theoretical results, namely that traditional models of belief-based learning (*e.g.*, Bayesian updating) among rational players do not in general give rise to equilibrium behavior in SFBP. The second half of this chapter, beginning with Section 3.3, presents the results of simulations, and demonstrates that models of low-rationality learning in fact give rise to equilibrium behavior.

3.2 Theoretical Investigations

The Santa Fe bar problem is a repeated game of negative externalities.[10] We now formally define both the one-shot strategic form game, and the corresponding repeated game. The players are the inhabitants of Santa Fe; notation $\mathcal{N} = \{1, \ldots, N\}$, with $n \in \mathcal{N}$. For player n, the strategy set $S_n = \{0, 1\}$, where 1 corresponds to *go to the bar* while 0 corresponds to *stay home*. Let Q_n denote the set of probability distributions over S_n, with mixed strategy $q_n \in Q_n$. The expected payoffs obtained by player n depend on the particular strategic choice taken by player n, the value to player n of attending the bar, and a negative externality, which are defined as follows.

Let s_n denote the realization of mixed strategy q_n of player n; thus, $s = \sum_{n \in \mathcal{N}} s_n$ is the realized attendance at the bar. In addition, let $c \in \{0, \ldots, N\}$ denote the capacity of the bar. The externality f depends on s and c as follows: if the bar is undercrowded (*i.e.*, $s \leq c$), then $E(s) = 0$; on the other hand, if the bar is overcrowded (*i.e.*, $s > c$), then $E(s) = 1$. Finally, let $0 \leq \alpha_n \leq 1$ denote the value to player n of attending the bar, and without loss of generality assume $\alpha_n \leq \alpha_{n+1}$. Now the payoff function for player n for pure strategies $s_n \in S_n$ is given by $\pi_n(s_n, s) = \alpha_n - E(s)$, if $s_n = 1$, and $\pi_n(s_n, s) = 0$, otherwise: *i.e.*, $\pi_n(s_n, s) = s_n[\alpha_n - E(s)]$.[11] As usual, the

[10] An externality is a third-party effect. An example of a negative externality is pollution; an example of a positive externality is standardization. Although externalities are so-called because they are external to the game, it is natural to consider payoffs in terms of externalities when there are large numbers of players.

[11] Our results also hold in the case where $\pi_n(s_n, s) = E(s) - \alpha_n$, if $s_n = 0$.

expected payoffs $\mathbb{E}_{q_n}[\pi_n(s_n, s)]$ obtained by player n via mixed strategy q_n are given by $\mathbb{E}_{q_n}[\pi_n(s_n, s)] = \sum_{s_n \in S_n} q_n(s_n)\pi_n(s_n, s)$. In this formulation, SFBP is a discretization of an ordered externality game in the sense of Friedman [40, 41].

The one-shot strategic form SFBP is described by the tuple $\Gamma = (\mathcal{N}, (S_n, \pi_n)_{n \in \mathcal{N}}, c)$, and the infinitely repeated SFBP is given by Γ^∞. Following Foster and Young [38], a history h^t of length $t \in \mathbb{N}$ is defined to be a sequence of t outcomes drawn from the set $S = \{0, 1, 2, \ldots, N\}$;[12] the history $h^t = (s^1, \ldots, s^t)$ indicates the number of players who attended the bar during periods 1 through t. Let h^0 denote the null history, let H^t denote the set of all histories of length t, and let $H = \bigcup_0^\infty H^t$. A behavioral strategy[13] for player n is a function from the set of all possible histories to the set of mixed strategies for that player: *i.e.*, $g_n : H \to Q_n$. Now player n's play at time t is given by $q_n^t = g_n(h^{t-1})$, which is contingent on the history through time $t-1$.

A belief-based learning algorithm is a function from the set of all possible histories to the set of possible beliefs. We assume that beliefs in the repeated SFBP take the form of aggregate statistics, with a belief as a subjective probability over the space of possible externality effects $\mathcal{E} = \{\textit{undercrowded}, \textit{overcrowded}\}$.[14] The event *undercrowded* obtains whenever $s^t \leq c$; otherwise, the event *overcrowded* obtains. Let $\Delta(\mathcal{E})$ be the set of probability distributions over the set $\mathcal{E}$. Formally, a belief-based learning algorithm for player n is a function $f_n : H \to \Delta(\mathcal{E})$.[15] Since the event space $\mathcal{E}$ is of cardinality 2, the private sequence of probability distributions $\{(p_n^t, 1-p_n^t)\}$ is denoted simply $\{p_n^t\}$, where p_n^t is the probability that player n attributes to the bar being undercrowded at time t.

[12] Implicit in this notion of history is the assumption that players are anonymous: *i.e.*, they cannot distinguish other players, and moreover, they also cannot distinguish between themselves and others.

[13] A behavioral strategy determines one-shot strategies throughout the repeated game. When it is clear from context that we are referring to strategies of the one-shot game, we omit the word behavioral.

[14] Implicit in this belief structure is the assumption that beliefs are *deterministic* in the sense of Foster and Young [38]. For example, we do not consider beliefs of the form: with probability P, player n assigns probability p_n^t to the bar being undercrowded at time t, while with probability $1-P$, player n assigns probability $p_n'^t$ to the bar being undercrowded at time t instead.

[15] A belief-based learning algorithm is *not* a function $f_n : H \to Q_{-n}$, where $Q_{-n} \equiv \prod_{m \neq n} Q_m$. If it were, this would violate the assumption that players are anonymous, as players could distinguish their own play from that of the aggregate. The given definition precludes any notion of correlated beliefs, in which players might attempt to correlate the behavior of an individual, such as oneself, with attendance at the bar.

The expected payoff for player n at time t is computed in terms of the beliefs that player n holds at time t:

$$\mathbb{E}_{p_n^t}[\pi_n(s_n, s)] = \begin{cases} p_n^t \alpha_n - (1 - p_n^t)(1 - \alpha_n) & \text{if } s_n = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $p_n^* \equiv 1 - \alpha_n$. Player n is indifferent between his two pure strategies whenever $p_n^t = p_n^*$, since this implies that $\mathbb{E}_{p_n^t}[\pi_n(1, s)] = \mathbb{E}_{p_n^t}[\pi_n(0, s)] = 0$. In order to simplify notation, in what follows we often write $\pi(q_n, s^t)$ for $\mathbb{E}_{q_n}[\pi(s_n, s^t)]$ and $\pi(q_n, p_n^t)$ for $\mathbb{E}_{p_n^t}[\pi(q_n, s^t)]$. The actual probability that the bar is undercrowded at time t as determined by the players' strategies is denoted p_0^t. The existence of such objective probabilities is implied by the fact that in general players employ mixed strategies.[16]

Definition 3.2.1 SFBP is *uniform* iff for all $n, m \in \mathcal{N}$, $\alpha_n = \alpha_m \equiv \alpha$, and thus, $p_n^* = p_m^* \equiv p^*$.

3.2.1 An Example

This section formalizes an argument pertaining to the oscillatory behavior that is well-known to arise via Cournot best-reply dynamics [24] in congestion games that resemble the Santa Fe bar problem. Note that since best-reply dynamics can be viewed as "as-if" Bayesian learning, the theorem presented in this section follows as an immediate corollary of the more general results derived in later sections. We begin by reminding the reader of the assumptions implicit in best-reply dynamics.

Definition 3.2.2 A strategy $q_n^t \in Q_n$ is said to be a *best-reply* for player n at time t iff $q_n^t \in \arg\max_{q_n \in Q_n} \pi_n(q_n, p_n^t)$: *i.e.*, $\pi_n(q_n^t, p_n^t) \geq \max_{q_n \in Q_n} \pi_n(q_n, p_n^t)$.

In other words, strategy q_n^t is a best-reply for player n at time t iff it is utility maximizing. Now player n utilizes best-reply dynamics iff for all times $t+1$, player n assumes that the outcome that is realized during round t will again be the outcome of round $t+1$, and consequently plays a best-reply to the outcome of round t.

[16] We often refer to the sequence of actual probabilities $\{p_0^t\}$ as objective probabilities, but technically, this is a misnomer, since these probabilities are not truly independent of the individual players' decisions. They might rather be termed intersubjective.

Definition 3.2.3 A given player n is said to employ *best-reply dynamics* in SFBP iff for all times t, player n assumes that

$$p_n^{t+1} = \begin{cases} 1 & \text{if } s^t \leq c \\ 0 & \text{if } s^t > c \end{cases}$$

and moreover, player n plays only best-replies to these beliefs. In particular, if player n utilizes best-reply dynamics, then $q_n^t \in \arg\max_{q_n \in Q_n} \pi_n(q_n, s^t)$.

Theorem 3.2.4 *In the uniform repeated* SFBP*, best-reply dynamics do not converge:* i.e., $\forall n$, $\lim_{t,t' \to \infty} | p_n^t - p_n^{t'} | \neq 0$ *and* $\forall n$, $\lim_{t,t' \to \infty} | s^t - s^{t'} | \neq 0$.

Proof 3.2.5 Assume that all players employ best-reply dynamics. If at time t, $s^t \leq c$, then $p_n^{t+1} = 1$ for all n, to which the best response at time $t+1$ is pure strategy $s_n^{t+1} = 1$. But then $s^{t+1} > c$, so that $p_n^{t+2} = 0$ for all n, to which the best response at time $t+2$ is pure strategy $s_n^{t+2} = 0$. Now, it follows that $s^{t+2} \leq c$ and $p_n^{t+3} = 1$. This pattern repeats itself indefinitely, generating oscillatory behavior that is far from equilibrium. The argument is similar if $s^t > c$. □

Recall that SFBP falls into the class of ordered externality games. For such games, if best-reply dynamics converge, then it is known that the serially undominated set is a singleton [41, 78]. Consistent with this result, the serially undominated set obtained in the SFBP is not a singleton – on the contrary, it includes all the strategies of the game – and moreover, best-reply dynamics do not converge. Note, however, that stability results have been achieved regarding best-reply dynamics for special cases within the class of ordered externality games, some of which resemble non-uniform versions of SFBP [41].

3.2.2 A First Negative Result

This section presents a generalization of the results obtained in the previous section. It is argued that if all players are rational and if they learn according to Bayes' rule, then play does not converge to equilibrium behavior. In fact, this result is not contingent on the assumption of Bayesian learning and is readily applicable to any predictive belief-based learning mechanism in which players are rational.

Definition 3.2.6 A belief-based learning algorithm is *predictive* iff it generates a sequence of beliefs $\{p_n^t\}$ for player n *s.t.* $\lim_{t\to\infty} |\, p_n^t - p_0^t \,| = 0$.

In words, if player n utilizes a predictive learning algorithm, then the difference between player n's subjective beliefs p_n^t and the objective probabilities p_0^t converges to zero. Notice that this definition does not require that the objective probabilities themselves converge.

Definition 3.2.7 A player is *rational* iff she plays only best-replies to beliefs.

The following theorem states that in the uniform version of the repeated SFBP, whenever players exhibit rationality and predictivity, strategies converge to p^*. It follows by predictivity that beliefs converge to p^* as well. Thus, rational players who play best-replies to their beliefs, ultimately play best-replies to actual strategies: *i.e.*, play converges to Nash equilibrium. This intermediate result is later contested by noting that the assumptions of rationality and predictivity taken together yield conflicting conclusions.

Theorem 3.2.8 *In the uniform repeated* SFBP*, if players are rational and predictive, then* $\lim_{t\to\infty} |\, p_0^t - p^* | = 0$.

Proof 3.2.9 Suppose not. *Case 1:* Suppose $\exists \epsilon > 0$ *s.t.* $p_0^t > p^* + \epsilon$ infinitely often (i.o.). It follows by predictivity that for all n, $p_n^t > p^*$ i.o.. Now by rationality, all players play best-replies, which for such t is to go to the bar: *i.e.*, for all n, $s_n^t = 1$ i.o.. But this ensures that the bar will be overcrowded with probability 1, yielding $p_0^t = 0 < p^* + \epsilon < p_0^t$ i.o., which is a contradiction.

Case 2: Now suppose $\exists \epsilon > 0$ *s.t.* $p_0^t < p^* - \epsilon$ i.o.. In this case, it follows by predictivity that for all n, $p_n^t < p^*$ i.o.. Moreover, rationality implies that all players stay at home for such t: *i.e.*, for all n, $s_n^t = 0$ i.o.. But this ensures that the bar will be undercrowded with probability 1, which implies that $p_0^t = 1 > p^* - \epsilon > p_0^t$ i.o., which is again a contradiction. □

The following corollary states that whenever players are rational and predictive, beliefs converge to p^* as well as strategies. This result follows immediately from the definition of predictivity.

Corollary 3.2.10 *In the uniform* SFBP*, if players are rational and predictive, then for all* n*,* $\lim_{t\to\infty} | p_n^t - p^* | = 0$.

Proof 3.2.11 Consider an arbitrary player n. By the definition of predictivity, $\lim_{t\to\infty} | p_n^t - p_0^t | = 0$, and by Theorem 3.2.8, $\lim_{t\to\infty} | p_0^t - p^*| = 0$. Now by the triangle inequality, for all t, $| p_n^t - p^*| \leq | p_n^t - p_0^t | + | p_0^t - p^*|$. Thus, by taking limits, $\lim_{t\to\infty} | p_n^t - p^*| = 0$. □

The above theorem and corollary state that in SFBP, if players are rational and predictive, then both subjective beliefs and objective probabilities converge to p^*. It follows that rational players who play best-replies to their beliefs, ultimately play best-replies to actual strategies: *i.e.*, play converges to Nash equilibrium. The next theorem, however, states that in fact, no mechanism of rational, predictive, belief-based learning (including Bayesian updating) gives rise to objective probabilities that converge to p^*, except in unusual circumstances. As the assumptions of rationality and predictivity simultaneously give rise to conflicting conclusions, we deduce that together these assumptions are incompatible.

The next theorem constructs specific values of p^* for which beliefs and strategies do in fact converge to p^*. In these special cases, rational play indeed converges to Nash equilibrium. Before formally stating the theorem, we present an example of one such p^*. The negative results in this chapter rely on the assumption that indifferent players flip a fair coin; if, on the contrary, players were to flip a biased coin favoring one strategy or another, they would not be behaving as if they were truly indifferent.

Example 3.2.12 Assume $f(t)$ is a monotonically decreasing function of t: *e.g.*, $f(t) = 1/t$.

- Suppose G players (the optimists) hold beliefs $p^* + f(t)$. These players' beliefs converge to p^*, but by rationality, these players *always* go to the bar.
- Let H players (the pessimists) hold beliefs $p^* - f(t)$. These players' beliefs also converge to p^*, but by rationality, these players *never* go to the bar.
- Let I players (the realists) hold beliefs exactly p^* at all times t. These players are indifferent between going to the bar and not going, so they flip a fair coin.

Given that players' beliefs converge to p^*, we now consider the conditions under which players' strategies also converge to p^*. Let the excess capacity of the bar $d = c - G$ for the I indifferent players, after accomodating the G players who go to the bar in every period. Suppose indifferent players go to the bar iff their coin flips show heads. In this scenario, the probability p that the bar is undercrowded is the probability that with I flips of a fair coin,[17] at most d heads appear: *i.e.*,

$$p = \begin{cases} 0 & \text{if } d < 0 \\ 1 & \text{if } d \geq I \\ \frac{1}{2^I} \sum_{j=0}^{d} \binom{I}{j} & \text{otherwise} \end{cases} \tag{3.2}$$

Now as $t \to \infty$, $p_n^t \to p^*$ and $p_0^t \to p$. Thus, if $p^* = p$, then both beliefs and strategies converge to p^*. □

Using the layout of Example 3.2.12 and Equation 3.2, it is possible to describe all possible values of p. At fixed time t, let G denote the number of players who go to the bar; let H denote the number of players who stay at home; and let I denote the number of players who are indifferent and therefore flip a fair coin in deciding whether or not to attend the bar. The following set F describes all the realizable objective probabilities under these circumstances:

$$F = \{p \,|\, \exists\, G, H, I \in \{0, \ldots, N\} \text{ } s.t. \text{ } p \text{ is defined by Equation 3.2}\}$$

Note that F is a finite set since there are only finitely many possible values of G, H, and I. The following theorem states that objective probabilities cannot possibly converge to p^* if $p^* \notin F$.

Theorem 3.2.13 *In the uniform repeated* SFBP*, if players are rational and predictive, then* $\lim_{t\to\infty} |p_0^t - p^*| \neq 0$*, unless* $p^* \in F$*, provided indifferent players flip a fair coin.*

[17] Mathematically, this result holds in the more general case when players flip a coin of bias q. In particular, Equation 3.2 becomes

$$p = \begin{cases} 0 & \text{if } d < 0 \\ 1 & \text{if } d \geq I \\ \sum_{j=0}^{d} \binom{I}{j} q^j (1-q)^{I-j} & \text{otherwise} \end{cases} \tag{3.1}$$

We do not present this case, however, since this assumption is more difficult to justify.

Proof 3.2.14 Suppose to the contrary that $p^* \notin F$, but $\lim_{t\to\infty} |p_0^t - p^*| = 0$. Let $\delta = \min\{d(p^*, x) \,|\, x \in F\}$. Note that $\delta > 0$, since $p^* \notin F$. By the assumption of convergence, $\exists T \, s.t. \, \forall t > T, |p_0^t - p^*| < \delta$. But now, since $p_0^t \in F$, it follows that $d(p^*, F) < \delta$. Contradiction. □

In SFBP, assuming a bar of capacity c, if players are rational and predictive, then strategies can only converge to p^* if p^* happens to be an element of finite set F, but even then, it is not guaranteed unless beliefs also converge to p^*. Thus, it is only on rare occasions that players exhibit both rationality and predictivity, such that both beliefs and strategies converge to p^*: *i.e.*, play converges to Nash equilibrium. More often than not, play does not converge to Nash equilibrium in SFBP.

Example 3.2.15 Consider an instance of SFBP in which players are both rational and predictive. Let $N = I = 10$, and assume $c = 6$. In other words, there are 10 players, all of whom are indifferent and flip a fair coin. Now according to Equation 3.2, if $p^* \approx .828$, and if in addition beliefs converge to p^*, then strategies also converge to p^*. This implies that players are playing best-replies to actual strategies. Thus, in this particular instance of SFBP, if by chance $\alpha = 1 - p^* \approx .172$, and if beliefs indeed converge to p^*, then play converges to Nash equilibrium. □

As Theorems 3.2.8 and 3.2.13 yield contradictory conclusions in all but finitely many cases, the next corollary states that their assumptions are more often than not inconsistent. In particular, *there is no rational learning in* SFBP.

Corollary 3.2.16 *In the uniform repeated* SFBP*, players cannot be both rational and predictive, unless* $p^* \in F$.

This concludes the discussion of our first negative result in SFBP. It was argued that two conditions which together are sufficient for convergence to Nash equilibrium, namely rationality and predictivity, are incompatible in SFBP. In the next section, a second negative result is derived; similar analyses appeared in Greenwald, *et al.* [52] and Mishra [79].

3.2.3 A Second Negative Result

This section describes a second negative result which is based on the notion of strong predictivity. Strongly predictive learning yields instances of SFBP in which there are no optimists and no pessimists; all players are realists who flip a fair coin.

Definition 3.2.17 A belief-based algorithm is said to be *strongly predictive* iff it generates a sequence of beliefs $\{p_n^t\}$ *s.t.* $p_n^t = p_0^t$ almost always (a.a.).

The next theorem states that no means of rational, strongly predictive learning gives rise to objective probabilities that equal p^*, unless the capacity of the bar fortuitously lies between $N/2 - k_1\sqrt{N}$ and $N/2 + k_2\sqrt{N}$, for certain $k_1, k_2 > 0$. This result is explained intuitively as follows. Assume players are strongly predictive; then subjective probabilities equal objective probabilities a.a.. By reasoning that is analogous to Theorem 3.2.8, objective probabilities equal p^* a.a.; it follows that subjective probabilities equal p^* a.a.. This implies that the players are indifferent a.a., so by assumption, they flip a fair coin. Thus, attendance at the bar is likely to be near $N/2$ a.a.. The theorem states that unless the capacity of the bar is near $N/2$, rational, strongly predictive learning is ineffectual in SFBP.

Theorem 3.2.18 *In the uniform repeated* SFBP, *given* $0 < \alpha < 1$, *if players are rational and strongly predictive, then* $p_0^t \neq p^*$ *a.a., provided that indifferent players flip a fair coin*[18] *and that the capacity of the bar* $c \leq N/2 - \sqrt{[1 + \ln(1/(1-\alpha))]N}$ *or* $c \geq N/2 + \sqrt{[3 + 3\ln(1/\alpha)][N/2]}$.

Proof 3.2.19 Suppose not: *i.e.*, suppose that $p_0^t = p^*$ i.o.. Since the players are strongly predictive learners, for all n, $p_n^t = p_0^t$ a.a.. Together these statements imply $p_n^t = p^*$ i.o. At such times t, rational players are indifferent; by assumption, they flip a fair coin. It follows that attendance at the bar is binomially distributed $\sim S(N, 1/2)$. Two distinct cases arise, depending on the capacity of the bar. This proof utilizes the multiplicative variant of the Chernoff bound [23].

[18] Mathematically, this result holds for arbitrary probabilities p_n, p_m *s.t.* $|p_n - p_m| < \delta$, for small values of $\delta > 0$, where p_n and p_m denote the probabilities that players n and m, respectively, go to the bar. We do not present this case, however, since this assumption is more difficult to justify.

Case 3.2.19.1 Assume that $c \leq N/2 - \sqrt{[1 + \ln(1/(1-\alpha))]N}$. In this case,

$$\begin{aligned} p_0^t &= \Pr[S(N,1/2) < c] \\ &\leq \Pr[S(N,1/2) < \{1 - \sqrt{[4 + 4\ln(1/(1-\alpha))]/N}\}(N/2)] \\ &\leq e^{-([4+4\ln(1/(1-\alpha))]/2N)(N/2)} \\ &= (1-\alpha)/e \\ &< 1 - \alpha \end{aligned}$$

Therefore, $p_0^t < 1 - \alpha = p^*$. Contradiction.

Case 3.2.19.2 Assume that $c \geq N/2 + \sqrt{[3 + 3\ln(1/\alpha)][N/2]}$. In this case,

$$\begin{aligned} 1 - p_0^t &= \Pr[S(N,1/2) > c] \\ &\leq \Pr[S(N,1/2) > \{1 + \sqrt{[6 + 6\ln(1/\alpha)]/N}\}(N/2)] \\ &\leq e^{-([6+6\ln(1/\alpha)]/3N)(N/2)} \\ &= \alpha/e \\ &< \alpha \end{aligned}$$

Therefore, $p_0^t > 1 - \alpha = p^*$. Contradiction. □

This concludes the discussion of negative results on rational learning in SFBP. The remainder of this chapter focuses on positive results in SFBP that are obtained via low-rationality learning. While the discussion thus far has been purely theoretical, that which follows is simulation-based.

3.3 Practical Investigations

In the second half of this chapter, we study learning among computational agents who are not highly rational; instead, they exhibit low-rationality learning, which is discussed below. We present positive simulation results for which computational learning based on low-rationality yields convergence to equilibrium behavior in SFBP. These results are of interest because they demonstrate that learning in repeated games affords solutions to problems of resource allocation in decentralized environments.

The Santa Fe bar paradigm is applicable in a wide range of practical areas, such as network control and optimization and financial management. For example, SFBP is analogous to a network flow control problem which software agents might face in deciding whether or not to transmit data over a given communication link. If all the agents believe that current network delays are minor, then all the agents might decide to transmit simultaneously, causing congestion to in fact be major; but now, if the agents believe congestion delays are substantial, then all agents might decide *not* to transmit, causing congestion to once again be minimal; and so on. Similarly, SFBP also parallels an investment scenario which automated trading agents might face in deciding whether or not to buy a certain security. If the market price is low, then all the agents might decide to buy, but this increase in demand in turn causes an increase in market price; now if the market price is high, then all the agents might decide to sell, but this increase in supply causes the market price to fall once again. This pattern repeats itself indefinitely in this naive implementation of computational investors.

An interesting extension of SFBP, dubbed the *New York City bar problem* [33, 52] (NYCBP), considers this problem in a city with many bars. In this case, the networking analog is a routing problem which concerns the choice of route by which to transmit a fixed amount of data so as to minimize congestion. In financial terms, this problem corresponds to the management of an investment portfolio. This section describes simulations of both the Santa Fe bar problem in its original form, and the New York city bar problem in the case of two bars.

3.3.1 Learning Algorithms

The simulation experiments discussed in this section were conducted using low-rationality (*i.e.*, non-Bayesian) learning algorithms. According to these algorithms, players do not maintain belief-based models over the space of opponents' strategies or payoff structures. Instead, these algorithms specify that players *explore* their own strategy space by playing all strategies with some non-zero probability, and *exploit* successful strategies by increasing the probability of employing those strategies that generate high payoffs. Simple reinforcement techniques such as those utilized in this

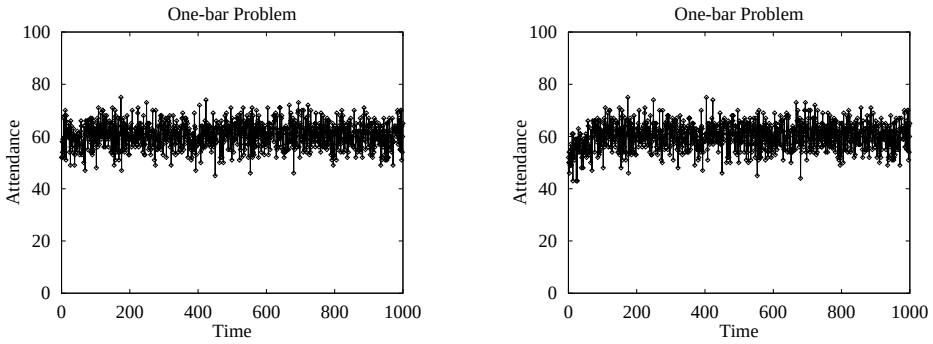
thesis are advantageous because unlike Bayesian learning, they do not depend on any complex modeling of prior probabilities over possible states of the world, and unlike the model of bounded rationality originally introduced by Arthur in his work on SFBP, they do not depend on non-deterministic beliefs. It is the simplicity of these learning algorithms that makes them potentially suitable for automated network control.

Specifically, the learning algorithms simulated in this chapter include the additive updating procedure described by Roth and Erev [30] and a simple variant introduced in Chapter 3, as well as two related multiplicative updating procedures due to Freund and Schapire [39] and Auer, Cesa-Bianchi, Freund, and Schapire [3]. These learning algorithms are simulated in two contexts. In the first, so-called *informed* settings, complete information is available regarding payoffs, including the surmised payoffs of strategies which are not played; in the second, so-called *naive* settings, the only information pertaining to payoff functions that is available at a given time is the payoff of the strategy that is in fact employed at that time. We now describe simulations of SFBP (one-bar problem) and NYCBP (two-bar problem) in informed and naive settings.

3.3.2 One-bar Problem

The one-bar problem was simulated assuming 100 computational agents and a single bar of capacity 60. Figure 3.1 depicts the results of simulations of both the additive and the multiplicative updating schemes by plotting attendance at the bar over time. (For simulation purposes, $\alpha = .5$ and $\beta = .01$.) Note that attendance centers around 60, which is in fact the capacity of the bar. Specifically, the mean is 60.51 and the variance is 4.91 for the additive updating algorithm, while the mean is 60.04 and the variance is 5.11 for the multiplicative updating algorithm. These learning algorithms, which do not necessitate perfectly rational behavior, yield equilibrium outcomes. These results are robust in that they hold regardless of the capacity of the bar. Figure 3.2 depicts attendance at the bar over time when the capacity of the bar begins at 60, but at time 1000 it changes to 40, and at time 2000 it changes to 80. In this scenario, attendance at the bar fluctuates accordingly[19] (see Table 3.1).

[19] In fact, this result was obtained for additive updating using a responsive variant of the algorithm (see [50]) in the spirit of Roth and Erev [30].



(a) Additive Updating (b) Multiplicative Updating

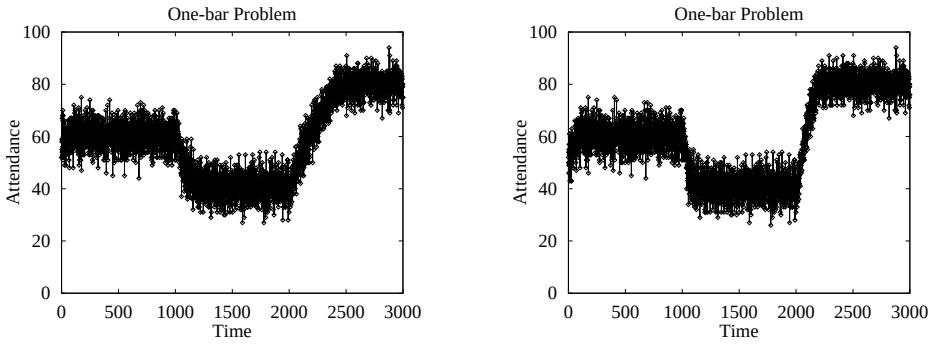
Figure 3.1: Attendance vs. Time in One-bar Problem

	Additive Updating		*Multiplicative Updating*	
Varied Capacity	Mean	Variance	Mean	Variance
Time 1 – 1000	60.51	4.91	60.04	5.11
Time 1001 – 2000	42.24	6.46	41.53	6.00
Time 1501 – 2000	40.50	4.90	40.53	4.92
Time 2000 – 3000	72.12	12.37	76.66	10.03
Time 2501 – 3000	80.02	16.48	80.16	4.04

Table 3.1: Mean and Variance for One-Bar Problem: Varied Capacity

Careful inspection of the mixed strategies which the agents employ reveals that the additive and multiplicative algorithms converge to (at least) δ-Nash equilibrium (see, for example, [47]) in the one-bar problem, and perhaps precisely to Nash equilibrium. In particular, when the capacity of the bar is fixed at 60, additive updating yields mixed strategies ranging from $(.397, .603)$ to $(.403, .597)$ (*i.e.*, $\delta = .003$), while the multiplicative updating algorithm yields mixed strategies ranging from $(.371, .629)$ to $(.423, .577)$ (*i.e.*, $\delta = .029$).[20] These algorithms generate a fair solution to the one-bar problem in which on average, all agents attend the bar 60% of the time.

[20] Note that in the multiplicative updating algorithm, the values of these extrema are directly correlated to the choice of β: *e.g.*, for $\beta = .001$, the mixed strategies range from $(.392, .608)$ to $(.404, .596)$ (*i.e.*, $\delta = .008$).

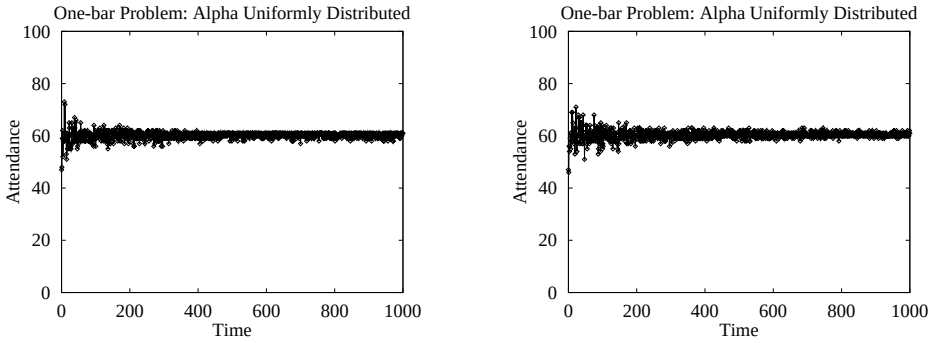


(a) Additive Updating (b) Multiplicative Updating

Figure 3.2: Attendance vs. Time in One-bar Problem: Varied Capacity

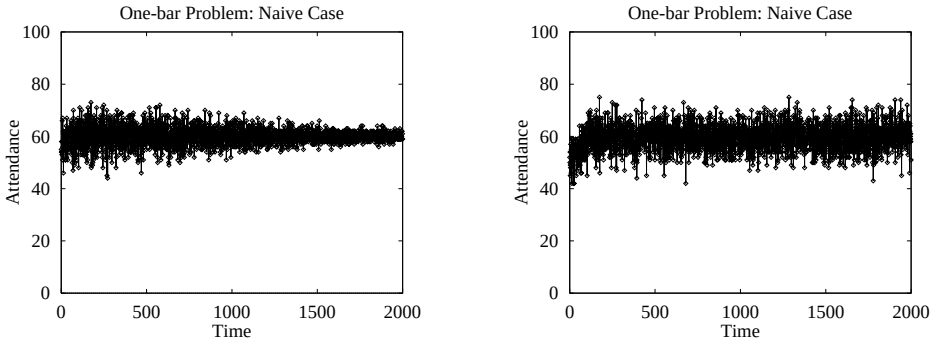
In addition to simulating the uniform Santa Fe bar game, with $\alpha = .5$ for all agents, the non-uniform problem was also simulated with the values of α_n uniformly distributed in the range $[0, 1]$. Figure 3.3 depicts the results of these simulations. In this scenario attendance once again centers around the capacity of the bar; moreover, we observe substantially more stability. Specifically, for additive updating, the mean is 60.09 and the variance is 1.59, while the mean is 60.37 and the variance is 1.32 for multiplicative updating. This is a result of the fact that those agents with small values of α learn not to attend the bar at all, while those agents with large values α learn to always attend the bar, and only those on the cusp vary between sometimes going to the bar and sometimes staying at home. In fact, theoretical results on stability have been obtained for best-reply dynamics in related non-atomic, non-uniform games [40].

The last set of one-bar simulations considers the problem in a naive setting. The agents learned via naive versions of the various algorithms (with $\epsilon = .02.$), and they were provided with incomplete information: they were informed of the attendance at the bar only if they themselves attended. Naive algorithms approximate informed algorithms such that play converges to δ-Nash equilibrium in an informed setting if and only if play converges to $(\delta + \epsilon)$-Nash equilibrium in the corresponding naive case. Figure 3.4 (a) depicts the results of learning via naive multiplicative updating for $\beta = .01$ for which attendance once again centers around the capacity of the bar. (See Table 3.2 for precise values of the mean and variance.)



(a) Additive Updating (b) Multiplicative Updating

Figure 3.3: Attendance vs. Time in One-bar Problem: α_n Uniformly Distributed



(a) Multiplicative Updating ($\beta = 0.01$ (b) Multiplicative Updating ($\beta = 0.1$)

Figure 3.4: Attendance vs. Time in One-bar Problem: Naive Case

Naive Setting	Mean	Variance
$\beta = 0.01$	59.96	2.95
$\beta = 0.1$	59.66	5.00

Table 3.2: Mean and Variance for One-Bar Problem: Multiplicative Updating, Naive Setting

Finally, it is interesting to note that in naive settings, for alternative choices of β in the multiplicative updating algorithm, it is possible to decrease variance and thereby increase stability (see Table 3.2). In particular, if the agents' rates of learning are increased, learning becomes less accurate, and some number of (say m) agents stop attending the bar while (say l) others settle on always attending the bar. The remaining $N - m - l$ agents consequently adjust their behavior as if the total population were $N - m$ and the capacity of the bar were $c - l$, which yields lower variance in the attendance (see Figure 3.4 (b) where $\beta = 0.1$). After a while, however, since the agents are continually experimenting, agents who were once fixed on a certain strategy learn that their payoffs could be improved by altering their strategy, and they do so. Thus, in the long-run, all agents attend the bar an equivalent number of times, as in the informed setting.

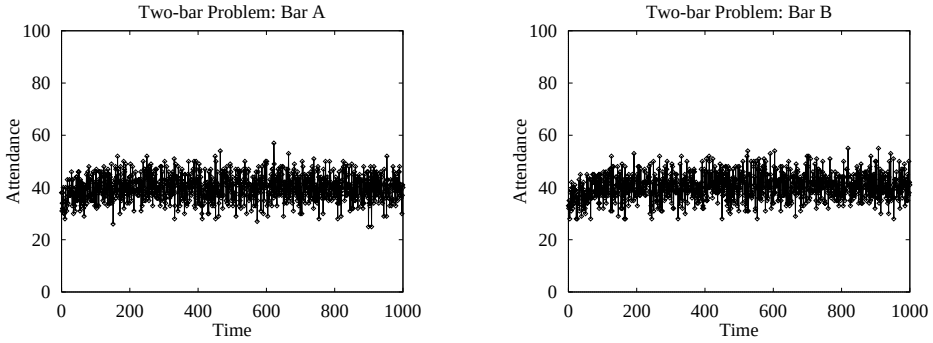
3.3.3 Two-bar Problem

In this section, we discuss simulations of the two-bar problem. We show that in spite of the additional complexity of this problem, the equilibrium behavior observed in the one-bar problem extends to the two-bar problem because of the robust nature of the computational learning algorithms. The first scenario which we consider is analogous to the one-bar problem; in particular, there is excess demand. Figure 3.5 depicts the attendance at two bars, say bar A and bar B, each of capacity 40, with a population of 100 agents who learn according to the informed version of multiplicative updating.[21] Note that attendance in each of the bars centers around 40, with about 20 agents choosing to stay at home each round. This is once again a fair solution in that each agent attends each bar 40% of the time and stays at home 20% of the time.

In the two-bar problem, in addition to the study of excess demand, it is also of interest to consider the case of excess supply: *e.g.*, two bars, each of capacity 60, and a population of 100 agents. Figure 3.6 depicts the results of simulation of this scenario where agents learn according to the informed version of additive updating. In this case, agents learn to play mixed strategies of approximately $(1/2, 1/2)$, and

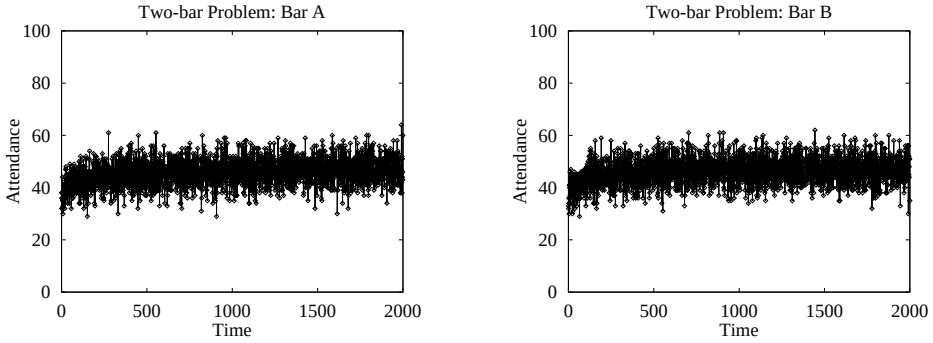
[21] No qualitative differences between additive and multiplicative updating were observed in simulations of the two-bar problem. Consequently, we present results in this section which depict attendance at two bars, rather than attendance resulting from two algorithms.

consequently each agent attends each bar approximately 50% of the time. The space of equilibria, however, in this instantiation of the two-bar problem with excess supply, ranges from 40 agents at bar A and 60 agents at bar B to 60 agents at bar A and 40 agents at bar B.



(a) Multiplicative Updating: Bar A (b) Multiplicative Updating: Bar B

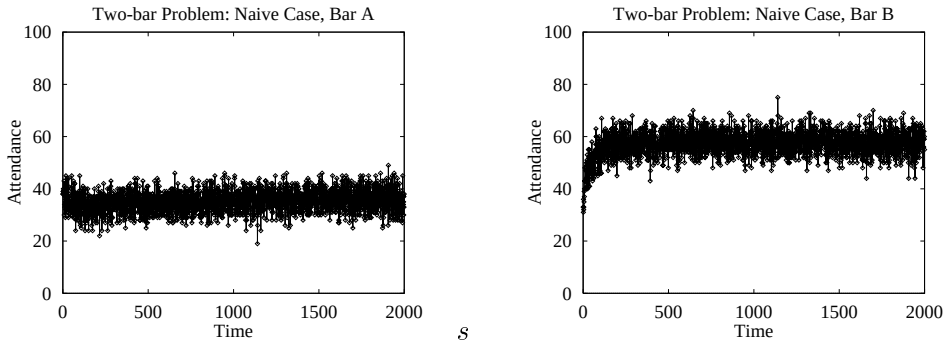
Figure 3.5: Attendance vs. Time in the Two-bar Problem: Excess Demand



(a) Additive Updating: Bar A (b) Additive Updating: Bar B

Figure 3.6: Attendance vs. Time in the Two-bar Problem: Excess Supply

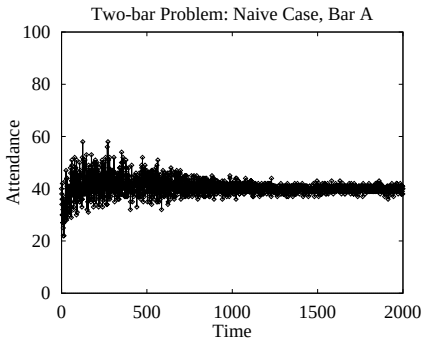
In naive settings, equilibria other than the symmetric outcome which is achieved in informed settings can persist. Figure 3.7 depicts the results of simulations of naive additive learners who experiment until they arrive at a situation in which they appear satisfied, namely 40 regular attendees at bar A, and 60 regular attendees at bar B, and there they seem to remain. In fact, the population at bar B is slightly declining over time, while the attendance at bar A is correspondingly increasing ever so slightly, because of the downward pressure which exists when attendance at bar B exceeds capacity. It is likely that attendance at each bar would ultimately settle close to 50, but this warrants further investigation.



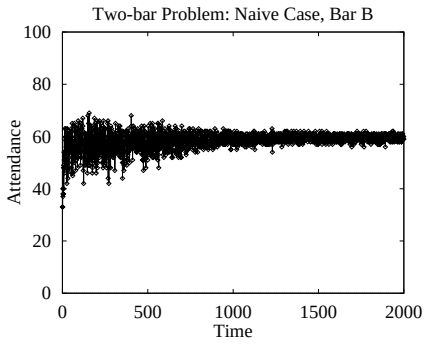
(a) Additive Updating: Bar A (b) Additive Updating: Bar B

Figure 3.7: Attendance vs. Time in the Two-bar Problem: Naive Case

Finally, in the case of excess supply, naive multiplicative learners who learn quickly ($\beta = 0.1$) arrive at a stable solution in which the variance is quite low (see Figure 3.8). In contrast with the observed behavior that results via learning according to the additive updating algorithm, naive multiplicative learners experiment until they arrive at a situation in which they are satisfied, namely 40 regular attendees at bar A, and 60 regular attendees at bar B, and there they remain. In this case, it seems less likely that further iterations will ultimately lead to equal attendance in both bars, since the lower variance eliminates some of the downward pressure which was apparent in Figure 3.7. The precise values of the variances for the various algorithms in simulations of the two-bar problem appear in Table 3.3.



(a) Multiplicative ($\beta = 0.1$): Bar A



(b) Multiplicative ($\beta = 0.1$): Bar B

Figure 3.8: Attendance vs. Time in the Two-bar Problem: Naive Case

	Bar A		*Bar B*	
Two-Bar Problem	Mean	Variance	Mean	Variance
Excess Supply	46.86	5.13	47.17	5.12
Excess Demand	40.03	4.94	40.33	4.97
Naive: Additive	36.32	4.19	57.44	4.44
Naive: Multiplicative	40.56	3.17	58.36	3.34

Table 3.3: Mean and Variance for Two-Bar Problem

3.4 Conclusion

The first half of this chapter demonstrated that it is inconsistent to conclude that belief-based learning among rational players in the Santa Fe bar problem exhibits properties of learning. In light of this negative result, it appeared necessary to relax the assumption of rationality in order that learning might converge to equilibrium behavior, as Arthur did in his original work on this problem, in the study of bounded rationality, and inductive learning. In the second half of this chapter, we simulated low-rationality learning behavior, and we found that such learning indeed yields stable outcomes in SFBP. This work aids in our understanding of the process by which agents collectively learn through repeated interactions, particularly in the non-stationary environments that prevail when multiple agents endeavor to learn simultaneously.

Chapter 4

Network Experiments

4.1 Introduction

While much of classical game theory relies on the assumption of common knowledge, there are many contexts in which this assumption does not apply. Correspondingly, in recent years there has been increased interest in the process by which a set of initially naive agents *learn* through repeated play of a game. The central question concerns the nature of asymptotic play; what set of strategies do the agents learn to play in the long-time limit? (The recent review by Fudenberg and Levine [46] provides an overview of the literature.)

In this chapter we focus our attention on learning that occurs in what we call a *network context*. In network contexts, agents interact through the common use of a resource, such as a communication link or a shared database, which is accessed over a network. The interactions of Internet congestion control algorithms where agents share network bandwidth, as described in [100], is perhaps the most studied example of a repeated game in a network context. As the Internet continues to grow, and more resources are shared by remote users, we expect the network context to become increasingly common. As discussed in [43], the network context differs from the traditional game-theoretic context in four important ways.

I. First, agents have very limited *a priori* information. In general, agents are not aware of the underlying characteristics of the shared resource. In other words, they do not know the payoff structure of the game; they know neither their own payoff

function, nor that of the other players. In addition, agents are not explicitly aware of the existence of other players. While agents are aware that there may be other agents simultaneously using the shared resource, they do not have any way of directly observing their presence and, consequently, they know nothing about the number or the characteristics of their opponents. In particular, this implies that players *cannot* observe the actions of other players, a standard assumption in many classical models of learning in economics.

II. Second, the payoff function and the agent population are subject to change over time. Shared resources like network links and computer servers periodically crash, and often experience other unpredictable changes in their capabilities, such as upgrades or route changes. In addition, users of these shared resources come and go quite frequently. Thus, when an agent detects a change in his payoff while keeping his own strategy fixed, the agent cannot tell whether this change is due to changing strategies of the other players, changes in the players themselves, or variations in the characteristics of the shared resource.

III. Third, in many cases, learning is actually carried out by an automated agent, rather than a human user. For instance, congestion control algorithms (*e.g.*, TCP) embedded in a computer's operating system control the sharing of network links. Similarly, automated algorithms can control the retry behavior for query submission to a database. Consequently, the learning that takes place in these contexts is specified in the form of a well-defined algorithm. Moreover, these algorithms are intended to be quite general in nature, and do not depend on the detailed specifics of any particular situation. In particular, this means that Bayesian learning algorithms are inappropriate, because the initial beliefs depend on the specific context. In any event, the complexity of prior probabilities is such that it is not possible to use Bayesian updating in any realistic network setting.

IV. Fourth, in network contexts, games can be played in an asynchronous fashion. There need not be any notion of definable "rounds of play"; users can update their strategies at any time. Moreover, the rates at which agents update their strategies can vary widely, although in general these rates are determined by circumstances and are not a strategic variable. Due to the geographic dispersion of users of the Internet, for example, there can be varying communication delays to a shared resource, which

in turn can lead to updating rates that differ by several orders of magnitude.[1] In addition, automated agents can learn at very different rates, depending on processor speeds and the nature of the learning algorithms. Thus, asynchrony does not arise from Stackelbergian-type strategic manipulation, but rather from inherent properties of communication and computation. Agents closer to the shared resource or those who have faster processors have the potential to learn more rapidly and effectively.

We focus on contexts that have these four properties: low information content, non-stationary payoffs, automated learning, and asynchrony. We are interested in what happens when a set of automated agents play a game repeatedly in such a context, and we investigate this behavior empirically. We consider a small sampling of learning algorithms, some of which have been well-studied in the literature; for each algorithm we numerically simulate a set of agents using that algorithm and we observe the set of strategies played in the long-time regime. Our simulation experiments can be seen as a natural counterpart to human economic experiments; in particular, Chen [21] investigates some issues closely related to those considered here using human subjects rather than automated learning algorithms.[2]

We concentrate on the extent to which the asymptotic play depends on the amount of information available to the agents, the degree of responsiveness of the learning algorithm, and the level of asynchrony of play. Of particular interest is the extent to which the asymptotic play is contained in the various solution concepts including Nash equilibria, the set of serially undominated strategies (D^∞), and less traditional concepts such as serially unoverwhelmed strategies (O^∞) and serially Stackelberg-undominated strategies (S^∞) which are discussed below. Our findings suggest that the asymptotic play of games in network contexts can be quite different from that in standard contexts, where play is typically contained within D^∞. These results have important implications for the networking community, where it is often assumed that players are rational and that the network operating point is Nash equilibrium (see, for example, Shenker [100]).

[1] As discussed in [43], standard control theoretic results imply that the frequency at which strategies are updated should not be greater than the inverse of the round-trip communication delay to the shared resource; otherwise, instability may result.

[2] Although our present focus is solely on automated agents, experimental evidence (see [21], [30], and [80]) suggests that our results are also relevant in describing human/human and human/machine interactions.

4.1.1 Learning Algorithms

The literature is replete with learning algorithms, but not all of them are applicable in network contexts. Because knowledge of the payoff structure and the other agents is extremely limited, games in a network context are, from a single agent's perspective, most naturally modeled as *games against nature* in which each strategy has some random (and possibly time-varying) payoff about which the agent has no *a priori* knowledge. Consequently, in contrast with belief-based approaches to learning (*e.g.*, Bayesian updating) adopted in much of the literature, learning algorithms for network contexts typically utilize simple updating schemes that do not rely on any detailed assumptions about the structure of the game. Instead, these algorithms employ "trial-and-error" experimentation in an attempt to identify optimal strategies: *i.e.*, these algorithms seek to optimize given the trade-off between exploration and exploitation.

The learning algorithms which we simulate are distinguished first of all by their varying degrees of experimentation; for convenience, we denote by parameter $\epsilon \in [0, 1]$ this level of experimentation. In static environments, where the payoff structure and the set and characteristics of the other agents is fixed, it may be reasonable to decrease the level of experimentation over time, with experimentation ceasing in the infinite-time limit (*i.e.*, $\epsilon \to 0$ as $t \to \infty$). [3] Many learning algorithms proposed in the literature have this property. In network contexts, however, the environment is not static; the underlying payoff structure, and the population of agents, are subject to change at any time without explicit notification. As a result, agents should be prepared to respond to changing conditions at all times, and should do so in a bounded amount of time. This requires that a non-zero level of experimentation be maintained in the long-time limit, and that future play be more heavily influenced by payoffs obtained in the recent, rather than the distant, past. This second point can be achieved via a parameter $\gamma \in (0, 1]$ which dictates the rate (and typically inverse accuracy) of learning. We call the ability to respond to changes in the environment in bounded time *responsiveness*, and posit that this property is fundamental to learning in network contexts. As we shall see, responsiveness has important implications for the resulting asymptotic play.

[3] This is apparent in decision problems such as classic bandit problems [48, 65].

The learning algorithms we discuss also differ in the particular criteria that they are designed to satisfy. Perhaps the simplest criterion is that, when playing a static game-against-nature, the algorithm rapidly learns to play (with high probability) the strategy with the highest average payoff.[4] When combined with responsiveness, and a certain monotonicity property, this leads to the class of so-called *reasonable* learning algorithms introduced in [43]. One example of such an algorithm is the *stage* learning algorithm. Stage learners partition a repeated game into *stages*, which consist of $1/\gamma$ rounds of a game. At each round of play, a stage learner chooses its strategy at random based on the probabilities, or weights, it has assigned to each of its strategies. Weights are updated upon termination of each stage, with weight $1-\epsilon$ assigned to the pure strategy that obtained the highest average payoffs during the previous stage, and weight $\epsilon/(n-1)$ assigned to all other strategies. Another example of a reasonable learning algorithm, so-called *responsive learning automata* introduced in [42], is a responsive version of simple learning automata (see, for example, Narendra and Thathachar [82]). This algorithm updates weights after every round of play using quite a different method. Another reasonable learning algorithm (for certain choices of parameters) is defined by Roth and Erev [30], and has been used to model human behavior in game-theoretic experiments.

A second criterion, which is a worst-case measure of performance, involves the concept of *regret*. Intuitively, a sequence of plays is optimal if there is no regret for playing a given strategy sequence rather than playing another possible sequence of strategies. Regret comes in two forms: external and internal. A sequence of plays is said to exhibit no external regret if the difference between the cumulative payoffs that are achieved by the learner and those that could be achieved by any other pure strategy is insignificant. The no internal regret optimality criterion is a refinement of the no external regret criterion where the difference between the performance of a learner's strategies and any *remapped* sequence of those strategies is insignificant. By remapped we mean that there is a mapping f of the strategy space into itself such

[4] The formal definition of probabilistic converge in finite time is described in [43]. In this paper we do not formally define convergence, but take a more pragmatic approach which is appropriate for simulations. That is, we say that play has converged when the numerical properties are unchanged by additional iterations as evidenced by simulations.

that for every occurrence of a given strategy s in the original sequence the mapped strategy $f(s)$ appears in the remapped sequence of strategies. The learning procedures described in Foster and Vohra [37] and Hart and Mas-Colell [57] satisfy the property of no internal regret. Some early no external regret algorithms were discovered by Blackwell [14], Hannan [53], and Banos [10], and Megiddo [77]; recently, no external regret algorithms appeared in Cover [25], Freund and Schapire [39], and Auer, Cesa-Bianchi, Freund and Schapire [3].

We investigate six learning algorithms: the reasonable learners discussed above (see [42, 43, 30]), two based on external regret (see [3, 36]), and one which exhibits no internal regret (see [57]). Some of these algorithms were initially proposed for quite different settings, in which responsiveness is not necessary and the level of information is significantly higher (*e.g.*, agents know their own payoff function). We have extended these learning algorithms for use in network contexts. We call the versions designed for low-information settings *naive*, and those designed for higher information contexts *informed*. We also consider both *responsive* and *non-responsive* variants. Lastly, each agent has a time-scale parameter A that determines the rate at which it updates its strategies. A player updates its strategy (*i.e.*, runs its learning algorithm) only every A rounds, and treats the average payoff during those A rounds as its actual payoff. We are interested in how the asymptotic play depends on whether agents are responsive, whether they are informed, and on the degree of asynchrony (differences in the A values) among the agents.

4.1.2 Solution Concepts

In order to describe the asymptotic play, we introduce several solution concepts. We begin with some notation. Throughout this chapter, our attention is restricted to finite games. Let $\mathcal{N} = \{1, \ldots, N\}$ be a finite set of *players*, where $N \in \mathcal{N}$ is the number of players. The finite set of strategies available to player $i \in \mathcal{N}$ is denoted by S_i, with element $s_i \in S_i$. The set of pure strategy profiles is the Cartesian product $S = \prod_i S_i$. In addition, let $S_{-i} = \prod_{j \neq i} S_j$ with element $s_{-i} \in S_{-i}$, and write $s = (s_i, s_{-i}) \in S$. Finally, the payoff function $\pi_i : S \to \mathbb{R}$ for player i is a real-valued function defined on S.

Recall that strategy $s_i \in S_i$ is strictly dominated for player i if there exists some strategy $s_i^* \in S_i$ such that $\pi_i(s_i, s_{-i}) < \pi_i(s_i^*, s_{-i})$ for all $s_{-i} \in S_{-i}$. Let D^∞ denote the serially undominated strategy set: *i.e.*, the set of strategies that remains after the iterated elimination of strictly dominated strategies. Milgrom and Roberts [78] show that the asymptotic play of a set of *adaptive* learners – learners that eventually learn to play only undominated strategies – eventually lies within D^∞. In addition, it is shown in [42] that certain responsive learners playing synchronously also converge to D^∞. The set D^∞ is widely considered to be an upper bound in terms of solution concepts; that is, it is commonly held that the appropriate solution concept that arises via learning through repeated play is a subset of the serially undominated set.[5] This may indeed be true in standard game-theoretic contexts.

In [42, 43], however, it is shown that in network contexts, where there is the potential for asynchrony and responsive learning, play can asymptotically remain outside the serially undominated set. A more appropriate solution concept for such settings is based on the concept of *overwhelmed* strategies. We say that strategy $s_i \in S_i$ is strictly overwhelmed if there exists some other strategy $s_i^* \in S_i$ such that $\pi_i(s_i, s_{-i}) < \pi_i(s_i^*, s'_{-i})$ for all $s_{-i}, s'_{-i} \in S_{-i}$. Let O^∞ denote the set of strategies that remains after the iterated elimination of strictly overwhelmed strategies. It is shown in [43] that the asymptotic play of a set of reasonable learners lies within O^∞, regardless of the level of asynchrony. However, it is conjectured that O^∞ is not a precise solution concept, only an upper bound.

A refinement of O^∞, called S^∞ is defined in [43]. Because it is rather cumbersome, we do not present the precise definition of S^∞, but here is some intuition. The set S^∞ extends the set D^∞ by allowing for the possibility that play is asynchronous, rather than synchronous as is standard in repeated game theory. In particular, dominated strategies are iteratively deleted, assuming all possible orderings among player moves. In two player games, for example, this amounts to three orderings, those in which each of the two players plays the role of leader, while the other follows, as well as the ordering in which both players move simultaneously. More formally, the computation of S^∞ is as follows: pick a specific (non-strict) ordering of the players

[5] Note that this also holds for one-shot games with common knowledge, as the set D^∞ contains all the rationalizable strategies [11, 85].

and construct the extensive form game arising from players moving according to that order; now compute the set of actions which survive the iterated deletion of strictly dominated strategies in the new game; finally, take the union of these actions for all orderings. Since the selected ordering is non-strict, asynchronicity defined in this way incorporates synchronicity, from which it follows that $D^\infty \subseteq S^\infty \subseteq O^\infty$.

Another result of great interest, due to Foster and Vohra [37], is that a set of no internal regret learners converges to a correlated equilibrium. Note that the support of a set of correlated equilibria is a subset of D^∞; in other words, correlated equilibria do not assign positive probabilities to strategies outside D^∞, but neither do they necessarily converge to Nash equilibria. In contrast, the asymptotic play of a set of no external regret learners need not remain inside D^∞, as is shown in Chapter 2. Note that this remark pertains only to the no external regret criterion, but says nothing about the convergence properties of specific algorithms which are defined to satisfy this criterion, such as those considered in this study.

In the remainder of this paper, we present the result of simulations of the six learning algorithms on various games. We ask whether the asymptotic play of these games in network contexts converges within the sets D^∞, S^∞, or O^∞. Recall that the term convergence is used informally, both because of experimentation, which precludes true convergence, and our interest in achieving results in finite time. We are interested in determining which of these concepts, if any, represents an appropriate solution concept for games in network contexts.

4.2 Simulations in Network Contexts

We consider three sets of games: simple games (two players, two or three strategies), the congestion game (two players, many strategies), and an externality game (many players, two strategies). The simulations were conducted with varying degrees of asynchrony, ranging from synchronous play to extreme asynchrony with one player acting as the leader (*i.e.*, we vary the value of A from 1 to 10,000 for the leading player and we set $A = 1$ for all other players). The degree of responsiveness is determined by parameters ϵ and γ.

4.2.1 Simple Two-Player Games

This subsection presents the results of simulations of four simple two-player games with either two or three strategies per player. The row player is taken to be the leader. The parameter γ was set to .01 for all algorithms, while the degree of experimentation ϵ was set to .025 for the reasonable learning algorithms and .05 for the no-regret algorithms.[6] In addition, the no-regret algorithms depend on tuning parameters; for the mixing method, $\alpha = 100$, for multiplicative updating, $\beta = 1$, and for the no internal regret algorithm, $\kappa = 2$.[7] Unless otherwise stated, the simulations described in this chapter were run for 10^8 iterations.[8] Initially, all strategies were assigned equal weights.

Game D

The game depicted in Figure 4.1 is referred to as Game D since it is D-solvable, but it is not S-solvable or O-solvable: *i.e.*, $D^\infty \neq S^\infty = O^\infty$. More specifically, the set D^∞ is a singleton that contains only the strategy profile (T, L), which is the unique Nash equilibrium. On the other hand, $S^\infty = O^\infty = \{T, B\} \times \{L, R\}$. Note that (B, R) is a Stackelberg equilibrium in which the row player is the leader.

The graph depicted in Figure 4.2 describes the overall results of simulations of Game D, assuming responsive learning in a naive setting. In particular, Figure 4.2 (a) plots the percentage of time in which the Nash equilibrium solution arises as the degree of asynchrony varies. Asynchrony of 100, for example, implies that the column player is learning 100 times as fast as the column player; thus, the row player is viewed as the leader and the column player the follower. Notice that when play is synchronous, all the algorithms converge to the unique Nash solution. However, in

[6] The choice of parameter values reflects the trade-off between exploration and exploitation. Increasing the rate of responsiveness γ and the rate of experimentation ϵ leads to increased error in stationary environments, but increased accuracy in non-stationary environments. In our experience, the results are fairly robust to small changes in parameter settings, although we have not formally measured this robustness.

[7] These parameters represent the learning rates in the no regret algorithms. As usual, slower learning rates correspond to higher degrees of accuracy in stationary environments; in non-stationary environments, faster learning rates induce more responsive behavior.

[8] Although play generally converged in far fewer iterations, this rather lengthy simulation time eliminated the transient effects of initial conditions in the final long-run empirical frequency calculations.

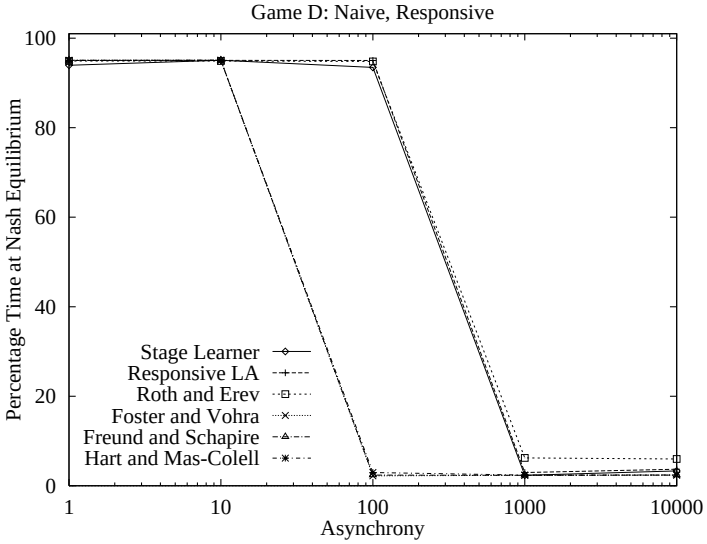
1 \ 2	*L*	*R*
T	1,2	3,0
B	0,0	2,1

Figure 4.1: Game D

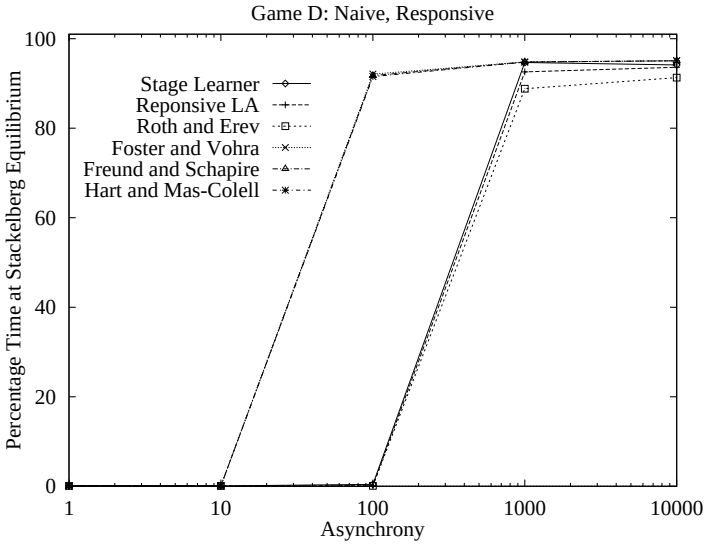
the presence of sufficient asynchrony, play does not converge to the Nash solution for any of the algorithms studied. Instead, play converges to the Stackelberg equilibrium, as depicted in Figure 4.2 (b). These results demonstrate that D^∞ does not always contain the asymptotic play. Note that these results are robust; in particular, the results are unchanged even when the game is studied with "noisy" payoffs $\hat{\pi}_i$, where $\hat{\pi}_i = \pi_i \pm \delta$, for small $\delta > 0$.

The transition from Nash to Stackelberg equilibrium depicted in Fig. 4.2 is rather abrupt. This observation prompted us to conduct further simulations at a series of intermediate values to more precisely determine the impact of asynchrony. For the reasonable learning algorithms, the transition between equilibria takes place when A falls between 100 and 1000; for the no regret algorithms, this transition takes place when A lies between 10 and 100. Fig. 4.3 depicts the details of these transitions in the respective ranges of asynchrony for the two sets of algorithms. Only reasonable learning algorithms ever clearly exhibit out-of-equilibrium behavior; the no regret algorithms transition directly from one equilibrium to the other.

Recall that (B, R) is the Stackelberg equilibrium in Game D. Fig. 4.4 plots the changing weights over time of strategy B for player 1 and strategy R for player 2 for the no external regret algorithm due to Freund and Schapire. The individual plays are also plotted; marks at 100 signify play of Stackelberg strategies, and marks at 0 signify play of Nash strategies. For comparison purposes, the synchronous case,

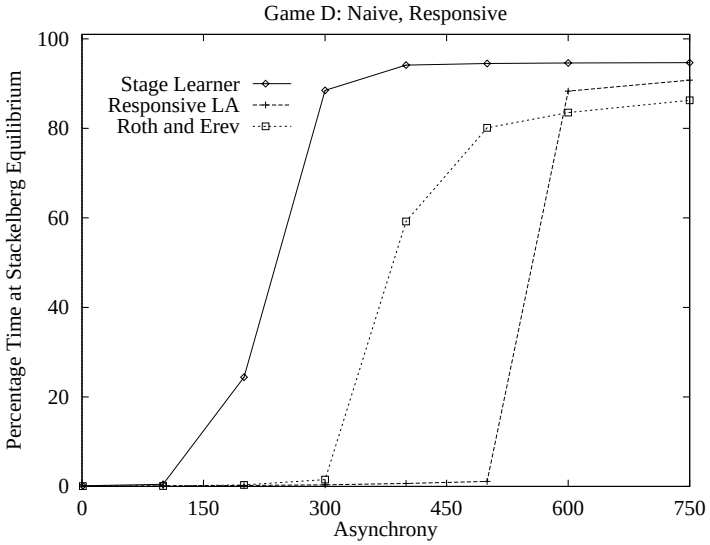


(a) Nash Equilibrium

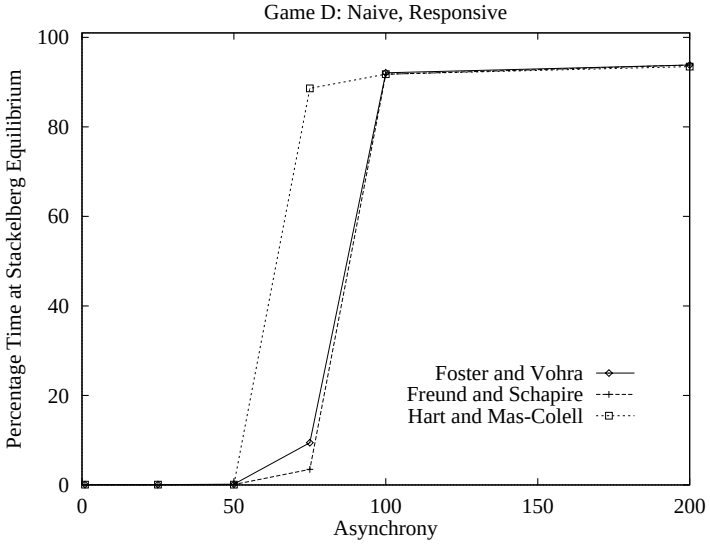


(b) Stackelberg Equilibrium

Figure 4.2: *Convergence to Equilibria in Game D.* (a) Percentage of time during which Nash equilibrium arises as the degree of asynchrony varies. (b) Percentage of time during which Stackelberg equilibrium arises.

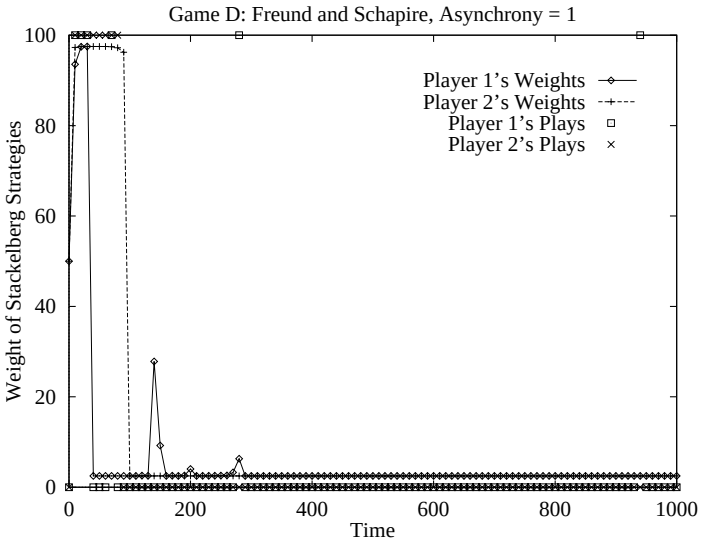


(a) Reasonable Learning Algorithms

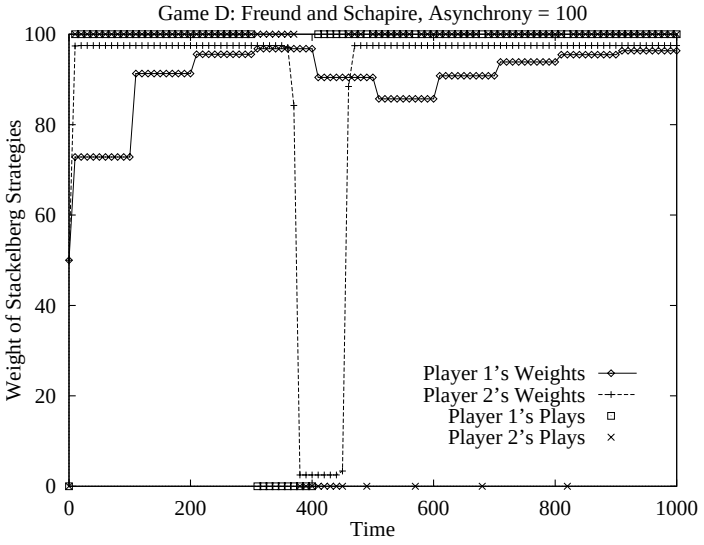


(b) No Regret Learning Algorithms

Figure 4.3: *Detail of Convergence to Stackelberg Equilibrium in Game D.*



(a) Nash Equilibrium ($A = 1$)



(b) Stackelberg Equilibrium ($A = 100$)

Figure 4.4: *Convergence to Equilibria in Game D: Algorithm due to Freund and Schapire.* (a) Weights of the Stackelberg equilibrium strategies over time when $A = 1$; play converges to Nash equilibrium. (b) Weights when $A = 100$; play converges to Stackelberg equilibrium.

where play converges to Nash equilibrium, as well as the asynchronous case with $A = 100$, where play converges to Stackelberg equilibrium, are depicted. Notice that play converges in the former case after roughly 300 iterations, while it converges in the latter case after roughly 1000 iterations (for the given settings of parameters ϵ and γ). In Fig 4.4(b), the leader (player 1) slowly learns to play the Stackelberg solution, and because the follower (player 2) is responsive, his weights follow the leader's plays. This behavior is representative of all the learning algorithms considered.

Game O

The next game that is studied in this section is depicted in Figure 4.5. This game is referred to as Game O, since it is O-solvable. In this game, $\{(T, L)\}$ is the unique Nash equilibrium and $\{(T, L)\} = D^{\infty} = S^{\infty} = O^{\infty}$.

1 \ 2	L	R
T	2,2	3,1
B	1,3	0,0

Figure 4.5: Game O

Simulations of all the algorithms, for levels of asynchrony ranging from 1 to 10,000, show that Nash equilibrium is played over 95% of the time. In particular, play does not diverge from the Nash equilibrium solution in this O-solvable game, as it did in Game D, regardless of the degree of asynchrony. It has been established that, for reasonable learners, O^{∞} is an upper bound on the solution concept. Our data is consistent with the same result holding for the other classes of algorithms considered, although this is far short of a proof that the O^{∞} solution concept applies to them

as well. The next game addresses the question of whether the O^∞ solution concept might in fact be too large a set.

Prisoners' Dilemma

This section presents the results of simulations of the repeated Prisoners' Dilemma (see Figure 4.6). In this game, $\{(D, D)\}$ is the unique Nash (as well as Stackelberg) equilibrium, and $\{(D, D)\} = D^\infty = S^\infty \neq O^\infty$, since O^∞ is the entire game. The Prisoner's Dilemma provides a simple test of the conjecture that the outcome of responsive learning in network contexts is described by the S^∞ solution concept, rather than the larger solution set O^∞.

1 \ 2	C	D
C	2,2	0,3
D	3,0	1,1

Figure 4.6: Prisoners' Dilemma

Simulations of all the algorithms, for levels of asynchrony ranging from 1 to 10,000, show that in this game the Nash equilibrium is played over 95% of the time. Since play does not diverge significantly from the Nash (and Stackelberg) equilibrium, the asymptotic play is not spread throughout O^∞; on the contrary, it is confined to S^∞.

Game S

The last simple two-player game that is studied is a game in which the players have three strategies. The game is depicted in Figure 4.7, and is referred to as Game S. In Game S, $D^\infty = \{T, L\}$; $S^\infty = \{T, L\} \times \{B, R\}$; and O^∞ is the entire game;

thus, $D^\infty \neq S^\infty \neq O^\infty$. The results of simulations of Game S resemble the results of simulations of Game D. [9] Figure 4.8 shows that the learning algorithms do not converge to the Nash equilibrium solution of this game when there is asynchrony. Instead, play converges to the Stackelberg equilibrium, as in Game D. This game provides a second test of the conjecture that the outcome of responsive learning in network settings is a strict subset of O^∞.

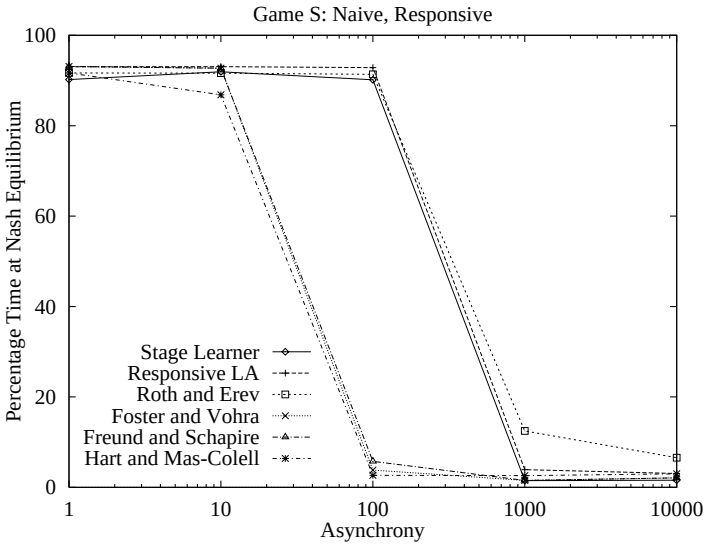
1 \ 2	L	C	R
T	2,2	4,0	2,0
M	1,1	3,3	0,2
B	0,0	3,0	1,1

Figure 4.7: Game S

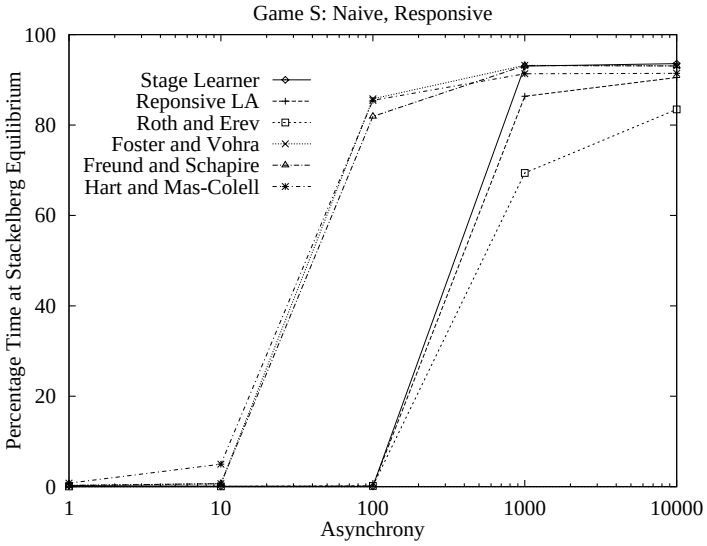
4.2.2 Externality Games

To test whether similar results apply to games with more than two players, we also experimented with externality games. An externality game, as defined in [40], is one in which each agent can choose either to participate or not to participate (in some joint venture) and where the payoffs obtained by a given player depend only on whether that player participates and on the total number of participating players. We now study a related class of games which are D-solvable and, for certain choices of the parameters, the games in this class are S-solvable and O-solvable as well.

[9] Note that in these simulations, the reasonable learning algorithms utilized $\epsilon = .1667$.



(a) Nash Equilibrium



(b) Stackelberg Equilibrium

Figure 4.8: *Convergence to Equilibria in Game S.* (a) Percentage of time during which Nash equilibrium arises as the degree of asynchrony varies. (b) Percentage of time during which Stackelberg equilibrium arises.

The class of games which we consider (class EG) is a discretization of the non-atomic games discussed in [41]. The set of players $\mathcal{N} = \{0, \ldots, N-1\}$, with $i \in \mathcal{N}$. The players have two possible strategies, namely 0 and 1, where 1 corresponds to participation, and 0 corresponds to non-participation. The number of participants, therefore, is given by $\lambda(s) = \sum_{i \in \mathcal{N}} s_i$, where s_i denotes the strategic choice of player i. Payoffs are determined as follows. The value to player i of participation is $v_i \in \mathbb{R}$, and the cost of participation is $C_i(\lambda)$, where C_i is a nondecreasing function of the externality. Thus, if player i participates, then $s_i = 1$ and $\pi_i(1, s_{-i}) = v_i - C_i(\lambda(s))$. Otherwise, if player i does not participate, then $s_i = 0$, and $\pi_i(s) = \phi\pi_i(1, s_{-i})$, for $\phi \in [0, 1)$. Intuitively, ϕ measures the extent to which players can opt out.

Note that the parameter ϕ does not affect the standard strategic elements of a given game in this class, such as best-replies or dominated strategies. In particular, if the game is D-solvable for $\phi = 0$ then it is D-solvable for all $\phi \in [0, 1)$. Similarly, varying ϕ does not change the set of Nash equilibria. Moreover, it is straightforward to show that when $\phi = 0$, if the game is D-solvable, then it must also be O-solvable (and therefore, also S-solvable). In contrast, for ϕ sufficiently close to 1, the game is not S-solvable (and therefore, not O-solvable). Thus, by varying ϕ we can create a class of games which are D-solvable but not necessarily S-solvable or O-solvable.[10]

In our simulations, we consider eight players (*i.e.*, $\mathcal{N} = \{0, \ldots, 7\}$), and we set $v_i = i$ and $C_i(\lambda) = \lambda/\mu$, for $\mu \in \mathbb{R}$. In the first set of simulations, we choose $\mu = 1.9$; we call this Game $\text{EG}_{1.9}$. This game is D-solvable and therefore has a unique Nash equilibrium. Moreover, this implies that for $\phi = 0$, this game must also be O-solvable; however, for ϕ sufficiently close to 1, it is neither S-solvable nor O-solvable. More specifically, when $\phi > 6/11 = .\overline{54}$, Game $\text{EG}_{1.9}$ has a Stackelberg equilibrium with player 2 as the leader which differs from the Nash equilibrium: the Nash equilibrium for all $\phi \in [0, 1)$ is $s = (0, 0, 0, 1, 1, 1, 1, 1)$, while the Stackelberg equilibrium (with player as the 2 leader and $\phi > 6/11$) is $s = (0, 0, 1, 0, 1, 1, 1, 1)$.

Simulations of Game $\text{EG}_{1.9}$ were conducted using the naive, responsive variants of the no-regret learning algorithms.[11] The convergence results in the asynchronous case

[10] Proofs of these claims appear in [50] for the class of games considered.

[11] The payoffs were translated by N/μ in simulations of responsive learning automata and the algorithm due to Roth and Erev in order to avoid negative payoffs.

(for $A = 5,000$) are listed in Table 4.1, for $\epsilon = .02$ and $\gamma = .002$. [12] Simulations of all algorithms show rapid convergence to the Nash equilibrium (NE) for all values of β in the synchronous case, and to the Stackelberg equilibrium (SE) when $\beta = .6$ and $\beta = .9$ in the asynchronous case. In particular, in Game $\text{EG}_{1.9}$, for certain choices of the parameters, the asymptotic play is not contained in the set D^∞.

Algorithm	$\beta = 0$	$\beta = .5$	$\beta = .6$	$\beta = .9$
Foster and Vohra	NE	NE	SE	SE
Freund and Schapire	NE	NE	SE	SE
Hart and Mas-Colell	NE	NE	SE	SE

Table 4.1: Game $\text{EG}_{1.9} : \gamma = .002, A = 5,000$

Another game which we simulated in the class EG is Game $\text{EG}_{2.1}$; this game is identical to Game $\text{EG}_{1.9}$, except $\mu = 2.1$. Like Game $\text{EG}_{1.9}$ this game is D-solvable. This implies that for $\phi = 0$, this game is also O-solvable; however, for ϕ sufficiently close to 1, this game is not O-solvable. Lastly, unlike Game $\text{EG}_{1.9}$, Game $\text{EG}_{2.1}$, is S-solvable. This game was simulated assuming all the same parameter values as the previous set of simulations, as well as some additional choices for γ. Selected results of these simulations appear in Table 4.2. Notice that regardless of the values of γ and ϕ, and even in the presence of extreme asynchrony, play converges to Nash equilibrium. Thus, as $D^\infty = S^\infty \neq O^\infty$, this game provides further evidence for the conjecture that asymptotic play of learning in network contexts is contained in S^∞.

Algorithm	$\beta = 0$	$\beta = .5$	$\beta = .6$	$\beta = .9$
Foster and Vohra	NE	NE	NE	NE
Freund and Schapire	NE	NE	NE	NE
Hart and Mas-Colell	NE	NE	NE	NE

Table 4.2: Game $\text{EG}_{2.1} : \gamma \in \{.01, .005, .002, .001\}, A = 10,000$

[12] In the algorithm due to Foster and Vohra, $\alpha = 1,000$; in the algorithm due to Freund and Schapire, $\beta = 1$; finally, in the algorithm due to Hart and Mas-Colell, $\kappa = 5$.

4.2.3 Congestion Games

Thus far, we have considered games with relatively small strategy spaces. In this section, we experiment with a larger strategy space, using an example that arises in computer networks. Consider several agents simultaneously sharing a network link, where each agent controls the rate at which she is transmitting data. If the sum of the transmission rates is greater than the total link capacity, then the link becomes congested and the agents' packets experience high delays and high loss rates. The transmission rates are controlled by each agent's congestion control algorithm, which vary the rates in response to the level of congestion detected.

One can model the interaction of congestion control algorithms as a cost-sharing game where the cost to be shared is the congestion experienced. That is, we can model this as a game where the strategies are the transmission rates r_i and the outcomes are the pairs (r_i, c_i), where c_i is the congestion experienced as function of the strategy profile r. The allocations must obey the sum rule $\sum_i r_i = F(\sum_i c_i)$, where F is a constraint function (*i.e.*, the total congestion experienced must be a function of the total load).

Most current Internet routers use FIFO packet scheduling algorithms, which result in congestion proportional to the transmission rate: $c_i = [r_i / \sum_j r_j] F(\sum_j c_j)$. FIFO implements the average cost pricing (ACP) mechanism. In contrast, the fair queuing packet scheduling algorithm can be modeled as leading to allocations such that c_i is independent of r_j as long as $r_j \geq r_i$ (this condition, plus anonymity, uniquely specifies the allocations). Fair queuing implements the Serial mechanism (see [100] for a detailed description).

Chen [21] studies the two-player congestion game defined under the following conditions: (1) linear utility functions $U_i(r_i, c_i) = \alpha_i r_i - c_i$, (2) quadratic congestion $F(x) = x^2$, and (3) discrete strategy space $S_i = \{1, 2, \ldots, 12\}$. For parameters $\alpha_1 = 16.1$ and $\alpha_2 = 20.1$, the game defined by the ACP mechanism is D-solvable, but it is not S-solvable or O-solvable. The unique Nash equilibrium is $(4, 8)$ and the Stackelberg equilibrium with player 2 leading is $(2, 12)$. In contrast, the game defined by the Serial mechanism is O-solvable, with unique Nash equilibrium $(4, 6)$.

We conducted simulations of both the Serial and the ACP mechanism using the naive and responsive variants of the no-regret learning algorithms, with the degree of experimentation $\epsilon = .02$. [13] In our simulations of the Serial mechanism, all of the learning algorithms concentrate their play around the Nash equilibrium. In the ACP mechanism, when play is synchronous, the asymptotic behavior again centers around the Nash equilibrium. However, in the presence of sufficient asynchrony (*e.g.*, $A = 5,000$, when $\gamma = .002$), play converges to the Stackelberg equilibrium.

4.2.4 Discussion

Our results thus far indicate that when responsive learning algorithms play repeated games against one another, their play can reach outside the serially undominated set, given sufficient asynchrony. In our examples, however, the outcome is either largely inside the serially undominated set, or with sufficient asynchrony, converges to the Stackelberg equilibrium. We did not observe more general behavior, with probabilities spread over a wider set of strategies, although, based on work by Foster and Vohra [37], we conjecture that such behavior arises in more complex games.

4.3 Simulations in Non-network Contexts

Network contexts differ from standard learning contexts considered in the literature in three important ways, namely, responsive learning, limited information access, and asynchronous play. In previous sections, we have looked at how varying asynchrony affects the convergence properties of learners. In this section, we briefly consider the remaining two properties of learning in network contexts.

First we augment the information structure by considering contexts in which play is *informed*; in other words, in informed settings, learners know the payoffs that would have occurred had they chosen an alternative action. This typically arises when players (i) know the payoff matrix, and (ii) can observe the actions of the other players. Not surprisingly, in this setting play is confined within D^∞. Unfortunately

[13] In the algorithm due to Foster and Vohra, $\alpha = 5,000$; in the algorithm due to Freund and Schapire, $\beta = 1$; finally, in the algorithm due to Hart and Mas-Colell, $\kappa = 2,000$.

in network contexts, informed learning is not an option; the basic structure of the Internet is such that learning is inherently uninformed.

Our second consideration, namely responsiveness, is not inevitable, but instead reflects a common (and appropriate) design choice on the Internet. We find the behavior of naive and non-responsive learners, in the presence of asynchrony, to be complex and do not claim to understand such asymptotic behavior; for example, we do not know whether play always converges to D^∞. We demonstrate some of this complexity through simulations of the Shapley game, a classic example for which fictitious play does not converge. Irrespective of these complexities, non-responsive learning is not viable in network contexts due to the non-stationarity of payoffs.

The simulation results in this section show that the seemingly obvious conjecture that asynchrony alone leads to Stackelberg behavior is not true in general. In our simulations, this conjecture only when we consider naive *and* responsive learning algorithms, as are relevant for network contexts. If the algorithms are informed, or non-responsive, we do not observe Stackelbergian behavior.

4.3.1 Informed Learning

Recall that the simulations that lead to asymptotic play outside D^∞ utilize the naive and responsive variants of the set of learning algorithms. In our simulations of Game D (see Figure 4.1), responsive but informed learning does *not* lead to play outside D^∞, even in the presence of extreme asynchrony, for enhanced versions of reasonable learning that utilize full payoff information and the original no regret algorithms. Specifically, for levels of asynchrony between 1 and 10,000, simulations of responsive and informed algorithms show that Nash equilibrium is played over 95% of the time.[14]

Intuitively, this occurs because the set of informed learning algorithms compares the current payoff with the potential payoffs of the other strategies, *assuming that the other agents keep their strategies fixed.* The key to the Stackelberg solution is that the leader evaluates his payoff in light of the probable responses of other agents. Naive learners, when learning at a slow rate, do this implicitly; that is, they only receive

[14] The simulations of Game D discussed in this section and the next depend on the same set of parameter values as in Section 4.2.1; specifically, $\gamma = .01$ for all algorithms, while $\epsilon = .025$ for the reasonable learning algorithms and $\epsilon = .05$ for the no-regret algorithms.

their payoffs after the other players respond to their play. The informed learning algorithms which we consider do not take this reaction into account.

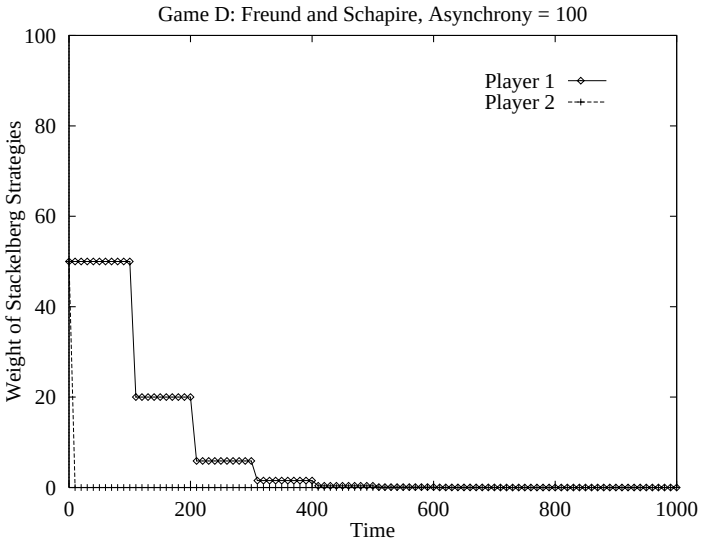
If informed and responsive learning algorithms do indeed converge in general to D^∞, this might be seen as an argument to consider only informed learning algorithms. However, in network contexts this is not an option; the information about other payoffs is not available and so we are forced to use naive learning algorithms. However, agents do have a choice as to whether to use responsive or non-responsive learning algorithms, a subject to which we now turn.

4.3.2 Non-responsive Learning

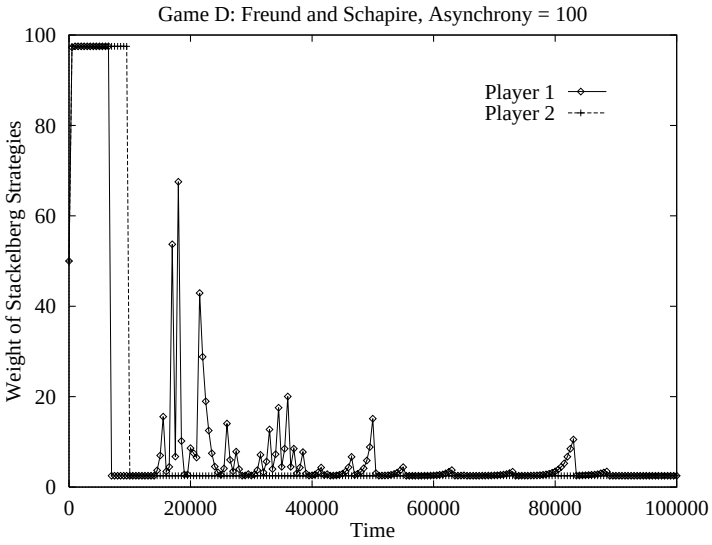
We now consider naive but non-responsive learners. Simulations of Game D (see Figure 4.1) using non-responsive algorithms, for levels of asynchrony between 1 and 10,000, show that the Nash equilibrium is played over 99% of the time for the set of informed algorithms and over 95% of the time for the naive set. In particular, the behavior of informed but non-responsive learners is similar to that of informed and responsive learners, in that they learn to eliminate dominated strategies, resulting in convergence to D^∞. This seems reasonable because the informed and non-responsive algorithms which we study are approximately adaptive in the sense of Milgrom and Roberts [78], who prove that such adaptive learners converge to D^∞.

The case of naive, non-responsive learners is slightly more complicated. What appears to be happening is that while initially the follower responds to the play of the leader, eventually the follower becomes less responsive and therefore stops following, which causes the leader to lose its advantage. Figure 4.9 depicts the weight over time of strategy B for player 1 and strategy R for player 2 in simulations of the no external regret learning algorithm due to Freund and Schapire with level of asynchrony 100. Note that values are recorded only every 500 rounds. Notice that player 1 (the leader) is inclined to increase his weight, but in the absence of a noticeable response from player 2 (the follower), player 1 is forced to settle at the Nash equilibrium.

While in our simulations of relatively simple games, the asymptotic play of non-responsive learning algorithms is confined to D^∞, we have no proof at present that naive and non-responsive learning algorithms must remain inside D^∞. This subject



(a) Informed Version



(b) Naive Version

Figure 4.9: *Asynchronous, Non-responsive Learning in Game D:* Algorithm due to Freund and Schapire. (a) Stackelberg equilibrium strategy weights in informed case; play quickly converges to Nash equilibrium. (b) Same weights in naive case; play again converges to Nash equilibrium.

warrants further study. Regarding network contexts, however, the analysis of such questions is moot, since non-responsive learners are unsatisfactory for a much simpler reason: their performance is sub-optimal in non-stationary settings.

Non-stationary Environments

Consider a simple, one-player two-strategy game where the payoffs initially are 1 for strategy A and 0 for strategy B. We simulate a non-stationary version of this game where the payoffs are reversed every 5,000 rounds, and we compare the performance of responsive and non-responsive learning algorithms.

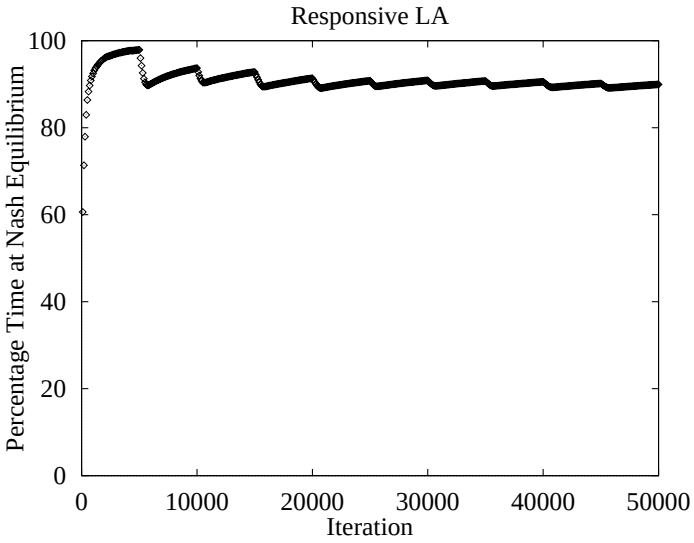
Figures 4.10 and 4.11 (a) plot the cumulative percentage of time spent playing the optimal strategy in simulations of sample reasonable learning algorithms.[15] All the reasonable learning algorithms – namely stage learning, responsive learning automata, and the algorithm due to Roth and Erev – spend over 90% of their time at the current optimal strategy in the simulated quasi-static environment. In addition, the resulting fluctuations in the weight of strategy A in this game are depicted in (b); observe that the weight of strategy A changes with the state of the environment.

In contrast to the reasonable learning algorithms, the non-responsive, no-regret algorithms (both the naive and informed versions) perform poorly in non-stationary environments. Figure 4.12 plots the cumulative percentage of time spent playing the Nash equilibrium for Freund's and Schapire's no external regret algorithm, in both its responsive and non-responsive forms.[16] Note that the non-responsive version of the algorithm spends only about 50% of its time playing the currently optimal strategy. This behavior is representative of all the no-regret algorithms studied. This is because the non-responsive no-regret algorithms fixate on one strategy early on – the one that is initially optimal – and are unable to adjust to future changes in environmental conditions.

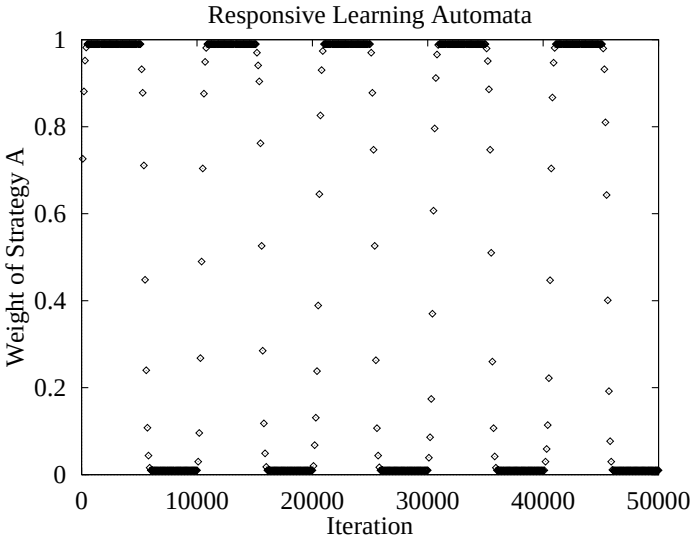
Thus, the criterion of no-regret, while perhaps appropriate for learning in static environments, is not sufficient for learning non-stationary payoff functions. Since network contexts are typically non-stationary and since this non-stationarity can be

[15] The algorithmic parameters for the reasonable learning algorithms were chosen as follows: $\epsilon = \gamma = .01$.

[16] For simulation purposes, in the algorithm due to Freund and Schapire, $\beta = 1$, and in the responsive case, γ was set equal to .01.

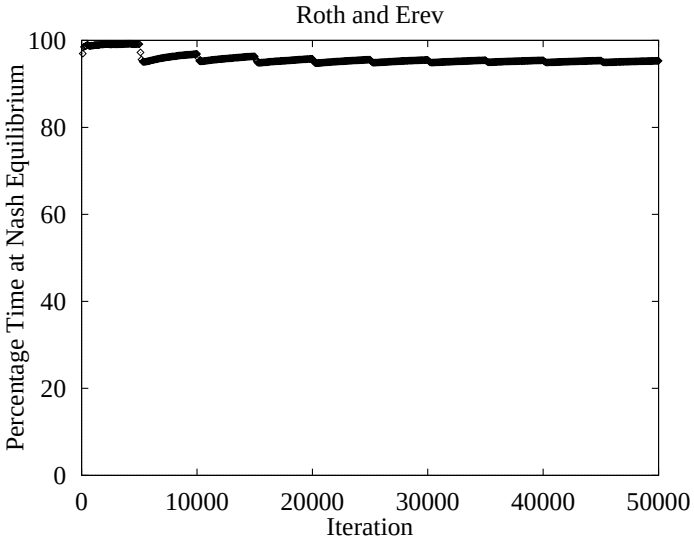


(a) Percentage Time at Nash Equilibrium

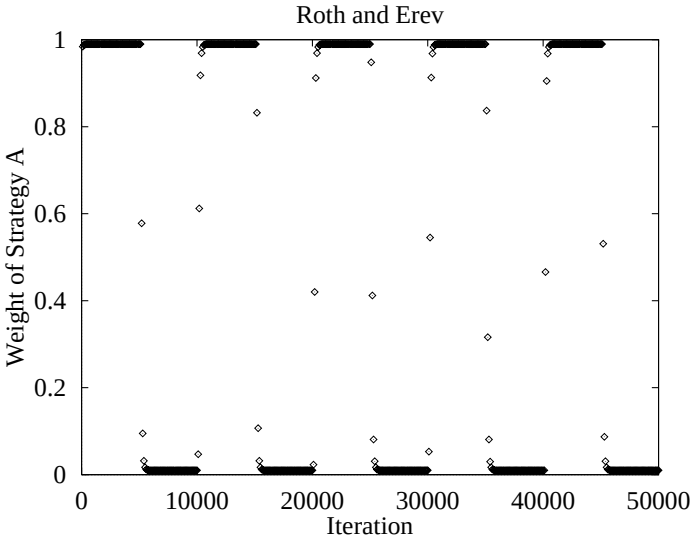


(b) Weight of Strategy A

Figure 4.10: *Responsive LA in Quasi-static Environment.*

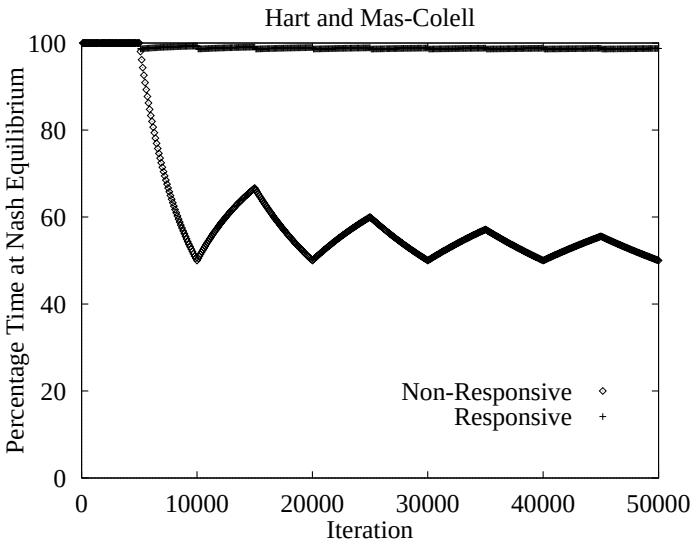


(a) Percentage Time at Nash Equilibrium

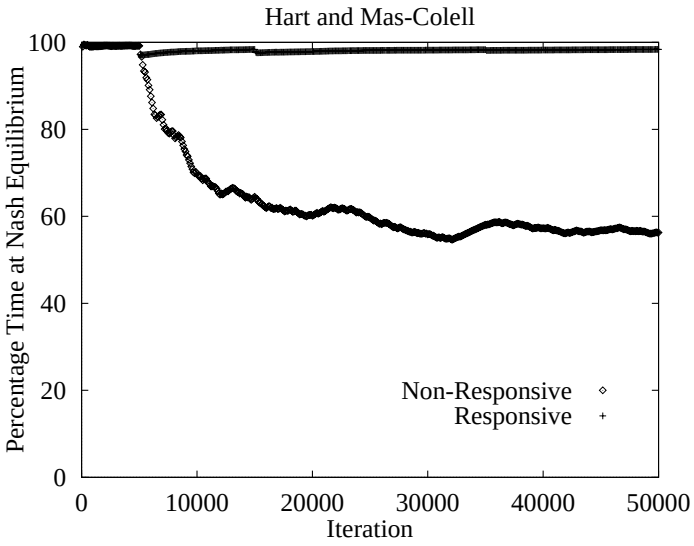


(b) Weight of Strategy A

Figure 4.11: *Algorithm due to Roth and Erev in Quasi-static Environment.*



(a) Informed Version



(b) Naive Version

Figure 4.12: *Algorithm due to Hart and Mas-Colell in Quasi-static Environment.* Non-responsive vs. responsive learning.

detected only via experimentation (one cannot observe the change in the structure of the game directly), the learning algorithms employed should be responsive.

Shapley Game

In this section, we compare the behavior of responsive and non-responsive learners in the Shapley game (see Figure 4.3.2), a well-known game in which fictitious play, an informed, non-responsive algorithm, does not converge. In this game, fictitious play results in cycling through the cells with 1's in them (cells *1, 2, 5, 6, 7,* and *9* in Figure 4.3.2), with ever-increasing lengths of play in each such cell [98]. One is led to conjecture that this fascinating behavior arises because of the clear-cut choices made by fictitious play – the strategy with the highest expected payoff is chosen with probability 1, leading to abrupt transitions in the trajectory of play.

1 \ 2	*L*	*C*	*R*
T	1,0 *1*	0,1 *2*	0,0 *3*
M	0,0 *4*	1,0 *5*	0,1 *6*
B	0,1 *7*	0,0 *8*	1,0 *9*

Figure 4.13: Shapley Game

Surprisingly, in our simulations,[17] we observe behavior which is similar to that of fictitious play for most of the non-responsive learning algorithms[18] – both informed

[17] The graphs present the results of simulations of the algorithm due to Foster and Vohra with $\epsilon = .03$.

[18] The only exception is the algorithm of Hart and Mas-Colell which is known to converge to correlated

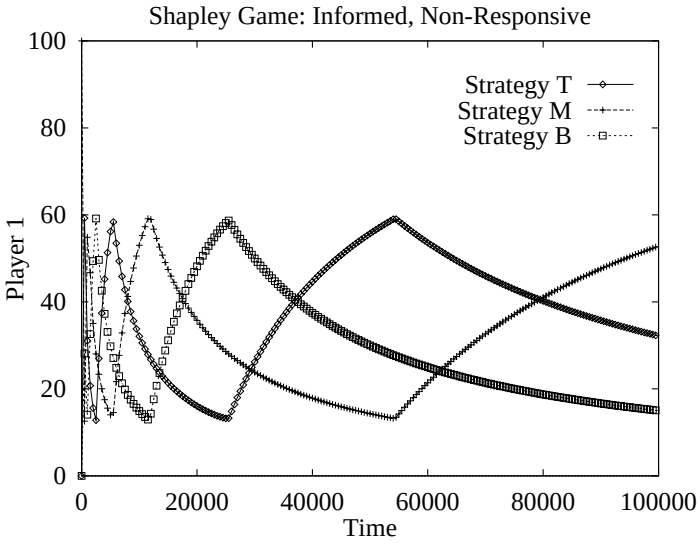
(see Figures 4.14 (a) and 4.15 (a)) and naive (see Figure 4.14 (b) and 4.15 (b)) – even though these algorithms do not in general induce discrete changes. In particular, Figure 4.14 plots the cumulative percentage of time player 1 plays each of the three strategies; although not depicted, the behavior patterns of player 2 are identical. In addition, Figure 4.15 depicts the joint empirical frequencies of the various strategy profiles after 10^6 iterations, where the x-axis is labeled *1,...,9* corresponding to the cells in Figure 4.3.2. Via this joint perspective, we see that both informed and naive, non-responsive learners spend very little time playing in cells without a 1 in them. Specifically, in the informed case, the likelihood of play in cells *3, 4*, and *8* approaches 0, and in the naive case this likelihood approaches ϵ/N.

In contrast, the responsive algorithms, while they *do* display the same cycling behavior, the duration of play in each cell does *not* continue to grow. Instead the responsive algorithms spend equal amounts of time in each of the distinguished cells. This is depicted in Figure 4.16 (a) (the informed case) and Figure 4.16 (b) (the naive case), which plot the cumulative percentage of time player 1 spends playing each of the three strategies. Notice that these graphs converge to 33% for all strategies. In addition, Figure 4.17 depicts the joint empirical frequencies of the various strategy profiles after 10^6 iterations. One interesting feature of the set of responsive learning algorithms is that their empirical frequencies converge to that of the fair and Pareto optimal correlated equilibrium; in particular, both players have expected average payoff of 1/2. This follows from the bounded memory of these algorithms.

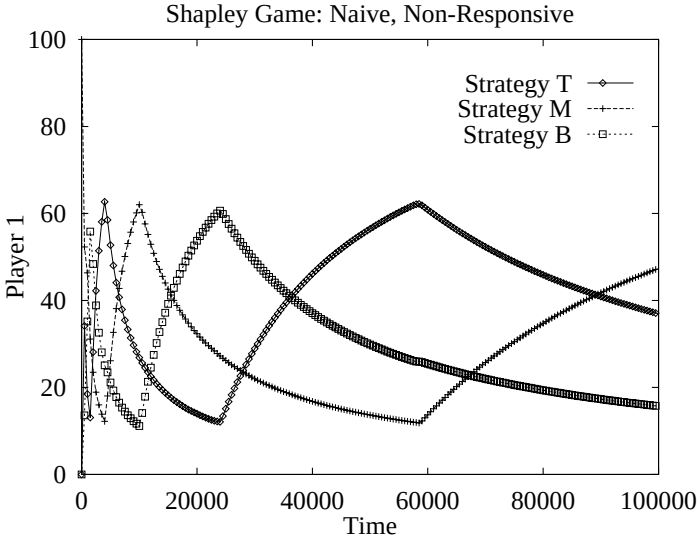
4.4 Related Work

There is a vast body of economics literature on learning through repeated play of games, and we make no attempt here to provide a detailed review; for a comprehensive discussion, see the review by Fudenberg and Levine [46]. There is also substantial interest within the artificial intelligence community in the area of multi-agent learning; see the recent special issue of *Machine Learning* on this topic. In this section, we place our work in the context of the varying approaches to learning taken by economics and artificial intelligence researchers.

equilibrium, and in fact converges to the mixed strategy Nash equilibrium in the Shapley game.

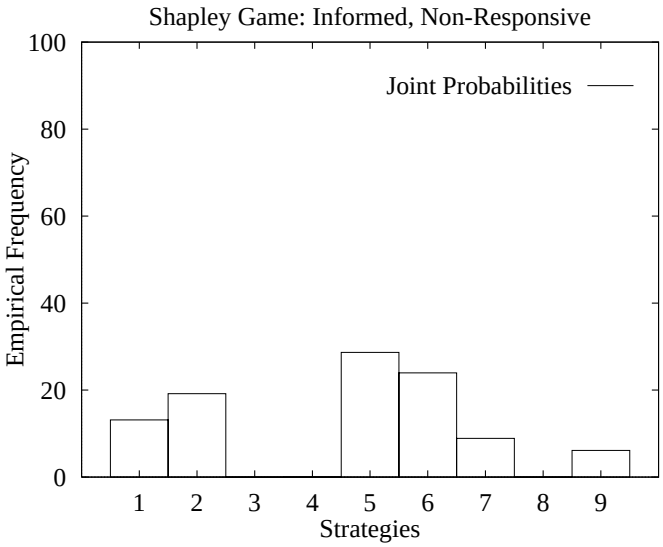


(a) Informed, Non-Responsive

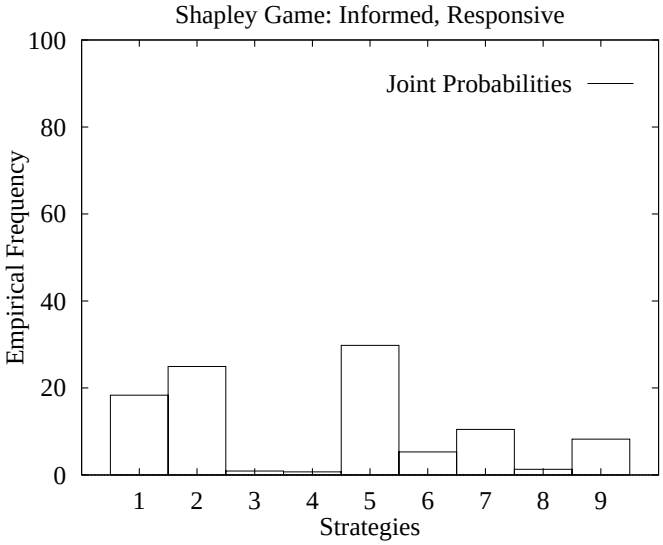


(b) Naive, Non-Responsive

Figure 4.14: *Non-Responsive Learning in Shapley Game.* Cumulative percentage of time player 1 plays each of his strategies assuming non-responsive learning.

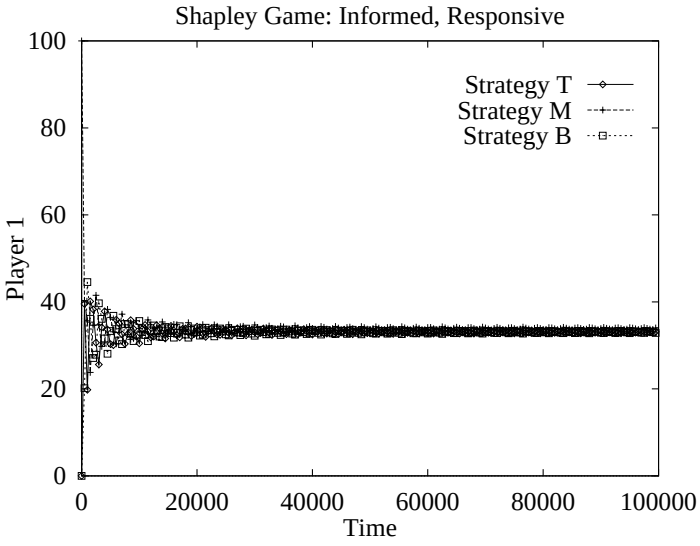


(a) Informed, Non-Responsive

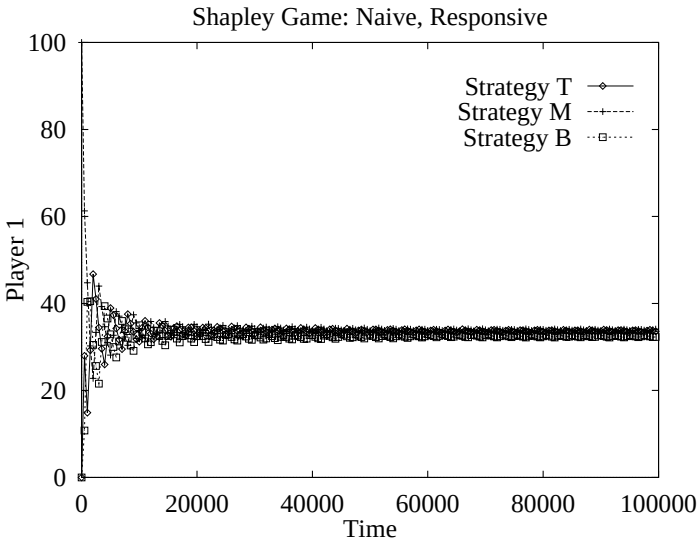


(b) Naive, Non-Responsive

Figure 4.15: *Joint Empirical Frequencies of Strategy Profiles in the Shapley Game via Non-responsive Learning.* The x-axis labels *1,...,9* correspond to the cells in Figure 4.3.2.

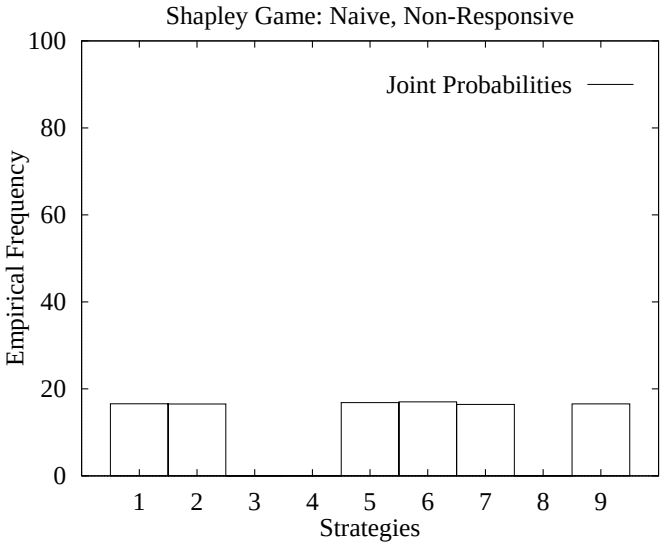


(a) Informed, Responsive

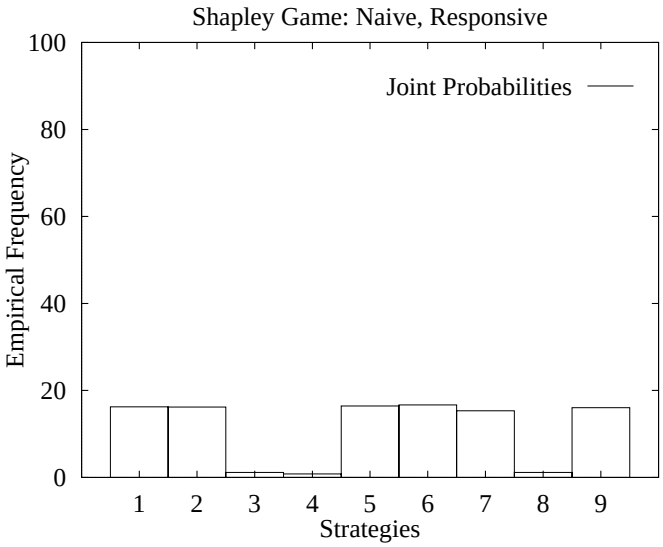


(b) Naive, Responsive

Figure 4.16: *Responsive Learning in Shapley Game.* Cumulative percentage of time player 1 plays each of his strategies assuming responsive learning.



(a) Informed, Responsive



(b) Naive, Responsive

Figure 4.17: *Joint Empirical Frequencies of Strategy Profiles in the Shapley Game using Responsive Learning.* The x-axis labels *1,...,9* correspond to the cells in Figure 4.3.2.

4.4.1 Relevance to Economics

Economic research on learning falls roughly into two camps. The "high-rationality" approach involves learning algorithms that aim to predict their opponents' strategies, and optimize with respect to those predictions. The prediction methods might be for example, Bayesian (as in Kalai and Lehrer [65]), calibrated (as in Foster and Vohra [37]), or consistent (as in Fudenberg and Levine [45, 44]). Typically, the asymptotic play of high-rationality learning is either correlated or Nash equilibrium. Since these algorithms depend on knowledge of the underlying structure of the game, however, they are not applicable in network contexts.

In contrast, the features of "low-rationality" approaches to learning are similar to those which concern us here; in particular, agents have no information about the game other than the payoffs they receive. Examples of such work include Roth and Erev [89], Erev and Roth [31], Borgers and Sarin [15], Mookerji and Sopher [80], and Van Huyck *et al.* [61]; most of these algorithms as they were initially proposed are not responsive, but as we show in Chapter 2, they can be made responsive via slight modifications. The focus of these papers is typically on matching the results of human experiments, whereas we focus instead on the nature of the asymptotic play. Chen [21], however, performed experiments on the congestion game discussed in Section 4.2.3 in which she compared synchronous and asynchronous play, as well as learning in full information one-shot games (zipper design, where play is repeated, but is against different opponents) versus naive learning in repeated games.

4.4.2 Relevance to Artificial Intelligence

In the context of artificial intelligence research, the present study is related to work in machine learning, where our sample set of algorithms classify as reinforcement learning schemes. Our angle is unique, however, since we consider multi-agent learning in network contexts. Recall that such contexts are characterized by naiveté regarding payoff information, non-stationary environments, and asynchronous interactions.

Traditional reinforcement learning algorithms (see survey by Kaelbling, *et al.* [63] and text by Sutton and Barto [106]) were designed for single agent learning that took place in stationary environments where payoffs are determined by state. Like learning

in network contexts, reinforcement learning is naive in the sense that complete payoff information is not assumed to be available to the learner. In contrast with learning in network contexts, however, reinforcement learning algorithms are not designed for use in non-stationary environments and therefore aim to maximize the discounted sum of expected future returns, whereas the algorithms considered here behave myopically. The theoretical bounds that are known for reinforcement learning algorithms such as Q-learning [113] do not apply in the non-stationary settings that are typical of network contexts.

More recently, reinforcement learning research has expanded its focus to multi-agent learning. Such learning, however, can be tricky simply because as agents learn, their actions change, and therefore their impact on the environment changes, often leading to non-stationarities. For simplicity, most studies in multi-agent learning consider settings where payoffs are either negatively correlated, as in zero-sum games (see, for example, Littman [73]), or positively correlated, as in coordination games (see, for example, Shoham and Tennenholtz [102]). Two notable exceptions include Wellman and Hu [114], who describe theoretical results on multi-agent learning in market interactions, and Sandholm and Crites [95], who conducted empirical studies of multi-agent reinforcement learning in the Prisoners' Dilemma. Similarly, this work considers multi-agent learning in positive sum games.

The results described here on convergence to Stackelberg equilibria were largely a result of the asynchronous nature of play and learning. Empirical investigations of asynchronous Internet interactions have been reported by Lukose and Huberman [75], but their studies concern time correlations that arise from such behavior rather than focus on convergence properties. The convergence of asynchronous machine learning algorithms where players' moves are determined by discrete random processes have recently been investigated in pricing models for electronic commerce by Greenwald and Kephart [51].

Finally, regarding theoretical computer science, the algorithms which we refer to as satisfying no regret optimality criteria, are also often described as achieving reasonable competitive ratios. Borodin and El-Yaniv [16] present a thorough discussion of the competitive analysis of on-line algorithms.

4.5 Conclusion

This chapter presented the results of simulation experiments conducted using six learning algorithms which embody three distinct notions of optimality: one average-case performance measure, and two worst-case performance measures. In the suite of relatively simple games examined here, all the algorithms exhibited qualitatively similar behavior. Thus, it seems that in network contexts, the key property is not which type of optimality is achieved, but rather, responsiveness. In low-information settings, where learners are necessarily naive and thus cannot detect changes in the structure of the game directly, algorithms should be able to respond to changes in environmental conditions in bounded time. It is shown in this chapter that when such naive and responsive learning algorithms operate in asynchronous settings, the asymptotic play need not lie within D^∞. The question of whether play outside D^∞ arises for naive but non-responsive players in asynchronous environments remains open, but presumably such behavior would only arise in games with more players and larger strategy sets than have been studied in this chapter.

It has been established previously that for reasonable learning algorithms the asymptotic play is contained within O^∞, and it was further conjectured that such play is contained within the smaller set S^∞. Our experimental results are consistent with this conjecture. While these simulation results are suggestive, they are in no way conclusive, and so we are left with the open question of what the appropriate solution concept is for naive and responsive learning algorithms in asynchronous settings.

Chapter 5

Shopbot Economics

5.1 Introduction

Shopbots, programs that search the Internet for advertised goods and services on behalf of consumers, herald a future in which autonomous agents will be an essential component of nearly every facet of e-commerce [20, 32, 68, 70, 109]. In response to a consumer's expressed interest in a specified good or service, a typical shopbot queries several dozen web sites, and then collates and sorts the available information for the user — all within seconds. For example, `www.shopper.com` claims to compare 1,000,000 prices on 100,000 computer-oriented products! In addition, `www.acses.com` compares the prices and expected delivery times of books offered for sale on-line, while `www.jango.com` and `webmarket.junglee.com` offer everything from apparel to gourmet groceries. Shopbots can out-perform and out-inform even the most patient, determined consumer, for whom it would otherwise take hours to obtain far less coverage of available goods and services.

Shopbots deliver on one of the great promises of the Internet and e-commerce: a radical reduction in the cost of obtaining and distributing information. Freer flow of information is expected to profoundly affect market efficiency, as economic frictions will reduce significantly [49, 72, 74]. Transportation costs, shopping costs, and menu costs (the costs of evaluating, updating, and advertising prices) should all decrease, as a consequence of the digital nature of information and the presence of autonomous agents that find, process, collate, and disseminate information at little cost.

In today's electronic marketplace, however, shopbots are a conundrum. On one hand, they are clearly a useful weapon for consumers; armed with up-to-the-minute information, consumers can demand that firms behave more competitively. Many of us would be happy to purchase our goods from the lowest-priced, highest-quality dealer, if only the cost and effort of obtaining complete and accurate information were not so monumental. Some vendors have responded to this threat by deliberately blocking automated agents from their sites; other vendors welcome shopbots as a means of attracting consumers who otherwise might not have known about them, or might not have thought to purchase from them [74]. Some of the vendors in this latter class even *sponsor* shopbots, by paying for the opportunity for their products to be listed on shopping sites such as `www.shopper.com`.

As the XML standardization effort gains momentum [107], one of the major barriers preventing the mass-production of shopbots is likely to be overcome — the headache associated with parsing the idiosyncrasies of individual vendor's `.html` files. Some vendors who oppose the use of shopbots are embedding their prices in text images, but there has been substantial research on the automated detection and recognition of such text, which aims to provide enhancements to commercially available OCR technology [117]. The outcome of progress in standardization and text extraction research may well be a great proliferation of shopbots, emerging as representatives of all forms of goods and services bought and sold on-line. What are the implications of the widespread use of shopbots? What sort of *sellbots* might firms implement to combat the increasing presence of shopbots, and how might the dynamic interplay of shopbots and sellbots evolve? In general, what is the expected impact of agent technology on the electronic marketplace?

DeLong and Froomkin [74] qualitatively investigate the ongoing emergence of shopbots; in particular, they note that short of violating anti-trust laws, firms will be hard pressed to prevent their competitors from sponsoring shopbots, in which case those who do not partake are likely to experience decreased sales. In this chapter, we utilize quantitative techniques to address the aforementioned questions. We propose, analyze, and simulate a simple economic model that captures some of the essence of shopbots, and aims to shed light on the potential impact of shopbots on markets. Looking ahead, we project that shopbots will evolve into economic entities in their

own right (*i.e.*, utility maximizers), interacting with billions of other economically-motivated software agents. Moreover, we predict the emergence of *pricebots*, a form of sellbots that set prices so as to maximize profits for firms, just as shopbots seek to minimize costs to consumers. Along these lines, we study adaptive price-setting algorithms, which pricebots might use to combat the growing community of shopbots, in a full-fledged agent-based electronic marketplace.

This chapter is organized as follows. The following section presents our model of shopbot economics. Section 5.3 analyzes this model from a game-theoretic point of view, proving first that there is no pure strategy Nash equilibrium, and then deriving the symmetric mixed strategy Nash equilibrium.[1] Section 5.4 describes a variety of adaptive price-setting algorithms and the results of their simulation under the prescribed model. An underlying theme throughout is whether or not adaptive learning yields the derived game-theoretic solution. A discussion of a possible future evolution of shopbots and pricebots follows in Section 5.5, while in Section 5.6, related work is described in which economists long ago predicted the emergence of today's shopbot-like services. Finally, concluding remarks and ideas for future research appear in Section 5.7.

5.2 Model

We consider an economy (see Figure 5.1) in which there is a single homogeneous good that is offered for sale by S sellers and of interest to B buyers, with $B \gg S$. Each buyer b generates purchase orders at random times, with rate ρ_b, while each seller s reconsiders (and potentially resets) its price p_s at random times, with rate ρ_s. The value of the good to buyer b is v_b, while the cost of production for seller s is r_s.

A buyer b's utility for a good is a function of its price as follows:

$$u_b(p) = \begin{cases} v_b - p & \text{if } p \leq v_b \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

[1] The material presented Sections 5.2 and 5.3 is closely related to well-known economic results; however, our work deviates from standard economic approaches (exceptions include Hopkins and Seymour [58] and Diamond [27]) in Section 5.4, where we consider price adjustment processes.

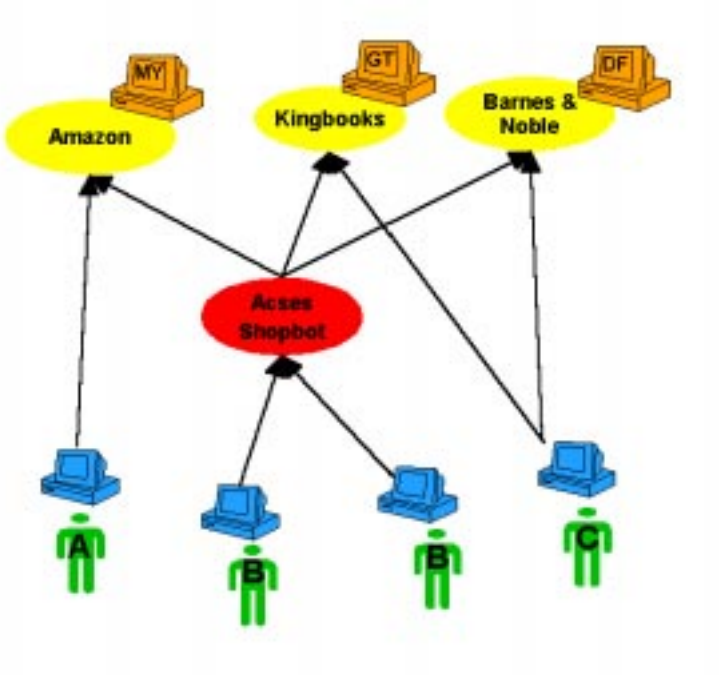


Figure 5.1: Shopbot Model

which states that a buyer purchases the good from a given seller if and only if the seller's offering price is less than the buyer's valuation of the good; if price equals valuation, we make the behavioral assumption that a transaction occurs. We do not assume that buyers are utility maximizers; instead we assume that buyers use one of the following *fixed sample size* search rules in selecting the seller from which to purchase:[2,3]

1. *Any Seller:* buyer selects seller at random, and purchases the good if the price charged by that seller is less than the buyer's valuation.

[2] It is also possible to consider all buyers as utility maximizers, with the additional cost of searching for the lowest price made explicit in the buyer utility functions. In particular, the search cost for bargain hunters is taken to be zero, while for those buyers who use the any seller strategy, its value is greater than v_b. The relationship between the exogenous model of the buyer distribution and the endogenous model which incorporates the cost of information acquisition and allows for explicit buyer decision-making is further explored in Section 5.6 on related work.

[3] In the economics literature (see, for example, Burdett and Judd [19]), buyers that employ strategies of this nature are said to use search rules of *fixed sample size* i; in particular, for buyers of type A, B, and C, respectively, $i = 1$, $i = S$, and $i = 2$.

2. *Bargain Hunter:* buyer checks the offer price of all sellers, determines the seller with the lowest price, and purchases the good if that lowest price is less than the buyer's valuation. (This type of buyer corresponds to those who take advantage of shopbots.)

3. *Compare Pair:* buyer selects two different sellers at random, determines which one is offering the good at the lower price, and purchases the good from that seller if that price is less than the buyer's valuation. (Ties are broken randomly.)

The buyer population is exogenously given as a mixture of buyers employing one of these strategies; specifically, fraction w_A using the *Any Seller* strategy, fraction w_B using the *Bargain Hunter* strategy, and fraction w_C using the *Compare Pair* strategy, where $w_A + w_B + w_C = 1$. Buyers employing these respective strategies are referred to as type A, type B, and type C buyers.

The profit function π_s for seller s per unit time as a function of the price vector $\vec{p}$ is expressed as follows:

$$\pi_s(\vec{p}) = (p_s - r_s)D_s(\vec{p}) \tag{5.2}$$

where $D_s(\vec{p})$ is the rate of demand for the good produced by seller s. This rate of demand is determined by the overall buyer rate of demand, the likelihood of buyers selecting seller s as their potential seller, and the likelihood that the chosen seller's price p_s will not exceed the buyer's valuation v_b. If $\rho = \sum_b \rho_b$, and if $h_s(\vec{p})$ denotes the probability that seller s is selected, while $g(p_s)$ denotes the fraction of buyers whose valuations satisfy $v_b \geq p_s$, then $D_s(\vec{p}) = \rho B h_s(\vec{p}) g(p_s)$. Note that $g(p_s) = \int_{p_s}^{\infty} \gamma(x)dx$, where $\gamma(x)$ is the probability density function describing the likelihood that a given buyer has valuation x. For example, suppose that the buyers' valuations are uniformly distributed between 0 and v; then the integral yields $g(p_s) = 1 - p_s/v$. Alternatively, if $v_b = v$ for all buyers b, then $\gamma(x)$ is the Dirac delta function $\delta(v - x)$, and the integral yields a step function $g(p_s) = \Theta(v - p_s)$ as follows:

$$\Theta(v - p_s) = \begin{cases} 1 & \text{if } p_s \leq v \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

Without loss of generality, define the time scale *s.t.* $\rho B = 1$. Now $D_s(\vec{p}) = h_s(\vec{p})g(p_s)$, and π_s is interpreted as the expected profit for seller s per systemwide unit sold.

The probability $h_s(\vec{p})$ that buyers select seller s as their potential seller depends on the distribution of the buyer population, namely (w_A, w_B, w_C). In particular,

$$h_s(\vec{p}) = w_A h_{s,A}(\vec{p}) + w_B h_{s,B}(\vec{p}) + w_C h_{s,C}(\vec{p}) \tag{5.4}$$

where $h_{s,A}(\vec{p}), h_{s,B}(\vec{p})$, and $h_{s,C}(\vec{p})$ are the probabilities that seller s is selected by buyers of type A, B, and C, respectively. It remains to determine these probabilities.

Given that the buyers' strategies depend on the relative ordering of the sellers' prices, it is convenient to define the following functions:

- $\mu_s(\vec{p})$ is the number of sellers charging a higher price than s,
- $\tau_s(\vec{p})$ is the number of sellers charging the same price as s, excluding s itself, and
- $\lambda_s(\vec{p})$ is the number of sellers charging a lower price than s.

Note that $\mu_s(\vec{p}) + \tau_s(\vec{p}) + \lambda_s(\vec{p}) = S - 1$, for all s.

The probability that a buyer of type A select a seller s is independent of the ordering of sellers' prices; in particular, $h_{s,A}(\vec{p}) = 1/S$. Buyers of type B, however, select a seller s if and only if s is one of the lowest price sellers: *i.e.*, s is *s.t.* $\lambda_s(\vec{p}) = 0$. In this case, a buyer selects a particular such seller s with probability $1/(\tau_s(\vec{p}) + 1)$. Therefore,

$$h_{s,B}(\vec{p}) = \frac{1}{\tau_s(\vec{p}) + 1}\, \delta_{\lambda_s(\vec{p}),0} \tag{5.5}$$

where $\delta_{i,j}$ is the Kronecker delta function, equal to 1, whenever $i = j$, and 0, otherwise.

Lastly, we determine the probability that a buyer of type C select a seller s. This probability is the product of the probability that seller s is one of the two sellers randomly selected by the buyer and the conditional probability that $p_s < p_{s'}$, given that the random pair is s and s'. The probability that seller s is one of the two sellers randomly selected by the buyer is simply the number of pairs including seller s divided by the number of ways in which two sellers can be chosen: *i.e.*, $(S-1)/\binom{S}{2} = 2/S$. Now the probability that $p_s < p_{s'}$, given that the random pair is s and s', is given by:

$$1\left(\frac{\mu_s(\vec{p})}{S-1}\right) + \frac{1}{2}\left(\frac{\tau_s(\vec{p})}{S-1}\right) + 0\left(\frac{\lambda_s(\vec{p})}{S-1}\right) \tag{5.6}$$

In words, this expression states that the probability that seller s is selected depends on three underlying probabilities as follows: (i) the probability that $p_s < p_{s'}$, namely $\mu_s(\vec{p})/(S-1)$, in which case seller s is selected with probability 1, (ii) the probability that $p_s = p_{s'}$, namely $\tau_s(\vec{p})/(S-1)$, in which case seller s is selected with probability 1/2, and (iii) the probability that $p_s < p_{s'}$, namely $\lambda_s(\vec{p})/(S-1)$, in which case seller s is selected with probability 0. Therefore,

$$h_{s,C}(\vec{p}) = \frac{2}{S}\left[\frac{\mu_s(\vec{p})}{S-1} + \frac{1}{2}\left(\frac{\tau_s(\vec{p})}{S-1}\right)\right] = \frac{2\mu_s(\vec{p}) + \tau_s(\vec{p})}{S(S-1)} \tag{5.7}$$

The preceding results can be assembled to express the profit function π_s for seller s in terms of the distribution of strategies and valuations within the buyer population. In particular, assuming all buyers have the same valuation v, such that $g(p_s)$ is a step function, if all sellers also have the same cost r, then

$$\pi_s(\vec{p}) = \begin{cases} (p_s - r)h_s(\vec{p}) & \text{if } p_s \leq v \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

where

$$h_s(\vec{p}) = w_A\frac{1}{S} + w_B\frac{1}{\tau_s(\vec{p})+1}\,\delta_{\lambda_s(\vec{p}),0} + w_C\frac{2\mu_s(\vec{p}) + \tau_s(\vec{p})}{S(S-1)} \tag{5.9}$$

In the next section, we present graphical representations of Eqs. 5.8 and 5.9 under varying assumptions about the distribution of the buyer population.

5.2.1 Profit Landscapes

With the goal in mind of deriving the price vectors that arise when all sellers aim to maximize profits (*i.e.*, the game-theoretic equilibria), we now discuss the topography of *profit landscapes*. A profit landscape is an S-dimensional plot of the sellers' profit functions, as computed in Eqs. 5.8 and 5.9, given the distribution of buyer demand. For example, letting $S = 5$ yields a 5-dimensional profit landscape of which Fig. 5.2 depicts 1-dimensional projections. These plots are generated by assigning 4 of the 5 sellers random prices, and then computing the profits of the remaining seller at all his possible price points given the others' prices and given various buyer distributions: $(w_A, w_B, w_C) \in \{(1,0,0), (1/2,1/2,0), (1/4,1/4,1/2)\}$.

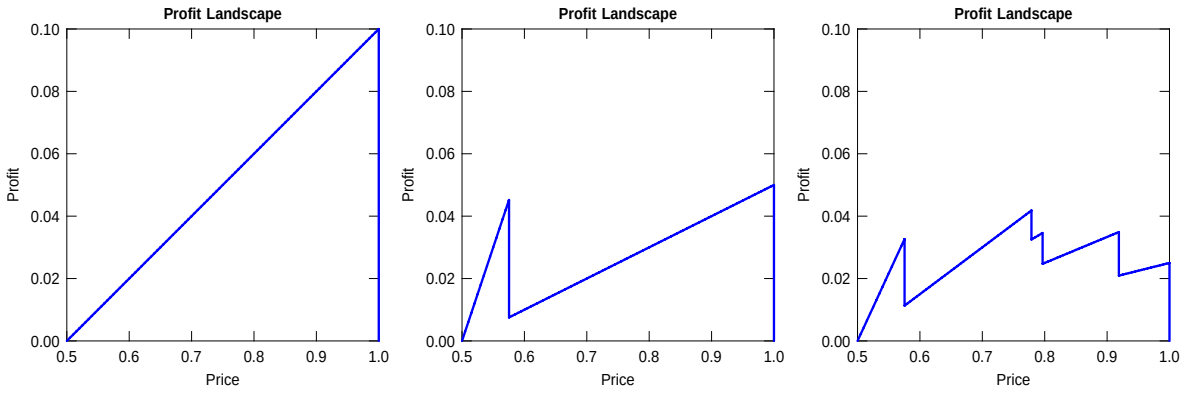


Figure 5.2: 1-dimensional projections of profit landscapes; $v = 1.0, r = 0.5$, $S = 5$, and (w_A, w_B, w_C) = (a) (1,0,0); (b) (1/2, 1/2, 0); (c) (1/4, 1/4, 1/2).

In Fig. 5.2(a), where all buyers are random shoppers, there is a unique maximum, implying that it is in the seller's best interest to charge the buyers' valuation v, since none of the buyers bother to search for lower prices. Fig. 5.2(b) depicts the profit landscape given that half the buyer population uses shopbots, buying only from lowest-priced sellers, while the other half buys from any seller. This figure reveals that the seller in question does well to set its price just below the lowest-priced among the others, but in fact such a price yields profits below those that are obtained by simply setting its price at v. Finally, Fig. 5.2(c) considers the presence of compare pair buyers in the market, in addition to shopbots and random shoppers. Such buyers create incentives for sellers to charge middle-of-the-road prices, since a seller need not be priced strictly below *all* others to attract buyers of this nature, as described by Eq. 5.7. In this diagram, profits are maximized by just undercutting the second lowest-priced alternative seller.

In the next section, we derive the price vectors that comprise the game-theoretic equilibria, given an exogenously determined buyer distribution. Later, we consider a series of learning processes by which the sellers' prices may or may not converge to these game-theoretic equilibrium prices.

5.3 Analysis

In this section, we determine the game-theoretic equilibrium price vectors, under various assumptions about the exogenous distribution of the buyer population. Recall that $B \gg S$; in particular, the number of buyers is assumed to be very large, while the number of sellers is a great deal smaller. In accordance with this assumption, it is reasonable to consider the strategic decision-making of the sellers alone, since their relatively small number suggests that the behavior of individual sellers indeed influences market dynamics, while the large number of buyers renders the effects of individual buyers' actions negligible. We first argue that there is no pure strategy Nash equilibrium, and we then derive the symmetric mixed strategy Nash equilibrium.

A Nash equilibrium is a vector of prices $\vec{p}^*$ at which sellers' profits are maximized given the other sellers' prices; in particular, no seller has any incentive to deviate from its said price [83]. We consider the case in which the number of sellers $S > 1$ and all sellers have identical production costs given by r. Moreover, we assume all buyers have the identical valuations v, with $v > r$. Throughout this exposition, we adopt the following notation: $\vec{p} = (p_s, p_{-s})$, which distinguishes the price offered by seller s from the prices offered by the remaining sellers.

Traditional economic models consider the case in which all buyers are bargain hunters: *i.e.*, $w_B = 1$, and $w_A = w_C = 0$. In this case, prices are driven down to marginal cost; in particular, $p_s^* = r$, for all sellers s. (For an explanation of this phenomenon, see, for example, Tirole [108]). Similarly, in the event that $w_C = 1$ and $w_A = w_B = 0$, the equilibrium price vector that results is again $p_s^* = r$, for all sellers s. Clearly, no seller sets its equilibrium price below c, since doing so yields negative profits. Moreover, no two sellers set their prices at either equal or unequal prices strictly greater than c at equilibrium. If, on the one hand, $p_s^* = p_{s'}^* > c$, then $\pi_s(p_s^* - \epsilon, p_{-s}^*) > \pi_s(p_s^*, p_{-s}^*)$, for small, positive values of ϵ,[4] and similarly for s', thereby negating the assumption that either p_s^* or $p_{s'}^*$ are equilibrium prices. On the other hand, if $p_s^* > p_{s'}^* > c$, then $\pi_s(p_{s'}^* - \epsilon, p_{-s}^*) > \pi_s(p_s^*, p_{-s}^*)$, which implies that p_s^* is not an equilibrium price. Therefore, if either $w_B = 1$ or $w_C = 1$, then all sellers set their equilibrium prices at cost, and earn zero profits.

[4] Unless otherwise noted, ϵ is assumed to be small and positive: *i.e.*, $0 < \epsilon < 1$.

In contrast, consider the case in which all buyers are of type A, meaning that they randomly select a potential seller: *i.e.*, $w_A = 1$ and $w_B = w_C = 0$. In this situation, tacit collusion arises, in which all sellers charge the monopolistic price equal to buyer valuations in the absence of explicit coordination; in particular, $p_s^* = v$, for all sellers s. The argument is straightforward. First note that $p_s^* \geq v$, for all s, since if $p_s^* < v$ for some s, then $\pi_s(p_s^* + \epsilon, p_{-s}^*) > \pi_s(p_s^*, p_{-s}^*)$, in which case p_s^* does not maximize profits. On the other hand, $p_s^* \leq v$, for all s, since if $p_s^* > v$ for some s, then $\pi_s(v, p_{-s}^*) > \pi_s(p_s^*, p_{-s}^*) = 0$, in which case again p_s^* does not maximize profits. Therefore, the monopolistic price $p_s^* = v$ prevails as the equilibrium price charged by all sellers.

Of particular interest in this study, however, are the price dynamics that result from a mixture of buyer types. In the following section, we consider the special case in which $0 < w_A, w_B < 1$, but $w_C = 0$; in other words, there are bargain hunters and there are buyers who consider sellers at random, but there are no compare pair buyers. Knowing that buyers of type A alone results in all sellers charging valuation v, we investigate the impact of buyers of type B, or shopbots, on the marketplace. We find that there are no pure strategy Nash equilibria in such an economy.[5] There does, however, exist a symmetric mixed strategy Nash equilibrium, which we derive in Section 5.3.2. The average equilibrium prices paid by buyers of various types are analyzed in Section 5.3.3. Although equilibrium prices remain above cost, the presence of shopbots in the marketplace leads to decreases in average prices for all buyers. We conclude that by decreasing the cost of obtaining information, shopbots have the potential to ameliorate market inefficiencies.

The material in this chapter is limited to endogenous decisions on the part of the sellers (either game-theoretic or via adaptive learning), with the behavior of buyers determined exogenously. In related work [67], we analyze endogenous buyer and seller decisions from a game-theoretic perspective. In future work, we intend to study the dynamics of interaction among adaptive buyers and sellers.

[5] The general case in which $w_A, w_B, w_C > 0$ is treated in Appendix A, where it is shown that there are again no pure strategy Nash equilibria in an economy of buyers of types A, B, and C.

5.3.1 Special Case: No Compare Pair Buyers

We begin our analysis of the special case which excludes compare pair buyers with the following observation: *at equilibrium, at most one seller s charges $p_s^* < v$.* Suppose that two distinct sellers $s' \neq s$ set their equilibrium prices to be $p_{s'}^* = p_s^* < v$, while all other sellers set their equilibrium prices at the buyers' valuation v. In this case, $\pi_s(p_s^* - \epsilon, p_{-s}^*) = [(1/S)w_A + w_B](p_s^* - \epsilon - r) > [(1/S)w_A + (1/2)w_B](p_s^* - r) = \pi_s(p_s^*, p_{-s}^*)$, which implies that p_s^* is not an equilibrium price for seller s. Now suppose that two distinct sellers $s' \neq s$ set their equilibrium prices to be $p_{s'}^* < p_s^* < v$, while all other sellers set their equilibrium prices precisely at v. In this case, seller s prefers price v to p_s^*, since $\pi_s(v, p_{-s}^*) = [(1/S)w_A](v - r) > [(1/S)w_A](p_s^* - r) = \pi_s(p_s^*, p_{-s}^*)$, which implies that p_s^* is not an equilibrium price for seller s. Therefore, at most one seller charges $p_s^* < v$.

On the other hand, *at equilibrium, at least one seller s charges $p_s^* < v$.* Given that all sellers other than s set their equilibrium prices at v, seller s maximizes its profits by charging price $v - \epsilon$, since $\pi_s(v - \epsilon, p_{-s}^*) = [(1/S)w_A + w_B](v - \epsilon - r) > [(1/S)(w_A + w_B)](v - r) = \pi_s(v, p_{-s}^*)$. Thus v is not an equilibrium price for seller s. It follows from these two observations that at equilibrium, exactly one seller s sets its price below the buyers' valuation v, while all other sellers $s' \neq s$ set their equilibrium prices $p_{s'}^* \geq v$. Note, however, that $\pi_{s'}(v, p_{-s'}^*) = [(1/S)w_A](v - r) > 0 = \pi_{s'}(v', p_{-s'}^*)$, for all $v' > v$, if $w_A > 0$, implying that all other sellers s' maximize their profits by charging price v. Thus, the unique form of pure strategy equilibrium which arises in this setting requires that a single seller s set its price $p_s^* < v$ while all other sellers $s' \neq s$ set their prices $p_{s'}^* = v$.

The price vector (p_s^*, p_{-s}^*), with $p_{-s}^* = (v, \ldots, v)$, however, is not in fact a Nash equilibrium. While v is an optimal response to p_s^*, since the profits of seller $s' \neq s$ are maximized at v given that there exists low-priced seller s, p_s^* is not an optimal response to v. On the contrary, $\pi_s(p_s^*, v, \ldots, v) < \pi_s(p_s^* + \epsilon, v, \ldots, v)$. In particular, the low-priced seller s has incentive to deviate. It follows that there is no pure strategy Nash equilibrium in the proposed model of shopbot economics, whenever $0 < w_A, w_B < 1$ and $w_C = 0$.

5.3.2 General Case: Symmetric Nash Equilibrium

There does, however, exist a symmetric mixed strategy Nash equilibrium, which we derive presently. We resort to the more general economic nomenclature in which buyers of type i ($0 \leq i \leq S$) search among i sellers chosen at random, and we derive the symmetric mixed strategy Nash equilibrium given an exogenous buyer distribution $(w_1, \ldots, w_S)$. Let $f(p)$ denote the probability density function according to which sellers set their equilibrium prices, and let $F(p)$ be the corresponding cumulative distribution function. Following Varian [110], we note that in the range for which it is defined, $F(p)$ has no mass points, since otherwise a seller could decrease its price by an arbitrarily small amount and experience a discontinuous increase in profits. Moreover, there are no gaps in the distribution, since otherwise prices would not be optimal — a seller charging a price at the low end of the gap could increase its price to fill the gap while retaining its market share, thereby increasing its profits.

The cumulative distribution function $F(p)$ is computed in terms of the probability $h_s(\vec{p}, \vec{w})$ that buyers select seller s as their potential seller. This quantity is the weighted sum of $h_{s,i}(\vec{p})$ over $0 \leq i \leq S$. The first component $h_{s,0}(\vec{p}) = 0$. Consider the next component $h_{s,1}(\vec{p})$. Buyers of type 1 select sellers at random; thus, the probability that seller s is selected by such buyers is simply $h_{s,1}(\vec{p}) = 1/S$. Now consider buyers of type 2. In order for seller s to be selected by a buyer of type 2, s must be included within the pair of sellers being sampled — which occurs with probability $(S-1)/\binom{S}{2} = 2/S$ — and s must be lower in price than the other seller in the pair. Since, by the assumption of symmetry, the other seller's price is drawn from the same distribution, this occurs with probability $1 - F(p)$. Therefore $h_{s,2}(\vec{p}) = (2/S)\,[1 - F(p)]$. In general, seller s is selected by a buyer of type i with probability $\binom{S-1}{i-1}/\binom{S}{i} = i/S$, and seller s is the lowest-priced among the i sellers selected with probability $[1 - F(p)]^{i-1}$, since these are $i - 1$ independent events. Thus, $h_{s,i}(\vec{p}) = (i/S)[1 - F(p)]^{i-1}$, and moreover,[6]

$$h_s(p) = \frac{1}{S} \sum_{i=1}^{S} i w_i [1 - F(p)]^{i-1} \tag{5.10}$$

[6] In Eq. 5.10, $h_s(p)$ is expressed as a function of seller s's scalar price p, given that probability distribution $F(p)$ describes the other sellers' expected prices.

The precise value of $F(p)$ is determined by noting that a Nash equilibrium in mixed strategies requires that all pure strategies that are assigned positive probability yield equal payoffs, since otherwise it would not be optimal to randomize. In particular, the expected profits earned by seller s, namely $\pi_s(p) = h_s(p)(p - r)$ are constant for all prices p. The value of this constant can be computed from its value at the boundary $p = v$; note that $F(v) = 1$ because no rational seller charges more than any buyer is willing to pay. This leads to the following relation:

$$h_s(p)(p - r) = h_s(v)(v - r) = \frac{1}{S} w_1(v - r) \tag{5.11}$$

Combining Eqs. 5.10 and 5.11, and solving for p in terms of F yields:

$$p(F) = r + \frac{w_1(v - r)}{\sum_{i=1}^{S} i w_i [1 - F]^{i-1}} \tag{5.12}$$

Eq. 5.12 has several important implications. First of all, in a population in which there are no buyers of type 1 (*i.e.*, $w_1 = 0$) the sellers charge the production cost c and earn zero profits; this is the traditional Bertrand equilibrium. On the other hand, if the population consists of just two buyer types, 1 and some $i \neq 1$, then it is possible to invert $p(F)$ to obtain:

$$F(p) = 1 - \left[\left(\frac{w_1}{i w_i} \right) \left(\frac{v - p}{p - r} \right) \right]^{\frac{1}{i-1}} \tag{5.13}$$

The case in which $i = S$ was studied previously by Varian [110]; in this model, buyers either choose a single seller at random (type 1) or search all sellers and choose the lowest-priced among them (type S).

Since $F(p)$ is a cumulative probability distribution, it is only valid in the domain for which its valuation is between 0 and 1. As noted previously, the upper boundary is $p = v$; the lower boundary p^* can be computed by setting $F(p^*) = 0$ in Eq. 5.12, which yields:

$$p^* = r + \frac{w_1(v - r)}{\sum_{i=1}^{S} i w_i} \tag{5.14}$$

In general, Eq. 5.12 cannot be inverted to obtain an analytic expression for $F(p)$. It is possible, however, to plot $F(p)$ without resorting to numerical root finding techniques. We use Eq. 5.12 to evaluate p at equally spaced intervals in $F \in [0, 1]$; this produces unequally spaced values of p ranging from p^* to v.

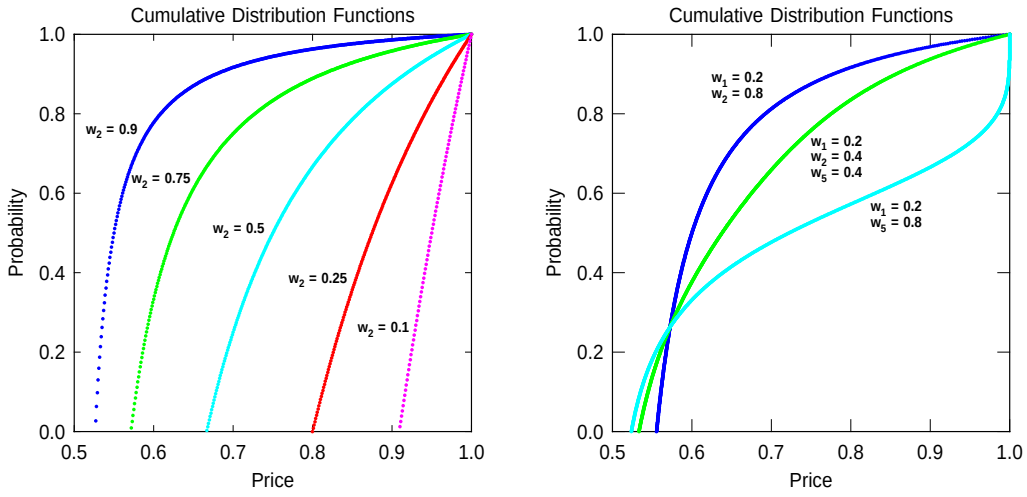


Figure 5.3: (a) CDFs: 2 sellers, $w_1 + w_S = 1$. (b) CDFs: 5 sellers, $w_1 = 0.2$, $w_2 + w_S = 0.8$.

Fig. 5.3 depicts the cumulative distribution functions that arise as the symmetric mixed strategy Nash equilibrium of the prescribed model under varying distributions of buyer strategies. For buyer valuation $v = 1$ and seller cost $r = 0.5$, Fig. 5.3(a) plots 5 CDFs for 5 values of w_1 and w_2, assuming 2 sellers, while Fig. 5.3(b) plots 3 CDFs given $w_1 = 0.2$ and $w_2 + w_S = 0.8$ assuming 5 sellers. We now turn our attention to the probability density function $f(p)$; after deriving and plotting the corresponding PDFs we include further explanation of the distinguishing features of this Nash equilibrium.

Differentiating both (extreme) sides of Eq. 5.11 with respect to p, and substituting Eq. 5.10, we obtain an expression for $f(p)$ in terms of $F(p)$ and p that is conducive to numerical evaluation:

$$f(p) = \frac{w_1(v-r)}{(p-r)^2 \sum_{i=2}^{S} i(i-1)w_i[1-F(p)]^{i-2}} \tag{5.15}$$

The values of $f(p)$ at the boundaries p^* and v are as follows:

$$f(p^*) = \frac{\left[\sum_{i=1}^{S} iw_i\right]^2}{w_1(v-r)\left[\sum_{i=2}^{S} i(i-1)w_i\right]} \quad \text{and} \quad f(v) = \frac{w_1}{2w_2(v-r)} \tag{5.16}$$

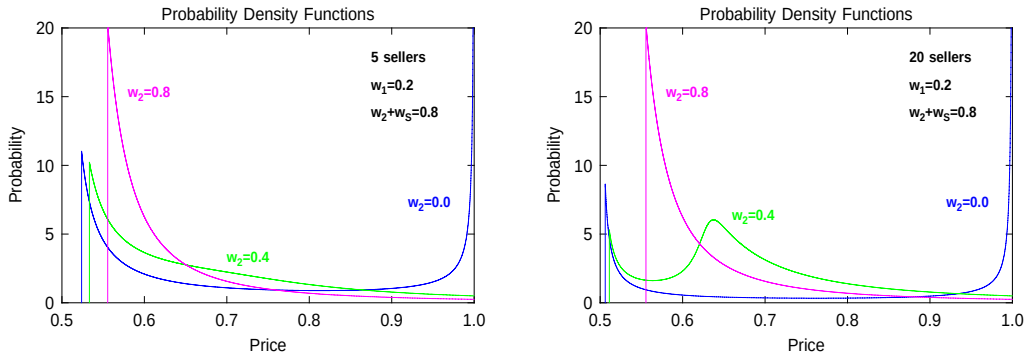


Figure 5.4: PDFs: $w_1 = 0.2$, $w_2 + w_{20} = 0.8$.

Fig. 5.4(a) and 5.4(b) depict the PDFs in the prescribed model under varying distributions of buyer strategies — in particular, $w_1 = 0.2$ and $w_2 + w_S = 0.8$ — when $S = 5$ and $S = 20$, respectively, given $v = 1$ and $r = 0.5$. In both figures, $f(p)$ is bimodal when $w_2 = 0$, as is derived in Eq. 5.16. Most of the probability density is concentrated either just above p^*, where sellers expect low margins but high volume, or just below v, where they expect high margins but low volume. In addition, moving from $S = 5$ to $S = 20$, the boundary p^* decreases, and the area of the no-man's land between these extremes diminishes. In contrast, when $w_2, w_S > 0$, a peak appears in the distribution. If a seller does not charge the absolute lowest price when $w_2 = 0$, then it fails to obtain sales from any buyers of type S. In the presence of buyers of type 2, however, sellers can obtain increased sales even when they are priced moderately. Thus, there is an incentive to price in this manner, as is depicted by the peak in the distribution.

Recall that the profit earned by each seller is $(1/S)w_1(v - r)$, which is strictly positive so long as $w_1 > 0$. It is as though only buyers of type 1 are contributing to sellers' profits, although the actual distribution of contributions from buyers of type 1 vs. buyers of type $i > 1$ is not as one-sided as it appears. In reality, buyers of type 1 are charged less than v on average, and buyers of type $i > 1$ are charged more than r on average, although total profits are equivalent to what they would be if the sellers practiced perfect price discrimination. In effect, buyers of type 1 exert negative externalities on buyers of type $i > 1$, by creating surplus profits for sellers.

5.3.3 Shopbot Savings

We now analyze the equilibrium distributions of prices paid by buyers of various types and their corresponding averages in order to quantify the benefit to buyers of shopbots in the marketplace. Recall that a buyer who obtains i price quotes pays the lowest of the i prices. (At equilibrium, the sellers' prices never exceed v since $F(v) = 1$, so a buyer is *always* willing to pay the lowest price.) The cumulative distribution for the minimal values of i independent samples taken from the distribution $f(p)$ is given by

$$Y_i(p) = 1 - [1 - F(p)]^i \tag{5.17}$$

Differentiation with respect to p yields the probability distribution:

$$y_i(p) = i f(p)[1 - F(p)]^{i-1} \tag{5.18}$$

The average price for the distribution $y_i(p)$ can be expressed as follows:

$$\begin{aligned} \bar{p}_i &= \int_{p^*}^{v} dp\, p\, y_i(p) \\ &= v - \int_{p^*}^{v} dp\, Y_i(p) \\ &= p^* + \int_0^1 dF\, \frac{(1-F)^i}{f} \end{aligned} \tag{5.19}$$

where the first equality is obtained via integration by parts, and the second depends on the observation that $dp/dF = [dF/dp]^{-1} = 1/f$. Combining Eqs. 5.12, 5.15, and 5.19 would lead to an integrand expressed purely in terms of F. Integration over the variable F (as opposed to p) is advantageous because F can be chosen to be equispaced, as standard numerical integration techniques require.

Fig. 5.5(a) depicts sample price distributions for buyers of various types: $y_1(p)$, $y_2(p)$, and $y_{20}(p)$, when $S = 20$ and $(w_1, w_2, w_{20}) = (0.2, 0.4, 0.4)$. The dashed lines represent the average prices $\bar{p}_i$ for $i \in \{1, 2, 20\}$ as computed by Eq. 5.19. The blue line labeled *Search–1*, which depicts the distribution $y_1(p)$, is identical to the green line labeled $w_2 = 0.4$ in Fig. 5.4(b), since $y_1(p) = f(p)$. In addition, the distributions shift toward lower values of p for those buyers who base their buying decisions on information pertaining to more sellers.

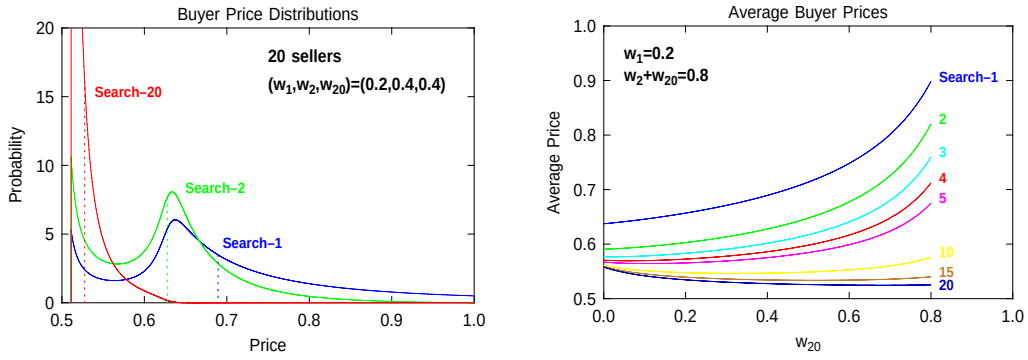


Figure 5.5: (a) Buyer price distributions: 20 sellers, $w_1 = 0.2, w_2 = 0.4$, $w_{20} = 0.4$. (b) Average buyer prices for various types: 20 sellers, $w_1 = 0.2$, $w_2 + w_{20} = 0.8$.

Fig. 5.5(b) depicts the average buyer prices obtained by buyers of various types, when w_1 is fixed at 0.2 and $w_2 + w_{20} = 0.8$. The various values of i are listed to the right of the curves. Notice that as w_{20} (*i.e.*, the number of shopbot users) increases, the average prices paid by those buyers who perform relatively few searches increases, rather dramatically for large values of w_{20}. This is because w_1 is fixed, which implies that the sellers' profit surplus is similarly fixed; thus, as more and more buyers perform extensive searches, the average prices paid by those buyers decreases, which causes the average prices paid by the less diligent searchers to increase. The situation is slightly different for those buyers who perform more searches but do not search the entire space of sellers: *e.g.*, $i = 10$ and $i = 15$. These buyers initially reap the benefits of increasing the use of shopbots, but eventually their average prices increase as well. Given a fixed portion of the population designated as random shoppers, Fig. 5.5(b) demonstrates that searching S sellers is a superior buyer strategy to searching only $1 < i < S$ sellers. Thus, there are savings to be had via price searches. Moreover, *shopbots offer added value* in markets in which there exist buyers who are willing to purchase from any seller.

5.4 Simulations

When sufficiently widespread adoption of shopbots by buyers forces sellers to become more competitive, it seems likely that sellers will respond by creating *pricebots* that mechanically set prices in attempt to maximize profitability. It is unrealistic, however, to expect that pricebots will simply compute the mixed strategy Nash equilibrium and distribute their prices accordingly. The business world is fraught with uncertainties that undermine the validity of traditional game-theoretic analyses: sellers lack perfect knowledge of buyer demands, and have an incomplete understanding of competitors' strategies. In order to be profitable for sellers, pricebots will need to continually adapt to changing market conditions. In this section, we present the results of simulations of various adaptive pricing strategies, and we compare the resulting price and profit dynamics with the game-theoretic equilibrium.

5.4.1 Simple Adaptive Pricing Strategies

We introduce three simple adaptive pricing strategies, each of which makes different demands on the required levels of informational and computational power of agents.

GT The *game-theoretic* strategy is designed to reproduce the mixed strategy Nash equilibrium computed in the previous section, provided that it is adopted by all sellers. It makes use of full information about the buyer population, and assumes that its competitors also use the GT strategy. It therefore generates a price chosen randomly from the probability density function derived in Section 5.3.2.

MY The *myoptimal*[7] (myopically optimal) strategy [70, 68, 92] makes use of full information about all characteristics that factor into buyer demand, as well as the competitors' prices, but makes no attempt to account for competitors' pricing strategies. Instead, it is based on the assumption of static expectations: even if one seller is contemplating a price change under myoptimal pricing, this seller does not assume that this will elicit a response from its competitors; instead it is assumed that competitors' prices will remain fixed.

[7] In the game-theoretic literature, this adaptive strategy is known as Cournot best-reply dynamics [24].

The myoptimal seller s uses all of the available information and the assumption of static expectations to perform an exhaustive search for the price p_s^* that maximizes its expected profit π_s. In our simulations, we compute π_s according to Eqs. 5.8 and 5.9, as is depicted in Fig. 5.2. The optimal price p_s^* is guaranteed to be either the valuation v or ϵ below some competitor's price, where ϵ is the *price quantum*, or the smallest amount by which one seller may undercut another, set to 0.002 in these simulations. This limits the search for p_s^* to S possible values.

DF The *derivative-following* strategy is less informationally intensive than either the myoptimal or game-theoretic pricing strategies. In particular, this strategy can be used in the absence of any knowledge or assumptions about one's competitors or the buyer demand function. A derivative follower simply experiments with incremental increases (or decreases) in its price, continuing to move its price in the same direction until the observed profitability level falls, at which point the direction of movement is reversed. The price increment δ is chosen randomly from a specified probability distribution; in the simulations described here the distribution was uniform between 0.01 and 0.02.

Simple Price and Profit Dynamics

We have simulated an economy with 1000 buyers and 5 sellers employing various mixtures of pricing strategies. In the simulations described below, a buyer's valuation of the good $v = 1.0$, and a seller's production cost $r = 0.5$. The mixture of buyer types is $w_A = 0.2$, $w_B = 0.4$, and $w_C = .4$: *i.e.*, 20% of the buyers visit sellers randomly, 40% are bargain hunters,[8] and the remaining 40% are compare pair shoppers.

The simulation is asynchronous: at each time step, a buyer or seller is randomly selected to carry out an action (*e.g.*, buying an item or resetting a price). The chance that a given agent is selected for action is determined by its rate; the rate ρ_b at which a given buyer b attempts to purchase the good is set to 0.001, while rate ρ_s at which sellers reconsider their prices is set to 0.00005 (except where otherwise specified). Each simulation was iterated for 10 million time steps, during which time we observe a total of approximately 9.9975 million purchases and 500 price adjustments per seller.

[8] Recall that bargain hunters in this model represent the use of shopbots.

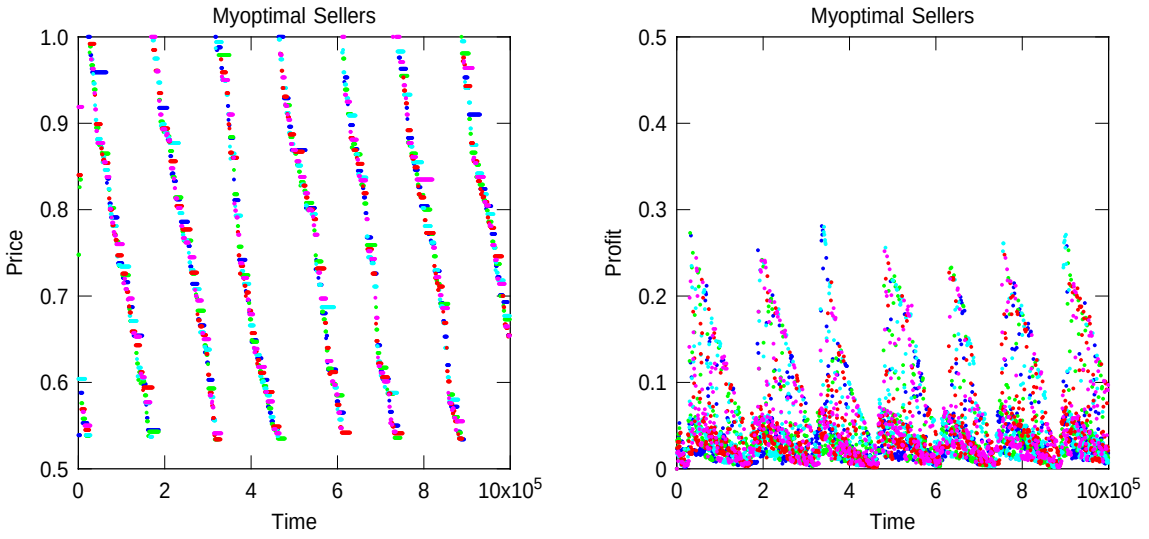


Figure 5.6: 1000 buyers and 5 MY sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Profit vs. time.

GT Pricing Strategy

Simulations verify that, if agents are game-theoretic strategists, then the cumulative distribution of prices closely resembles the derived $F(p)$ (to within statistical error). Moreover, the time-averaged profit of the sellers is $\bar{\pi} = 0.0202 \pm 0.0003$, which is nearly the theoretical value of 0.02. (Not shown.)

MY Pricing Strategy

Figure 5.6(a) illustrates the cyclical price wars that occur when all 5 sellers use the myoptimal pricing strategy.[9] Regardless of the initial value of the price vector, it quickly settles into a pattern in which all prices are initially near the monopolistic price $v = 1$, followed by a long episode during which the sellers successively undercut one another by ϵ. During this phase, no two prices differ by more than $(S-1)\epsilon$, and the prices fall linearly with time. Finally, when the lowest-priced seller is within ϵ above the value $p^* = 0.5\bar{3}$, the next seller finds it unprofitable to undercut, and instead resets its price to $v = 1$. The other sellers follow suit, until all but the lowest-

[9] The simulations depicted in Figure 5.6 were run for 1 million time steps, with $\rho_s = 0.0002$, yielding roughly 200 price adjustments per seller.

priced seller are charging $v = 1$. At this point, the lowest-priced seller finds that it can maintain its market share but increase its profit per unit dramatically — from $p^* - .5 = 0.0\bar{3}$ to $0.5 - \epsilon$ — by raising its price to $1 - \epsilon$. No sooner does the lowest-priced seller raise its price than the next seller who resets its price undercuts, thereby igniting the next cycle of the price war.

Cyclical price wars and generalized price/product wars have been observed in other models of software agents in information economies [54, 69]. Analysis ascribes this behavior to the topography of the profit landscape (see Section 5.2.1). Price wars and price/product wars occur whenever landscapes contains multiple peaks and agents behave myoptimally. Indeed, the model of shopbots adopted herein gives rise to profit landscapes $\pi_s(\vec{p})$ that are multi-peaked, due to the dependence in Eq. 5.9 upon the functions $\lambda_s(\vec{p})$, $\tau_s(\vec{p})$, and $\mu_s(\vec{p})$, which are discontinuous at any point where two or more sellers set their prices to be equal.

Fig. 5.6(b) depicts the average profits obtained by myoptimal sellers. The expected average profit over one price war cycle is given by the following expression:

$$\pi_s^{\mathrm{MY}} = \frac{1}{S}\left[\frac{1}{2}(v + p^*) - r\right] \tag{5.20}$$

In other words, on average each seller is expected to receive its fair share of a pie which ranges in value from $p^* - r$ to $v - r$. The upper points in Fig. 5.6 correspond to the profits obtained by the sellers who are first to abandon the low prices that prevail near the end of price wars. In fact, these values appear to be roughly 0.2667, which is precisely $(1/2)[(v + p^*) - r]$ for the stated parameters. The linear decrease in profits which follows these peaks reflects the corresponding price dynamics. Throughout, the bulk of the sellers' profits fall below 0.05, but even in this region prices decrease linearly. This is indicative of an attempt by sellers who suffer from low market share to rectify their position in exchange for lower margins.

Since prices fluctuate over time, it is also of interest to compute the distribution of prices. Fig. 5.8(a) plots the cumulative distribution function for myoptimal pricing. This cumulative density function, which is obtained via simulation, has precisely the same endpoints $p^* = 0.5\bar{3}$ and $v = 1$ as those of the derived mixed strategy equilibrium, but the linear shape between those endpoints (which reflects the linear price war) is quite different from what is displayed in Fig. 5.3(b).

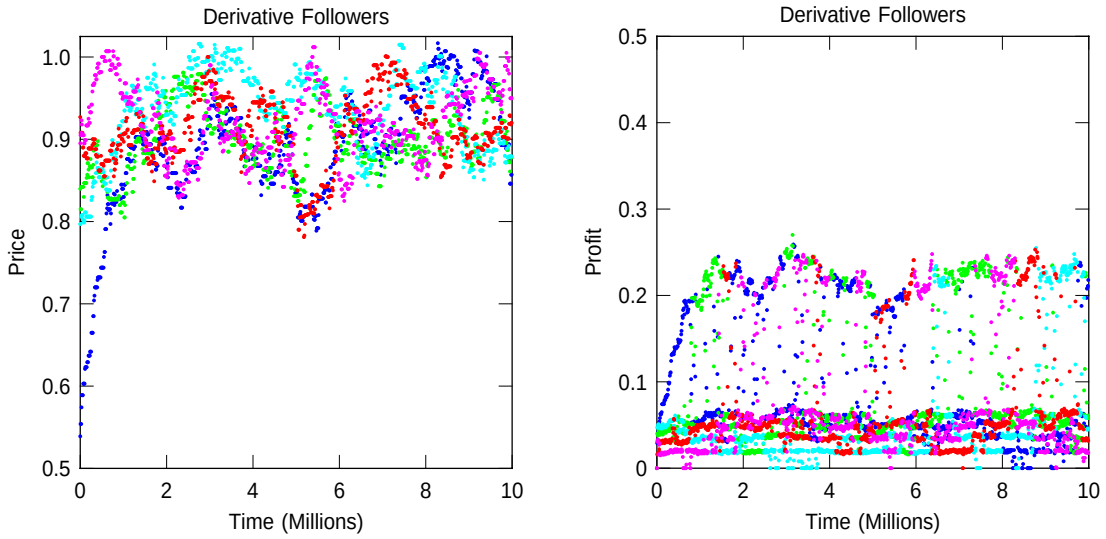


Figure 5.7: 1000 buyers and 5 DF sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Profit vs. time.

DF Pricing Strategy

Figure 5.7 shows the price dynamics that result when 5 derivative followers are pitted against one another. Recall that derivative followers do not base their pricing decisions on any information that pertains to other agents in the system — neither sellers' price-setting tendencies nor buyers' preferences. Nonetheless, their behavior tends towards what is in effect a collusive state in which *all* sellers charge nearly the monopolistic price. This is tacit collusion,[10] so-called because the agents do not communicate at all and there is consequently nothing illegal about their collusive behavior.[11] By exhibiting such behavior, derivative followers accumulate greater wealth than any of the sellers examined thus far. According to Fig. 5.7(b), the seller that is lowest-priced for a time earns profits roughly 0.25, while the others earn profits in the ranging from 0.01 to 0.05; this lower range contains linear profit regions due to the presence of compare pair buyers in the market. On average, the profit per time step for each of the five derivative followers is 0.075.

[10] See, for example, in Tirole [108]

[11] It has similarly been observed in Huck, *et al.* [60] that derivative followers tend towards a collusive outcome in models of Cournot duopoly.

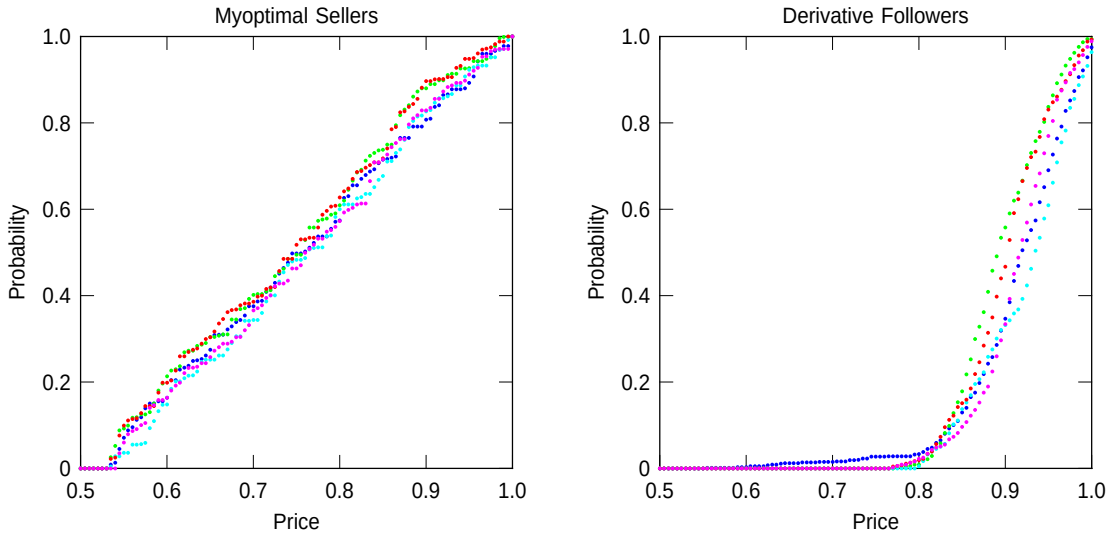


Figure 5.8: (a) CDF for 5 MY sellers. (b) CDF for 5 DF sellers.

How do derivative followers manage to collude? Like myoptimal sellers, derivative followers are capable of engaging in price wars; such dynamics are visible in Fig. 5.7(a). These price wars, however, are easily quelled, making upward trends more likely than downward trends. Imagine for simplicity that $w_A + w_B = 1$, and suppose that X and Y are the two lowest-priced sellers engaged in a mini-price war. Assume X's price is initially above Y's, but that X soon undercuts Y. This yields profits for seller X obtained from the entire population of type B buyers while it is lower-priced, and from its share of type A buyers all throughout. Now suppose Y undercuts X, but soon after X again undercuts Y. This yields profits for seller X once again obtained from the entire population of type B buyers during the period in which it is lower-priced, and from its share of type A buyers all throughout. In other words, given equivalent rates of price adjustment for both sellers, market share remains fixed during mini-price wars of this kind. Thus, the only variable in computing profits is price, leaving sellers with the incentive to increases prices more often than not. In the case in which $w_C > 0$, there is a tendency for sellers other than the two lowest-priced to engage in mini-price wars, but these price wars are once again easily quelled. The tendency of a society of DF sellers to reach and maintain high prices is reflected in the cumulative distribution function, shown in Fig. 5.8(b).

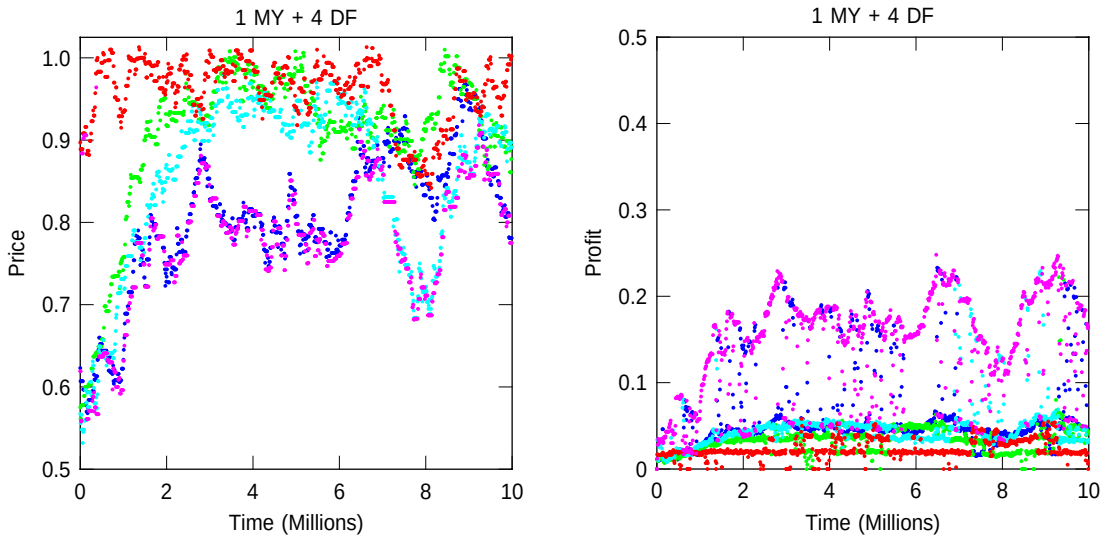


Figure 5.9: 1000 buyers, 1 MY + 4 DF sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Profit vs. time.

MY + DF Pricing Strategies

The fact that a group of derivative followers is able to extract a larger profit than groups of more clever and informed agents (both game-theoretic and myoptimal) may seem paradoxical. How is it that it could be so smart to be so ignorant? Fig. 5.9, in which one of the derivative followers is replaced by a myoptimal seller, suggests that in fact ignorance may not be bliss. During the simulation, the myoptimal seller averages a profit of 0.14! In contrast, 2 of the 4 derivative followers receive an average profit of 0.028, hovering around an average price near v during the bulk of the simulation. The remaining sellers have more interesting experiences: they (separately) engage in head-to-head combat with the myoptimal seller, managing in the process to do better than their cohorts, one obtaining an average profit of 0.047, and the other obtaining an average profit of 0.063. At time approximately 3 million, one of the derivative followers deviates from the established upward trend, finding that remaining lowest-priced in fact yields greater profits. The myoptimal agent immediately follows suit. For the next roughly 4 million time steps, the derivative follower is undercut regularly; however, this is more than compensated for by the additional profit that it achieves during those times when it cashes in by undercutting the myoptimal seller.

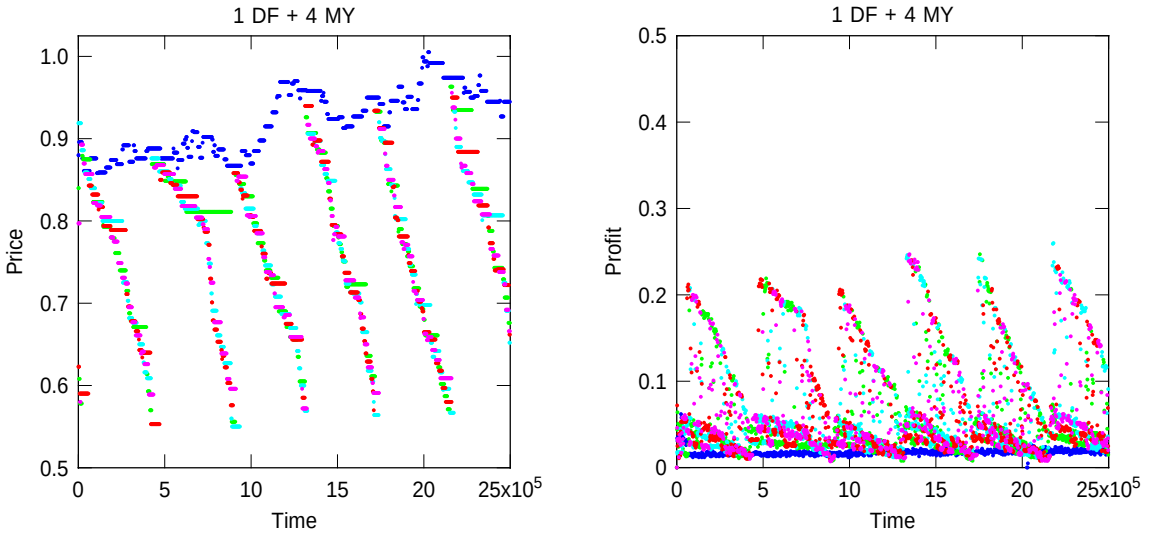


Figure 5.10: 1000 buyers, 1 DF + 4 MY sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Profit vs. time.

On the other hand, Fig. 5.10 depicts the results of simulation of one derivative follower competing with four myoptimal sellers. The presence of even one derivative follower in the market leads to lower profits for the myoptimal sellers than in the case in which all sellers are myoptimal because the derivative follower forms an upper bound that is generally less that v, above which no myoptimal seller is inclined to price. Thus, myoptimal sellers in this simulation do not achieve profits described by Eq. 5.20; they accumulate slightly less wealth, depending on the residing price of the derivative follower at the times in which price wars conclude.

5.4.2 Sophisticated Adaptive Pricing Strategies

In this section, we investigate the behavior of adaptive pricing strategies that fit into the regime of low-rationality learning defined in Chapter 2. These algorithms specify that players *explore* their strategy space by playing all strategies with some non-zero probability, and *exploit* successful strategies by increasing the probability of employing those strategies that generate high payoffs. Of particular interest in this study are the no external regret algorithm of Freund and Schapire [39] and the no internal regret algorithm of Foster and Vohra [37].

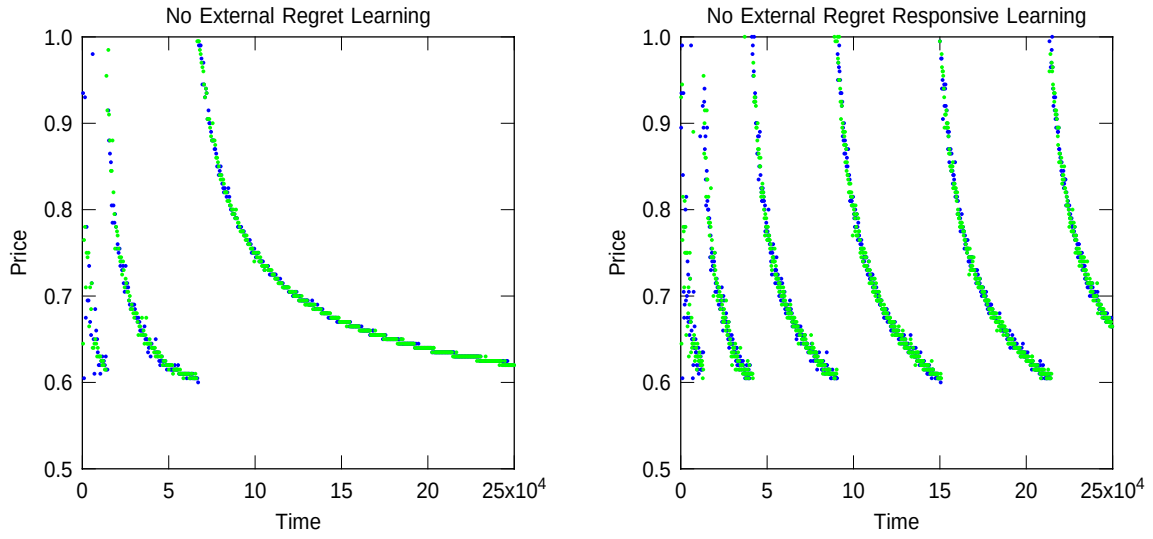


Figure 5.11: 1000 buyers, 2 NER sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Price vs. time, responsive learning.

No External Regret Learning

As described in Chapter 2, there are several algorithms that satisfy the no external regret optimality criteria: *e.g.*, Foster and Vohra [36] and Freund and Schapire [39]. This section discusses simulations of the latter in shopbot economics. We consider 2 no external regret sellers, with learning rate $\beta = 0.1$, given buyer distribution $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. We find that although this algorithm has been observed to converge to Nash equilibrium in games of two strategies (*e.g.*, the Santa Fe bar problem, see Greenwald, *et al.* [52]), play cycles exponentially in this game of many strategies[12] (see Fig. 5.11(a)). In fact, the outcome of play of no external regret learning in the prescribed model is reminiscent of its outcome, and the similar outcome of fictitious play, in the Shapley game (see Greenwald, *et al.* [50]). Fig. 5.11(b) depicts simulations of no external regret learning in which we limit the length of the cycles via a responsiveness parameter, namely $\gamma = 0.00005$. For appropriate choices of γ, the long term empirical frequency of plays is indeed near-Nash equilibrium.

[12] Technically, there a continuum of pricing strategies is the prescribed model of shopbot economics. For the purposes of simulating no regret learning, this continuum was discretized into 100 equally sized intervals.

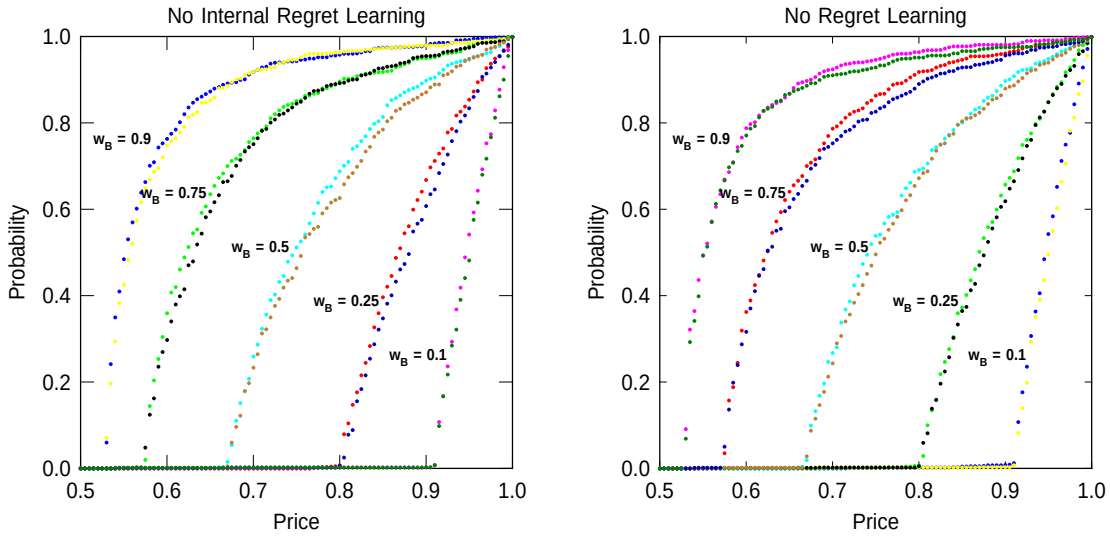


Figure 5.12: 1000 buyers, $w_A + w_B = 1$. (a) Price vs. time, 2 NIR sellers. (b) Price vs. time, 1 NER seller and 1 NIR seller.

No Internal Regret Learning

Learning that satisfies the no internal regret optimality criteria is known to converge to correlated equilibrium [37]. In this section, we investigate the properties of no internal regret learning in shopbot economics, and we observe convergence to the symmetric mixed strategy Nash equilibrium, one particular solution contained within the set of correlated equilibria. Despite several negative theoretical results on the rational learning of Nash equilibrium (*e.g.*, Foster and Young [38], Nachbar [81], and Greenwald, *et al.* [52]), in practice, sophisticated low-rationality algorithms tend to learn Nash equilibrium. Fig 5.12(a) depicts the results of simulations of the no internal regret learning algorithm due to Foster and Vohra [37]. In these simulations, there are 2 no internal regret sellers, and the buyer distributions range from $(w_A, w_B, w_C) = (0.1, 0.9, 0.0)$ to $(w_A, w_B, w_C) = (0.9, 0.1, 0.0)$; there are no compare pair buyers. These plots match the theoretical Nash equilibria depicted in Fig. 5.3(a). Similarly, the plots depicted in Fig 5.12(b), which convey the results of simulations of 1 no external regret learner and 1 no internal regret learner also match the theoretical Nash equilibria. It remains to further investigate the collective dynamics of myoptimal behavior, derivative following, and no regret learning.

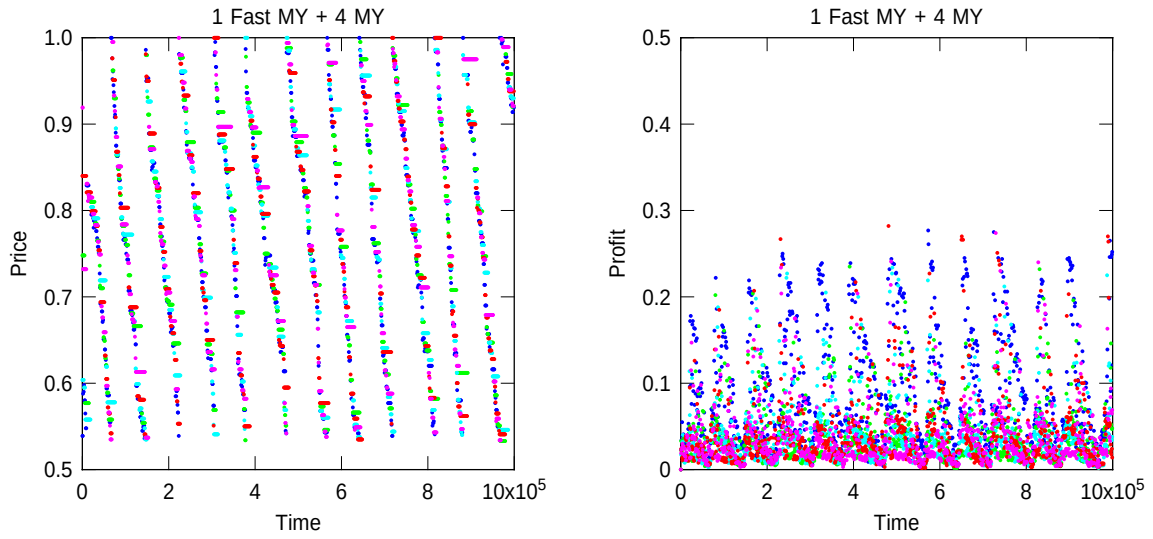


Figure 5.13: 1000 buyers, 1 Fast MY + 4 MY sellers; $(w_A, w_B, w_C) = (0.2, 0.4, 0.4)$. (a) Price vs. time. (b) Profit vs. time.

5.5 Discussion: Evolution of Shopbots and Pricebots

We now revisit the situation in which all five sellers utilize the myoptimal pricing strategy, but we allow one of the sellers to reset its price 5 times as quickly as the others. These price dynamics are illustrated in Fig. 5.13(a). As in Fig. 5.6(a), the sellers engage in price wars; in this case, however, they are accelerated, which is apparent from the increased number of cycles that occur during the simulation. The plots in Fig. 5.13(b) reveal that the fast myoptimal agent accumulates substantially more wealth than the others. The reason for this is simply that it undercuts far more often than it itself is undercut, and therefore maintains substantial market share.

It is evident from Fig. 5.13(b) that myoptimal sellers prefer to reset their prices faster rather than slower, particularly given a large proportion of shopbots: *i.e.*, bargain hunters. This could potentially lead to an arms race, in which sellers re-price their goods with ever-increasing frequency, resulting in arbitrarily fast price-war cycles. This observation is not specific to myoptimal agents. In additional simulations, we have observed sufficiently fast derivative followers who obtain the upper hand over slower myoptimal agents. In the absence of any throttling mechanism, it is advantageous for all types of sellers to re-price their goods as quickly as possible.

Let us carry the arms race scenario a bit further. In a world in which sellers reset their prices at ever-increasing rates, a human price setter would undoubtedly be too inefficient, and would quickly be replaced by a pricebot, perhaps a more sophisticated variant of one of the seller strategies proposed in Section 5.4. Almost certainly, this strategy would make use of information about the buyer population, which could be purchased from other agents. Even more likely, however, the strategy would require knowledge of competitors' prices. How would the pricebot obtain this information? From a shopbot, of course!

With each seller seeking to re-price its products faster than its competitors do so, shopbots would quickly become overloaded with requests. Imagine a scenario in which a large player like `amazon.com` were to use the following simple price-setting algorithm: every 10 minutes, submit 2 million or so queries to a shopbot (one for each title carried by `amazon.com`), then charge 1 cent below the minimum price obtained for each title! [13] Since the job of shopbots is to query individual sellers for prices, it would in turn pass this load back to `amazon.com`'s competitors: *e.g.*, `barnesandnoble.com`, `kingbooks.com`, etc. The rate of pricing requests made by sellers could easily dwarf the rate at which similar requests would be made by human buyers, eliminating the potential of shopbots to ameliorate market frictions.

One solution to an excess demand for shopbot services would be for shopbots to charge users for the information they provide. Today, shopbots tend to make a living from advertising revenues. This appears to be an adequate business model so long as requests are made by humans. Agents are unwelcome customers, however, because they are are not generally influenced by advertisements; as a result, agents are either barely tolerated or excluded intentionally. As economically-motivated agents who charge for the information services they provide, shopbots would create the proper incentives to deter excess demand. Such shopbots would gladly welcome business from humans as well as all kinds of autonomous agents, even those acting on behalf of the competition.

[13] It has been argued that shopbots could potentially increase the position of large market players and monopolists, who by utilizing such a strategy could limit the profitability of smaller players [101]. In the Internet domain, however, where small players boasting innovative technologies are springing up regularly, such a strategy would likely be deemed unprofitable.

If shopbots were to begin charging for pricing information, it would seem natural for sellers — the actual owners of the desired information — to themselves charge the shopbots for the use of their information. In turn, the pricing of this pricing information may itself involve the use of dynamic price-setting strategies by a related breed of pricebots. This scenario illustrates how the need for agents to dynamically price their services could quickly percolate through an entire economy of software agents. The alternative is a "meltdown" due to overload which could occur as agents become more prevalent on the Internet. Rules of etiquette followed voluntarily today by web crawlers and other autonomous agents [29] could be trampled in the rush for competitive advantage.

5.6 Related and Future Work

The study of the economics of information was launched in the seminal paper by Stigler [104] in 1961. In this research, Stigler cites several examples of observed price dispersion, which he attributes to costly search procedures faced by consumers. As a result, he notes the utility of trade journals and organizations that specialize in the collection and dissemination of product information, such as *Consumer Reports* and, of course, shopbots. Stigler reminds us that in medieval times, localized marketplaces thrived in spite of heavy taxes that were levied on merchants, demonstrating how worthwhile it was for sellers to participate in localized markets rather than search for buyers individually. Similarly, shopbots today serve as local marketplaces in the global information superhighway, and accordingly, we find sellers sponsoring shopbots and paying commissions on sales, as they essentially pay for the right to participate in the shopbot marketplace. This encapsulates the current business model of firms such as `www.acses.com` that design shopbots.

Since the seminal work of Stigler, many economists have developed and analyzed formal models that attempt to explain the phenomenon of price dispersion.[14] In contrast with our model that assumes an exogenous distribution of buyer types, many authors consider models in which buyers are viewed as rational decision-makers, with

[14] We mention only a handful of papers that make up this large body of literature, but refer the reader to the bibliography included in Hopkins and Seymour [58] for additional sources.

the cost of search defined explicitly, thereby giving rise to endogenous search behavior. For example, assuming buyers are rational, Burdett and Judd [19] find that if buyers use fixed sample size search rules, where cost is linear in the size of search, then if $0 < w_1 < 1$, the dispersed price equilibrium that arises is such that $w_1 + w_2 = 1$; in other words, buyers optimize by searching only once or twice. The assumption that cost increases linearly in the size of a search is not valid in today's world of shopbots, however. In related work [67], we study endogenous buyer decisions assuming non-linear search costs, and we find that if $0 < w_1 < 1$, one dispersed price equilibrium that arises is such that $w_1 + w_2 + w_3 = 1$. Moreover, if we impose a fixed lower bound on the proportion of random shoppers in the population, then depending on initial conditions, the possible equilibria that arise include $w_1 + w_i = 1$, for $2 \leq i \leq S$, given linear search costs. In the future, we plan to extend our model allowing shopbots themselves to act as economic agents, charging users for the information services they provide. This implies an additional strategic variable, namely the cost of search, to be determined dynamically by shopbots attempting to maximize profitability.

Salop and Stiglitz [93] consider a model of monopolistic competition[15] in which buyers are rational decision-makers, and there are two classes of buyers, depending on whether their associated search costs are low or high. In the non-degenerate case, the authors arrive at a two-priced equilibrium in which buyers with low search costs always search, buyers with (sufficiently) high search costs never search, and sellers charge either the competitive price c or the monopolistic price v. In addition, Wilde and Schwartz [115, 116] present a kind of equivalence theorem between a model in which buyers are rational decision-makers and a model in which the distribution of buyers types is exogenous. In particular, they argue that for all equilibria in the exogenous model, there exist information acquisition costs which "rationalize" the exogenous buyer distribution given the endogenous equilibrium prices. Their results lend some justification to the exogenous model that is studied in this chapter, which fails to account for buyer rationality.

[15] Monopolistic competition assumes an infinite number of firms who can freely enter and exit the market. In such a scenario, profits are driven to zero, since firms do not enter if doing so would yield negative profits, while if a firm, say A, earns positive profits, then another firm with lower production costs would enter the market and charge lower prices than A, thereby driving firm A out of business.

The aforementioned literature addresses the issue of spatial price dispersion. On the other hand, Varian [110] considers temporal price dispersion in the marketplace (*i.e.*, sales), and he provides an intuitive justification of this phenomenon as perhaps corresponding to mixed strategy Nash equilibria. In fact, his model closely resembles that which we have studied in this paper in that he assumes an exogenous buyer distribution (without compare pair buyers), but differs in that we do not assume monopolistic competition, and we consider only a finite number of firms. Moreover, we simulate the dynamics of pricebots who engage in price-setting behavior on the part of firms, and we find that the dynamics of sophisticated learning algorithms converge to the derived mixed strategy Nash equilibrium, thereby reproducing the phenomenon of sales in agent interactions. Other models of the dynamic price adjustment process are discussed in Diamond [27] and Hopkins and Seymour [58].

Another recent work of relevance which considers the potential impact of reduced buyer search costs on the electronic marketplace is Bakos [9]. This model is in some sense more general than ours in that it allows for product differentiation, but it does not allow for varying types among buyers. It remains to incorporate features of product differentiation within the present model of shopbot economics.

5.7 Conclusion

A well-known result in the theory of industrial organization, the so-called Bertrand paradox [12], states that the unique equilibrium price in an economy in which several firms produce non-differentiated goods is the marginal cost of production. This result is based on the assumption that in the absence of other forms of product distinction, consumers prefer the lowest-priced good. Our model of shopbot economics depends on more general assumptions in that we allow for the possibility that some consumers purchase from producers at random, or choose the lower-priced between two randomly selected producers; perhaps these consumers are impeded by the cost of searching for the lowest-priced good. These alternative assumptions provide a solution to the Bertrand paradox in which competition does *not* in fact drive prices down to marginal cost. Instead, firms achieve strictly positive profits, which provides some justification for their appearance in markets.

In addition to analytically computing equilibrium prices, our research program is also concerned with the the learning behavior exhibited by computational agents who employ price-setting algorithms on behalf of firms. In particular, do computational agents learn to set prices at or near the derived equilibrium, and do they generate strictly positive profits for the firms which they represent? This chapter investigated several simple and sophisticated price-setting schemes, and concluded that indeed computational agents do learn to price goods in such a way as to yield positive profits, sometimes well above the profits obtained at equilibrium. The situations in which profits are above equilibrium profits, however, are not likely sustainable since such price-setting schemes invite competition. It is our conjecture that in a realistic electronic marketplace inhabited by diverse learning agents, profits would inevitably be driven down to equilibrium profits. In future work, we intend to consider more diversity in the learning algorithms of agents, in an attempt to resolve this issue.

Finally, in this chapter we have investigated some of the likely consequences of a dramatic increase in the amount of information which is readily available to buyers via agentry. It is also of interest to consider the impact of a dramatic increase in the information available to sellers via agentry. Much of the relevant economic literature is highly dependent on the assumption that sellers are unable to distinguish among their various customers (see, for example, Salop and Stiglitz [94]). Companies like `amazon.com`, however, can readily build databases of customer profiles which classify not only likes and dislikes, but moreover, willingness to pay. It would be of interest to study the impact on the marketplace of increasing the information available to sellers via, say, *tastebots*.

A Appendix

In this appendix, it is shown that in the case of buyers of types A, B, and C, there is no pure strategy Nash equilibrium. More specifically, we assume $0 < w_A < 1$; if $w_A = 1$, then the unique equilibrium is the monopoly price, while if $w_B + w_C = 1$, then the unique equilibrium is the competitive price. We proceed by deriving the unique form of a possible pure strategy Nash equilibrium, if one were to exist, but we argue that this is in fact not an equilibrium.

Suppose that the sellers are ordered $s_1, \ldots, s_j, \ldots, s_S$ such that the indices $j < j+1$ whenever equilibrium prices $p_j^* \leq p_{j+1}^*$. First, note that equilibrium prices $p_j^* \in (r, v]$, since $p_j < r$ yields strictly negative profits, while $p_j = r$ and $p_j > v$ yield zero profits, but $p_j = v$ yields strictly positive profits. The following observation describes the form of a pure strategy Nash equilibrium whenever $w_A > 0$: *at equilibrium, no two sellers charge identical prices.*

Case A.1 Initially, suppose that two distinct sellers offer an equivalent lowest price: *i.e.,* $r < p_1^* = p_2^* < p_3^* < \ldots < p_S^* \leq v$. *In this case, seller 1 stands to gain by undercutting seller 2, implying that* p_1^* *is* not *an equilibrium price. In particular,*

$$\begin{aligned} \pi_1(p_1^* - \epsilon, p_{-1}^*) &= \left[\tfrac{1}{S} w_A + w_B + \tfrac{2}{S} w_C\right](p_1^* - \epsilon - r) \\ &> \left[\tfrac{1}{S} w_A + \tfrac{1}{2} w_B + \left(\tfrac{2}{S}\right)\left(\tfrac{S-3/2}{S-1}\right) w_C\right](p_1^* - r) = \pi_1(p_1^*, p_{-1}^*) \end{aligned}$$

Case A.2 Now suppose that two distinct sellers offer an equivalent intermediate price: *i.e.,* $r < p_1^* < \ldots < p_j^* = p_{i+1}^* < \ldots < p_S^* \leq v$. *In this case, seller* j *stands to gain by undercutting seller* $j+1$*, implying that* p_j^* *is* not *an equilibrium price. In particular,*

$$\begin{aligned} \pi_j(p_j^* - \epsilon, p_{-j}^*) &= \left[\tfrac{1}{S} w_A + \left(\tfrac{2(S-i)}{S(S-1)}\right) w_C\right](p_j^* - \epsilon - r) \\ &> \left[\tfrac{1}{S} w_A + \left(\tfrac{2(S-i)-1}{S(S-1)}\right) w_C\right](p_j^* - r) = \pi_j(p_j^*, p_{-j}^*) \end{aligned}$$

Case A.3 Finally, suppose that two distinct sellers offer an equivalent highest price: *i.e.,* $r < p_1^* < \ldots < p_{S-1}^* = p_S^* \leq v$. *In this case, seller* S *stands to gain by undercutting seller* $S-1$*, implying that* p_S^* *is* not *an equilibrium price. In particular,*

$$\begin{aligned} \pi_S(p_S^* - \epsilon, p_{-S}^*) &= \left[\tfrac{1}{S} w_A + \left(\tfrac{2}{S(S-1)}\right) w_C\right](p_S^* - \epsilon - r) \\ &> \left[\tfrac{1}{S} w_A + \left(\tfrac{1}{S(S-1)}\right) w_C\right](p_S^* - r) = \pi_S(p_S^*, p_{-S}^*) \end{aligned}$$

Further, we now observe that seller S charges price v at equilibrium, since for all $p_{S-1}^* < p_S < v$, $\pi_S(v, p_{-S}^*) = \frac{1}{S} w_A (v - r) > \frac{1}{S} w_A (p_S - r) = \pi_S(p_S, p_{-S}^*)$. Therefore, the relevant price vector consists of S distinct prices with $r < p_1^* < \ldots < p_{S-1}^* < p_S^* = v$.

The price vector $(p_1^*, \ldots, p_j^*, \ldots, p_S^*)$, however, is not a Nash equilibrium. While $p_S^* = v$ is in fact an optimal response to p_{-S}^*, since the profits of seller S are maximized at v given that there exists lower priced sellers $1, \ldots, S-1$, p_{S-1}^* is not an optimal response to $p_{-(S-1)}^*$. On the contrary, $\pi_{S-1}(p_1^*, \ldots, p_{S-1}^*, p_S^*) < \pi_{S-1}(p_1^*, \ldots, p_S^* - \epsilon, p_S^*)$. In particular, seller $S-1$ has incentive to deviate. Similarly, $\pi_{S-2}(p_1^*, \ldots, p_S^*) < \pi_{S-2}(p_1^*, \ldots, p_{S-1}^* - \epsilon, p_{S-1}^*, p_S^*)$, which implies that seller $S-2$ also has incentive to deviate, and so on. It follows that there is no pure strategy Nash equilibrium in the proposed model of shopbots, given buyers of type A, B, and C.

Chapter 6

Summary and Conclusions

In the last twenty years, there has been a merging of computer science, finance, and economics, exemplified by the development of program trading systems, real-time arbitraging, and sophisticated market models. These and related applications form an emerging field known as Computational Economics, in which numerical methods from scientific computing are applied to problems in financial economics. The reverse migration of ideas is also possible, namely the application of finance and economics to computer science, but the potential in this direction has not yet been fully exploited. Indeed, breakthroughs in computer science often arise from insights in related fields. A few noteworthy examples come to mind: Turing machines (mathematics), DNA-based computers (biology), Chomsky's hierarchy (linguistics), and neural networks (neuroscience). This thesis looked to ideas from economics – specifically, the theory of games – as inspirations for new computational paradigms. John Quarterman [86], an Internet statistician, describes several features of the Internet, which sum up the motivation for learning to play network games:

> The Internet is distributed by nature. This is its strongest feature, since no single entity is in control, and its pieces run themselves, cooperating to form the network of networks that is the Internet. However, because no single entity is in control, nobody knows everything about the Internet. Measuring it is especially hard because some parts choose to limit access to themselves to various degrees. So ... we have various forms of estimation.

Chapter 1

Chapter 1 presented an overview of one-shot games, including a number of game-theoretic equilibrium concepts. In order to determine the domain of applicability of the various solutions, this thesis considered the dynamics of learning during repeated games. Foster and Young [38] demonstrated that under traditional game-theoretic assumptions, including rationality, play either converges to Nash equilibrium in finite time, or play does not converge at all; in other words, there is no learning. In a similar vein, Nachbar [81] showed that repeated play of strategic form games among rational players does not generally converge to Nash equilibrium, unless players' initial beliefs coincide with a Nash equilibrium. In light of these negative results, Chapter 2 discussed a suite of optimality criteria and corresponding learning algorithms for which repeated play converges to various generalizations of Nash equilibrium.

Chapter 2

Chapter 2 described a sampling of so-called low-rationality learning algorithms found in the literature on game theory, machine learning, and stochastic control. Moreover, (responsive or non-responsive, and naive or informed) varieties of these algorithms were introduced in order to render these learning algorithms applicable in network games. In spite of their diverse origins, the algorithms were presented in a unified framework based on a series of no regret optimality criteria. In some cases, it is known that algorithms which satisfy one kind of no regret converge to some particular generalization of Nash equilibrium. It was shown, however, that other criterion do not correspond in any natural way to game-theoretic solution concepts. Nonetheless, it seems likely that some of the *algorithms* do converge to game-theoretic solutions. It remains to further investigate the properties of these learning algorithms in terms of the theory of games. Furthermore, it remains to extend the framework due to Milgrom and Roberts [78] of consistency with adaptive learning to consistency with *responsive* learning, in order to more accurately describe the properties of learning algorithms which are suitable in network domains.

Chapter 3

Chapter 3 was an investigation of the Santa Fe bar problem from both a theoretical and a practical perspective. Theoretically, it was argued that belief-based learning (*e.g.*, Bayesian updating) yields unstable behavior in this game. In particular, two conditions sufficient for convergence to Nash equilibrium, rationality and predictivity, are inherently incompatible. This result complements the earlier research of Foster and Young [38] and Nachbar [81]. On the practical side, it was shown via simulations that low-rationality learning algorithms do indeed give rise to equilibrium behavior. These conflicting outcomes are of particular interest in the design of computational agents for dissemination on the Internet; the negative theoretical results suggest that straightforward implementations of autonomous agents can give rise to outcomes far from the desired equilibrium. It remains to prove mathematically that low-rationality learning necessarily converges to Nash equilibrium in the Santa Fe bar problem.

Chapter 4

Chapter 4 reported on experimental simulations of learning in network contexts. The following questions were investigated: (i) what sort of collective behavior emerges via low-rationality, responsive learning among a set of automated agents who interact repeatedly in network contexts? (ii) do traditional game-theoretic solution concepts such as Nash equilibrium appropriately characterize the asymptotic play of network games? (iii) to what extent does the asymptotic play depend on three factors, namely the amount of information available to agents, the degree of responsiveness of learning, and the level of asynchrony of play? These questions were researched empirically, by simulating a sampling of responsive learning algorithms, and observing the set of strategies that was played in the long-run. The findings reported suggest that the asymptotic play of network games is rather different from that of standard game-theoretic contexts. In particular, Nash equilibrium is not generally the outcome of responsive learning in asynchronous settings of limited information. It remains to determine an appropriate solution concept with which to capture the outcome of learning in network contexts.

Chapter 5

Chapter 5 was a study of shopbot economics. This thesis proposed and analyzed an economic model of shopbots, and simulated an electronic marketplace inhabited by shopbots and pricebots, the latter being automated, price-setting agents that seek to maximize profits for sellers, just as shopbots seek to minimize costs for buyers. Analysis revealed that like the Santa Fe bar problem, rational learning in shopbot economics leads to instabilities that manifest themselves as price wars. In contrast, low-rationality learning yields behaviors ranging from tacit collusion to exponential cycling, with only sophisticated learning algorithms converging to mixed strategy Nash equilibrium. One possible extension to this research would be the study of the dynamics a full-fledged electronic marketplace, including adaptive shopbots as well as pricebots, the former acting as economic agents who dynamically charge users for the information services they provide.

Conclusions

This thesis advocated game theory and economics as frameworks in which to model the interactions of computational agents in network domains. On the one hand, the collective dynamics that arise in populations of learning agents were studied, where computational interactions were described via the theory of games. Understanding the collective behavior of agents is essential to the design of networking mechanisms that satisfy globally desirable properties. In the course of these investigations, the behavior of agents was mathematically prescribed by algorithms, as opposed to being described by hypotheses, as is often the approach in experimental economics. Having done so, this thesis came full circle, applying algorithmic ideology to game theory in the domain of shopbot economics. That which was learned by tailoring economics to computational interactions is readily applicable in the rapidly expanding world of e-commerce. Suddenly, the range of applications of computational game theory seems infinite.

Bibliography

[1] Internet Industry Almanac, 1998. `http://www.c-i-a.com/`.

[2] W.B. Arthur. Inductive reasoning and bounded rationality. *Complexity in Economic Theory*, 84(2):406–411, 1994.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. ACM Press, November 1995.

[4] R. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

[5] R. Aumann. Game theory. In John Eatwell, Murray Milgate, and Peter Newman, editors, *Game Theory*. W.W. Norton & Company, Inc., New York, 1987.

[6] R. Aumann and S. Hart, editors. *Handbook of Game Theory with Economic Applications*, volume 1. North-Holland, Amsterdam, 1992.

[7] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[8] R. Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, New Jersey, 1997.

[9] J.Y. Bakos. Reducing buyer search costs: Implications for electronic marketplaces. *Management Science*, 43(12), 1997.

[10] A. Banos. On pseudo games. *The Annals of Mathematical Statistics*, 39:1932–1945, 1968.

[11] B.D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028, 1984.

[12] J. Bertrand. Théorie mathématique de la richesse sociale. *Journal des Savants*, pages 499–508, 1883.

[13] K. Binmore. *Fun and Games: A Text on Game Theory.* D.C. Heath and Company, Massachusetts, 1992.

[14] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[15] T. Borgers and R. Sarin. Learning through reinforcement and replicator dynamics. Mimeo, 1995.

[16] A. Borodin and R. El-Yaniv. *Online Computation and Competitive Analysis.* Cambridge University Press, Cambridge, England, 1998.

[17] C. Boutilier, Y. Shoham, and M.P. Wellman. Economic principles of multi-agent systems. *Artificial Intelligence*, 94:1–6, 1997.

[18] L. Brouwer. Uber abbildungen von mannigfaltigkeiten. *Mathematische Annalen*, 38(71):97–115, 1912.

[19] K. Burdett and K.L. Judd. Equilibrium price dispersion. *Econometrica*, 51(4):955–969, 1983.

[20] A. Chavez and P. Maes. Kasbah: an agent marketplace for buying and selling goods. In *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, London, April 1996.

[21] Y. Chen. Asynchronicity and learning in cost-sharing mechanisms. Mimeo, 1997.

[22] J.Q. Cheng and M.P. Wellman. The WALRAS algorithm: A convergent distributed implementation of general equilibrium outcomes. *Computational Economics*, 12:1–24, 1998.

[23] H. Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[24] A. Cournot. *Recherches sur les Principes Mathematics de la Theorie de la Richesse*. Hachette, 1838.

[25] T. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1 – 29, 1991.

[26] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[27] P. Diamond. A model of price adjustment. *Economic Theory*, 3:156–168, 1971.

[28] *Digital Data Network Directory*. Defense Communications Agency, June 1984.

[29] D. Eichmann. Ethical web agents. In *Proceedings of the Second World Wide Web Conference '94: Mosaic and the Web*, 1994.

[30] I. Erev and A. Roth. On the need for low rationality cognitive game theory: Reinforcement learning in experimental games with unique mixed strategy equilibria. Mimeo, 1996.

[31] I. Erev and A. Roth. On the need for low rationality cognitive game theory: Reinforcement learning in experimental games with unique mixed strategy equilibria. Mimeo, 1996.

[32] J. Eriksson, N. Finne, and S. Janson. Information and interaction in MarketSpace — towards an open agent-based market infrastructure. In *Proceedings of the Second USENIX Workshop on Electronic Commerce*, November 1996.

[33] R. Even. *Distributed Intelligence with Bounded Rationality: Applications to Economies and Networks*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, New York, December 1998.

[34] R. Even and B. Mishra. CAFÉ: A complex adaptive financial environment. In *Proceedings of Conference of Computational Intelligence for Financial Engineering*, pages 20–25, March 1996.

[35] D.F. Ferguson, C. Nikolau, and Y. Yemini. An economy for flow-control in computer networks. In *Proceedings of Infocom '89*, pages 110–118. IEEE, 1989.

[36] D. Foster and R. Vohra. A randomization rule for selecting forecasts. *Operations Research*, 41(4):704–709, 1993.

[37] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 21:40–55, 1997.

[38] D. Foster and P. Young. When rational learning fails. Mimeo, 1998.

[39] Y. Freund and R. Schapire. Game theory, on-line prediction, and boosting. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 325–332. ACM Press, May 1996.

[40] E. Friedman. Dynamics and rationality in ordered externality games. *Games and Economic Behavior*, 16:65–76, 1996.

[41] E. Friedman. Learnability in a class of non-atomic games arising on the Internet. *Mimeo*, 1998.

[42] E. Friedman and S. Shenker. Synchronous and asynchronous learning by responsive learning automata. Mimeo, 1996.

[43] E. Friedman and S. Shenker. Learning and implementation on the Internet. Mimeo, 1997.

[44] D. Fudenberg and D. Levine. Conditional universal consistency. Mimeo, 1995.

[45] D. Fudenberg and D. Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[46] D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, 1998.

[47] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, 1991.

[48] J. Gittens. *Multi-armed Bandit Allocation Indices*. Wiley, New York, 1989.

[49] H. Green. Good-bye to fixed pricing? *Business Week*, pages 71–84, May 1998.

[50] A. Greenwald, E. Friedman, and S. Shenker. Learning in network contexts: Results from experimental simulations. Presented at *Fourth Informs Telecommunications Conference*, March 1998.

[51] A. Greenwald and J.O. Kephart. Shopbots and pricebots. In *Proceedings of Sixteenth International Joint Conference on Artificial Intelligence*, To Appear, August 1999.

[52] A. Greenwald, B. Mishra, and R. Parikh. The Santa Fe bar problem revisited: Theoretical and practical implications. Presented at *Stonybrook Festival on Game Theory: Interactive Dynamics and Learning*, July 1998.

[53] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

[54] J.E. Hanson and J.O. Kephart. Spontaneous specialization in a free-market economy of agents. Presented at *Workshop on Artificial Societies and Computational Markets at Autonomous Agents '98*, May 1998.

[55] G. Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[56] J. Harsanyi. Games with incomplete information played by Bayesian players. *Management Science*, 14:159–182, 320–334, 486–502, 1967–1968.

[57] S. Hart and A. Mas Colell. A simple adaptive procedure leading to correlated equilibrium. Technical report, Center for Rationality and Interactive Decision Theory, 1997.

[58] E. Hopkins and R.M. Seymour. Price dispersion: An evolutionary approach. Mimeo, 1996.

[59] M.T. Hsiao and A. Lazar. A game theoretic approach to decentralized flow control of markovian queueing networks. In Courtois and Latouche, editors, *Performance'87*, pages 55–73. North-Holland, 1988.

[60] S. Huck, H-T. Normann, and J. Oechssler. Learning in a cournot duopoly – an experiment. Mimeo, 1997.

[61] J. Van Huyck, R. Battalio, and F. Rankin. Selection dynamics and adaptive behavior without much information. Mimeo, 1996.

[62] V. Jacobson. Congestion avoidance and control. In *Proceedings of SIGCOMM*, pages 314–329. ACM Press, August 1988.

[63] L. Kaelbling, M. Littman, and A. Moore. Reinforcement learning: A survey. *Artificial Intelligence Research*, 4:237–285, 1996.

[64] S. Kakutani. A generalization of brouwer's fixed point theorem. *Duke Mathematical Journal*, 8:457–458, 1941.

[65] E. Kalai and E. Lehrer. Rational learning leads to Nash equilibrium. *Econometrica*, 61:1019–1045, 1993.

[66] E. Kalai and E. Lehrer. Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23:73–86, 1994.

[67] J.O. Kephart and A.R. Greenwald. Shopbot economics. In *Proceedings of Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, To Appear, July 1999.

[68] J.O. Kephart, J.E. Hanson, D.W. Levine, B.N. Grosof, J. Sairamesh, R.B. Segal, and S.R. White. Dynamics of an information filtering economy. In *Proceedings of the Second International Workshop on Cooperative Information Agents*, July 1998.

[69] J.O. Kephart, J.E. Hanson, and J. Sairamesh. Price and niche wars in a free-market economy of software agents. *Artificial Life*, 4(1), 1998.

[70] J.O. Kephart, J.E. Hanson, and J. Sairamesh. Price-war dynamics in a free-market economy of software agents. In C. Adami, R. K. Belew, H. Kitano, and C. Taylor, editors, *Proceedings of Sixth International Conference on Artificial Life*, Cambridge, June 1998. MIT Press.

[71] Y.A. Korilis, A. Lazar, and A. Orda. The designer's perspective to noncooperative networks. In *Infocom*, Boston, 1995.

[72] T.G. Lewis. *The Friction-Free Economy: Marketing Strategies for a Wired World*. Hardcover HarperBusiness, 1997.

[73] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann, 1994.

[74] J. Bradford De Long and A. Michael Froomkin. The next economy? In D. Hurley, B. Kahin, and H. Varian, editors, *Internet Publishing and Beyond: The Economics of Digital Information and Intellecutal Property*. MIT Press, Cambridge, 1998.

[75] R. Lukose and B. Huberman. A methodology for managing risk in electronic transactions over the internet. Third International Conference on Computational Economics, June 1997.

[76] J. Mackie-Mason and H. Varian. Pricing the Internet. In *Public Access to the Internet*, pages 269–314. MIT Press, Cambridge, 1995.

[77] N. Megiddo. On repeated games with incomplete information played by non-Bayesian players. *International Journal of Game Theory*, 9:157–167, 1980.

[78] P. Milgrom and J. Roberts. Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior*, 3:82–100, 1991.

[79] B. Mishra. A tale of two bars: A negative result for the Santa Fe bar problem. Mimeo, 1998.

[80] D. Mookherjee and B. Sopher. Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19:97–132, 1996.

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

[82] K. Narendra and M.A.L. Thathachar. Learning automata: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-4(4):323–334, 1974.

[83] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

[84] R. Parikh. A connection between Nash equilibria and propositional logic. Mimeo, 1998.

[85] D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

[86] J. Quarterman, 1997. `http://www.bltg.com/demograf.html`. BusinessLINK Technology Group Inc.

[87] A. Rapoport. *Two-Person Game Theory.* University of Michigan Press, Ann Arbor, 1966.

[88] Global Reach, April 1999. `http://www.euromktg.com/globstats/`.

[89] A. Roth and I. Erev. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.

[90] A. Rubinstein. The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review*, 79:385–391, 1989.

[91] B. Russell. Letter from Russell to Frege. In *From Frege to Gödel*, pages 124–125, Cambridge, 1867. Harvard University Press.

[92] J. Sairamesh and J.O. Kephart. Price dynamics of vertically differentiated information markets. In *Proceedings of First International Conference on Information and Computation Economies*, October 1998.

[93] S. Salop and J. Stiglitz. Bargains and rip-offs: a model of monopolistically competitive price dispersion. *Review of Economic Studies*, 44:493–510, 1977.

[94] S. Salop and J. Stiglitz. A theory of sales: A simple model of equilibrium price dispersion with identical agents. *American Economic Review*, 72(5):1121–1130, 1982.

[95] T. Sandholm and R. Crites. Multiagent reinforcement learning in the iterated prisoners' dilemma. *Biosystems*, 37:147–166, 1995. Special Issue on the Prisoners' Dilemma.

[96] T. Schelling. *Micromotives and Macrobehavior*. W.W. Norton and Company, New York, 1978.

[97] F. Schneider. What good are models and what models are good? In Sape Mullender, editor, *Distributed Systems*. ACM Press, New York, 1993.

[98] L.S. Shapley. A value for n-person games. In H. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*, volume II, pages 307–317. Princeton University Press, 1953.

[99] S. Shenker. Efficient network allocations with selfish users. In P.J.B. King, I. Mitrani, and R.J. Pooley, editors, *Performance '90*, pages 279–285. North Holland, New York, 1990.

[100] S. Shenker. Making greed work in networks: A game-theoretic analysis of switch service disciplines. *IEEE/ACM Transactions on Networking*, 3:819–831, 1995.

[101] S. Shenker, September 1998. Personal Communication.

[102] Y. Shoham and M. Tennenholtz. Co-learning and the evolution of social activity. Mimeo, 1993.

[103] J.M. Smith. *Evolution and the Theory of Games*. Cambridge University Press, New York, 1982.

[104] G.J. Stigler. The economics of information. *Journal of Political Economy*, 69(3):213–225, 1961.

[105] G. Sussman. Personal communication, April 1999.

[106] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Massachusetts, 1998.

[107] A. Tanenbaum. *Computer Networks*. Prentice Hall, Amsterdam, 3rd edition, 1996.

[108] J. Tirole. *The Theory of Industrial Organization*. The MIT Press, Cambridge, MA, 1988.

[109] M. Tsvetovatyy, M. Gini, B. Mobasher, and Z. Wieckowski. MAGMA: an agent-based virtual market for electronic commerce. *Applied Artificial Intelligence*, 1997.

[110] H. Varian. A model of sales. *American Economic Review, Papers and Proceedings*, 70(4):651–659, 1980.

[111] J. von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

[112] C.A. Waldsburger, T. Hogg, B.A. Huberman, J.O. Kephart, and W.S. Stornetta. SPAWN: A distributed computational economy. *IEEE Transactions on Software Engineering*, 18:103–117, 1992.

[113] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, 1989.

[114] M. Wellman and J. Hu. Conjectural equilibrium in multi-agent learning. *Machine Learning*, 33:179–200, 1998. Special Issue on Multi-agent Learning.

[115] L.L. Wilde and A. Schwartz. Equilibrium comparison shopping. *Review of Economic Studies*, 46:543–553, 1979.

[116] L.L. Wilde and A. Schwartz. Comparison shopping as a simultaneous move game. *Economic Journal*, 102:562–569, 1992.

[117] V. Wu, R. Manmatha, and E. Riseman. Finding text in images. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 1–10, July 1998.

[118] Z. Zhensheng and C. Douligeris. Convergence of synchronous and asynchronous greedy algorithms in a multiclass telecommunications environment. *IEEE Transactions on Communications*, 40(8):1277–1281, 1992.