University of California
Los Angeles

Distributed Learning and Efficient Outcomes in Uncertain and Dynamic Environments

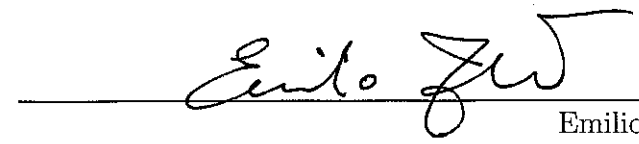
A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mechanical Engineering

by

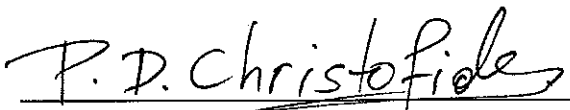
Georgios Christos Chasparis

2008

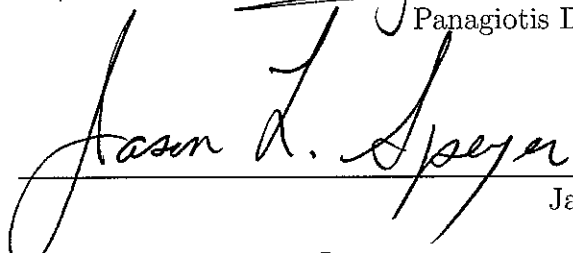
The dissertation of Georgios Christos Chasparis is approved.



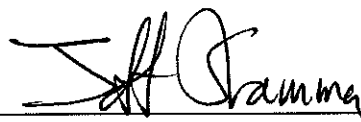
Emilio Frazzoli



Panagiotis D. Christofides



Jason L. Speyer



Jeff S. Shamma, Committee Chair

University of California, Los Angeles

2008

Table of Contents

List of Figures

List of Tables

List of Symbols

α	combination of actions of all agents or action profile
α_i	action of agent i
$\mathcal{A}$	cartesian product of the action spaces of all agents
$\mathcal{A}_i$	the set of actions of agent i
$\lvert\mathcal{A}_i\rvert$	cardinality of the set of actions of agent i
$\mathcal{B}$	set of inputs of a learning automaton
$\mathcal{B}_\varepsilon(\mathcal{V})$	the ε-neighborhood of set $\mathcal{V}$
$\mathrm{dist}(x, \mathcal{V})$	the distance from the point x to the set $\mathcal{V}$
D_i	payoff matrix of agent i
Δx_i	conditional expectation of the change in state x_i of agent i
ΔR_i	conditional expectation of the change in payoff R_i of agent i
$\Delta(\lvert\mathcal{A}_i\rvert)$	set of probability distributions over the set of actions $\mathcal{A}_i$
$\epsilon(k)$	step size sequence of a learning algorithm
e_j	unit vector whose jth entry is equal to 1 and the rest are 0
$E[X]$	expected value of the random variable X
$E[X\|Y]$	expected value of the random variable X given Y
$\mathcal{F}$	the update rule of a learning automaton
$\mathcal{G}$	the decision rule of a learning automaton
γ_i	gain of dynamic reinforcement for agent i
Γ	a (deterministic) game
$-i$	complementary set of agents $\mathcal{I}\backslash i$
$\mathcal{I}$	set of agents
λ	perturbation parameter or mutation rate
$P[A]$	probability of an event $A \in \mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, P)$
$\mathrm{P}(\cdot\|x)$	transition function of a Markov process
$\overline{r}_i(x)$	expected reward vector of agent i given state x

R	combination of payoffs of all agents or payoff profile
R_i	utility (or payoff) of agent i
$\overline{R}_i(x)$	expected reward of agent i given state x
$\mathbb{R}$	set of real numbers
$\mathcal{S}$	set of stationary points
$\overline{v}_i(\alpha_i, x)$	conditional reward of agent i when selects action α_i given state x
x_{ij}	probability that agent i selects action j
x_i	strategy of agent i
x_{-i}	strategy profile of all agents but i
$\mathcal{X}$	set of internal states or strategies of an automaton
$\mathbf{1}$	vector of ones of appropriate size

Acknowledgments

I would like to express my sincere gratitude to my advisor, Prof. Jeff S. Shamma, for his guidance and encouragement during my graduate studies. It was a unique experience to discuss research problems with him and I benefited greatly from his insight on controls that led to this work. I will always be grateful to him.

I would like also to thank my professors in the Henry-Samueli School of Engineering and Applied Sciences at UCLA, and especially Prof. Panagioti Christofides, Prof. Emilio Frazzoli and Prof. Jason Speyer for their support and advice during my graduate studies and their willingness to serve on my thesis committee.

During my graduate studies at UCLA, I was able to meet brilliant students, some of whom I was really lucky to have as best friends. I would really like to thank Vincent Seah, Ibrahim Al-Shyoukh, Jason Marden, Nestor Perez, and Ioannis Souldatos who were always there for me when I asked for their advice or help and who made my life and work at UCLA much more enjoyable.

Finally, it is fair for me to say that this work would have never been possible without the support and encouragement from my parents, Christos Chasparis and Evgenia Chaspari. I am grateful to them for everything I am and everything I have achieved so far.

Thesis supported by AFOSR/MURI grant #FA9550-05-1-0239 and NSF grant #ECS-0501394.

Vita

1978	Born, Athens, Greece.
2001	Diploma (Mechanical Engineering), National Technical University of Athens, Greece.
2002–2004	Research Assistant, Mechanical and Aerospace Engineering Department, UCLA.
2004	M.S. (Mechanical Engineering), UCLA, Los Angeles, California.
2004–2008	Research and Teaching Assistant, Mechanical and Aerospace Engineering Department, UCLA.

Publications and Presentations

Chasparis, G. and Shamma, J., *LP-based multi-vehicle path planning with adversaries*, pages 261-279. In Shamma, Jeff (ed.), Cooperative Control of Distributed Multi-Agent Systems, John Wiley & Sons, February 2008.

Chasparis, G. and Shamma, J., *Distributed Dynamic Reinforcement of Efficient Outcomes in Multiagent Coordination*, European Control Conference (ECC), Greece, Kos, July 2-5, 2007.

Chasparis, G. and Shamma, J., *The Emergence of Efficient Social Networks by Dynamic Reinforcement*, Paper presented at the 4th Lake Arrowhead Conference on Human Complex Systems, April 25-29, 2007.

Chasparis, G. and Shamma, J., *Linear-Programming-Based Multi-Vehicle Path Planning with Adversaries*, 24th American Control Conference (ACC), Portland, Oregon, Jun. 8-10, 2005.

Papadopoulos, E. and Chasparis, G., *Analysis and Model-based Control of Servomechanisms with Friction*, ASME J. Dynamic Systems, Measurement and Control, Vol. 126, No. 4, December 2004.

Papadopoulos, E. and Chasparis, G., *Analysis and Model-based Control of Servomechanisms with Friction*, Proc. of the 2002 IEEE/RSG Int. Conference on Intelligent Robots and Systems (IROS '02), Lausanne, Switzerland, October 2002.

Abstract of the Dissertation

Distributed Learning and Efficient Outcomes in Uncertain and Dynamic Environments

by

Georgios Christos Chasparis

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2008

Professor Jeff S. Shamma, Chair

This dissertation focuses on the problem of multiagent coordination. When agents have access to limited information about the environment (possibly other agents) and learn what to play through repeated experimentation, convergence to desirable equilibria might be challenging. The main contribution of the dissertation is the introduction of a learning adaptation method, similar to reinforcement learning techniques, accompanied with decision rules that are based on feedback control (dynamic reinforcement). This learning framework exploits transient phenomena of the dynamics (off-equilibrium behavior) to reinforce convergence to efficient outcomes when the induced stochastic process has multiple resting points. In particular, it is shown analytically that non-efficient outcomes can be destabilized when dynamic reinforcement is applied by even a single agent. The utility of the proposed framework is illustrated in coordination games and distributed network formation, where non-efficient resting points of the stochastic process can be destabilized. In the case of distributed network formation, which is of independent interest, we also illustrate the utility of the proposed learning adaptation method to incorporate multiple design criteria, usually met in topology control for ad-hoc networks, which can reinforce convergence to desirable outcomes.

CHAPTER 1

Introduction

1.1 Motivation

Recent research in autonomous control systems, such as robotic systems, has shown the importance of designing systems that are *adaptive* to *uncertain environments* while, in parallel, accomplish *desirable* high-level objectives, such as path-planning, formation or self-assembly. By uncertain environments, we usually mean other robotic systems with similar or conflicting interests. The problem becomes even harder when a robotic system consists of a large number of robots, due to the possibility of failures or the complexity of the problem (e.g., distributed estimation, multiplicity of objectives, etc.). Each robot may have access to only *local* information, while there might exist *global* objectives.

Problems that fit into the above framework include self-organization of robotic systems (where multiple robots need to form a desired structure) and motion formation of vehicles (where multiple vehicles need to move in a desirable formation). Applications also include topology formation of an ad-hoc sensor wireless network and coverage of an unknown area by autonomous sensors.

A common element of tasks of this form is the presence of multiple robots (usually called *agents* or *players*). The agents independently decide what to do, even without having the necessary available information, or even without knowing the team's objective. Given the inherent uncertainty in these problems and the absence of centralized control, these problems have been addressed within the framework of

learning, where agents *learn* their *behavior* by interacting with their environment, either this environment corresponds to other agents, or an adversarial entity.

Systems with multiple agents (or *multiagent systems*) are not only encountered in engineering problems. Multiagent systems have been used to model sociological phenomena or simulate collective behavior in societies. In sociological systems, agents interact with each other having only minimum available information about the behavior of the group, and applying generally naive rules of behavior. The behavior of the group seems to follow certain rules, in several cases more *rational* than the agents' naive behavior. Social scientists try to describe such phenomena and explain the underlying local interactions rules that could be responsible for certain collective behaviors. Most of these methods rely on learning approaches, where agents adopt behaviors that improve their performance with time.

Borrowing techniques and models used in different disciplines to describe collective behavior of multiagent systems can facilitate modeling and analyzing complex systems. From an engineer's perspective it may lead to new design techniques for systems that are adaptive and learn to behave in a certain (desirable) way. Such problems also require new *local* control techniques for reinforcing convergence to certain desirable *global* behaviors. This dissertation is a small effort towards these lines.

In conclusion, some of the most important challenges in *multiagent* systems are:

- modeling learning behavior at the agent-level based on local information;
- characterizing collective behavior based on the models of local interactions;
- introducing control techniques at the *agent*-level that will reinforce certain desirable outcomes at the *group*-level.

1.1.1 Coordination problems

We are going to restrict our attention to a class of problems for multiagent systems where each agent's objective is to *coordinate* (in some sense) with other agents. Such situations can be classified as *coordination problems*. Since the multiagent systems considered here do not assume any centralized control, we further assume that such a coordination will be the *result* of an interaction process among agents.

In order to define a *coordination problem*, we adopt the definition of [Lew02]. We first need to define a *strategic interaction* (or *game*) among multiple agents. In a strategic interaction, each agent must choose one of several alternative actions (finite in number). Often all agents have the same set of alternative actions, however this is not necessary. Also each agent has *preferences* over the joint actions of all agents, which implies that the outcome of any action an agent might choose depends on the actions of the other agents.

Some combinations of actions are *equilibria*, at which each agent has done as well as it can given the actions of the other agents.[1] In an equilibrium combination, there is no agent that it would have been better off had it alone acted otherwise. It is possible that some or all of the agents would have been better off if all or some had acted differently. Also, it is possible that an agent would have been better off had one or some of the other agents have acted otherwise.

We can illustrate equilibria through *payoff* (or *utility*) matrices for coordination problems between two agents, as shown in Table 1.1. Agent 1 selects *row* and agent 2 selects *column*. We also label alternative actions by labeled rows and columns (using capital letters, e.g., A and B). The squares then represent combinations of the agents' actions. We label these squares by two payoffs, where the left one corresponds to agent 1, and the right one corresponds to agent 2. These payoffs measure the desirability of the outcome for each agent, i.e., a high-payoff outcome is more desirable than a

[1] In other words, a *Nash equilibrium*.

low-payoff outcome. For example, in the game of Table 1.1, outcome (A, A) is more desirable than (B, B) for both agents.

	$2.A$	$2.B$
$1.A$	$2, 2$	$0, 0$
$1.B$	$0, 0$	$1, 1$

Table 1.1: A strategic interaction of two players and two actions.

Sometimes coordination problems are defined as situations where each agent is trying to achieve uniformity of actions by each doing whatever the others will do. What is important about uniform action combinations is that they are equilibria rather than they correspond to uniform action combinations. However, just the presence of equilibria does not make a strategic interaction a coordination problem.[2]

The main characteristic of a coordination problem that distinguishes it from other strategic interactions lies in the presence of *coordination equilibria*.

Definition 1.1.1 (Coordination equilibrium) *A coordination equilibrium is a combination of actions in which no one would have been better off had any one other agent acted otherwise.*

As clearly follows, coordination equilibria are equilibria by definition. For example, in the strategic interaction of Table 1.1, the combinations (A, A) and (B, B) are also coordination equilibria.

Thus, at coordination equilibria it is of every agent's interest to keep playing the equilibrium, i.e., there is a coincidence (or alignment) of interests among the agents. We specialize this property by assuming that coincidence of interest is not only restricted in coordination equilibria. Instead, we assume that *coincidence of interest predominates* in all possible outcomes of the game. More specifically, at any combination of actions, if there is one or some agents that can benefit by changing

[2]For example, a strategic interaction may correspond to a *pure conflict* in which the agent's payoffs sum to zero in every square (usually called a *zero-sum game*).

their actions, then all other agents do not get worse off by this change. We will refer to such a strategic interaction as a game of *aligned interests*.

Definition 1.1.2 (Game with aligned interests) *A game with aligned interests is a strategic interaction of two or more agents in which, for any combination of actions, if there exists an agent that can be better off by changing its action, then there is no agent that would be worse off by such a change.*

Another important characteristic of coordination problems is the multiplicity of coordination equilibria. Of course, there are situations where there is only one coordination equilibrium. However, in these situations the task of reaching a unique coordination equilibrium may be trivial when every agent makes the best choice given the actions of the other agents and when there is no considerable conflict of interest among agents.

Besides the trivial case of a unique coordination equilibrium, there are also situations in which there exist multiple coordination equilibria amongst which an agent is indifferent. We exclude this case, by defining the notion of a *proper* or *strict equilibrium*.[3] In particular, a *proper* or *strict equilibrium* is a combination of actions in which an agent's payoff is strictly greater than its payoff in any other choice it could have made, given the others' choices.

Summing up, we define *coordination problems* as follows:

Definition 1.1.3 (Coordination Problem) *Coordination problems are strategic interactions with aligned interests played by two or more agents in which there are two or more proper coordination equilibria.*

A special category of such games are games of *pure coordination*, where there is a perfect coincidence of interest (i.e., identical payoffs for every combination of actions).

[3]Also called *strict Nash equilibrium*.

For example, the game in Table 1.1 is a game of pure coordination. An example of a coordination problem, that is not a pure coordination, is the Stag-Hunt game shown in Table 1.2.

	$2.A$	$2.B$
$1.A$	$4, 4$	$1, 3$
$1.B$	$3, 1$	$3, 3$

Table 1.2: The Stag-Hunt game.

A common characteristic of the coordination problems of Tables 1.1–1.2 is that proper coordination equilibria coincide with uniform combinations of actions. We will refer to this special class of coordination problems as *coordination games*.

Definition 1.1.4 (Coordination game) *A coordination game is a coordination problem such that each agent has the same number of actions with every other agent, and every uniform combination of actions is a proper coordination equilibrium and vice versa.*

It can be easily verified, not all coordination problems can be modeled as a coordination game.

1.1.2 Examples of coordination problems

Coordination games are a special class of coordination problems. Several strategic interactions, mostly related to social phenomena, can be modeled in the form of a coordination game, including the adoption of new technologies and the adjustment of prices for products between oligopolists. In some social science literature, coordination games, and more specifically the Stag-Hunt game, are considered the simplest games to model the establishment of a social contract [Sky04]. In fact, establishing coordination equilibria with high payoffs can require the cooperation among more

than one agent, as easily seen in Table 1.2.[4]

However, coordination problems are not necessarily restricted to social sciences. In fact, many applications in engineering deal with convergence to coordination [JLM03, BHO05, Mor05, OFM07], synchronization [Str03], swarming or formation control [Olf06]. Several self-organization processes can also fit into the framework of coordination problems (or variations of them), such as the formation of a network among agents, where each agent decides which links to establish.[5]

Other examples that fit into this framework include self-organization of mechanical parts, where certain structures are more beneficial than others, and motion planning for multi-robot systems, where certain formation patterns may be more desirable than others. Note that in all these examples there exist multiple coordination equilibria. However, it is not necessarily true that all these problems can be formulated as aligned interest games. It strongly depends on the specifics of the design method and the underlying assumptions.

1.1.3 The role of strategic learning

Even though several problems related to multiagent systems can be formulated as coordination problems, the main question is *how* agents can solve such a coordination problem so that desirable outcomes are the solutions. This problem has been extensively studied by both economists and sociologists, as well as by computer scientists. The approach followed strongly depends on the underlying assumptions governing the information available to each agent and the communication constraints among agents.

In general, a learning algorithm is imposed that governs the off-equilibrium behavior (i.e., before any form of equilibrium arises). In this framework, agents learn

[4] In the coordination game of Table 1.2, either agent can guarantee payoff 3 by playing action B, but both agents can receive 4 only if they cooperate with each other.

[5] This problem will be analyzed in Chapter 5.

how to interact with the other agents either by trial-and-error (when, for example, the information given to each agent about the other agents' actions is limited), or by more demanding computations (when for example agents have access to the history of action combinations observed).

Examples of learning dynamics include *fictitious play* [FL98, SA05] (where each agent chooses an action that maximizes its expected utility against the empirical frequencies of the actions played by the other agents throughout history), *regret-based* algorithms [You04] (where each agent chooses an action than minimizes its regret), *reinforcement learning* [SP00] (where each agent's confidence playing an action depends on its success throughout history), and *satisficing* [CM05] (where each agent continues playing an action when it provides higher payoff than its aspiration level).

Note that different learning models assume different information structures for each agent. For example, fictitious play assumes full knowledge of the combination of actions played at each iteration by all agents, while reinforcement learning assumes that agents are only aware of their own action and payoff history. Therefore, the learning model used strongly depends on what is considered a *reasonable* model for interactions. For example, in sociological systems a model of fictitious play may be considered reasonable, however, in robotic systems reinforcement learning may be more appropriate.

1.2 Objective and contributions

This thesis is a small contribution towards *modeling*, *analysis* and *distributed control* in multiagent coordination problems under the framework of distributed learning. Starting from the fact that multiagent coordination problems accept many equilibria, some of which are not necessarily desirable, we focus on the problem of equilibrium selection and how desirable equilibria can be sustained through local decision rules.

To this end,

- we introduce a simple *reinforcement learning* to model interactions, which is of independent interest and allows for equilibrium selection in coordination problems;
- we introduce a *dynamic reinforcement* rule, inspired by *feedback control*, for local decisions that is based only on transient phenomena, but reinforces the emergence of efficient equilibria even when only a single agent applies it;
- we illustrate the applicability of this framework in coordination games and show how dynamic reinforcement can exclude convergence to risk-dominant equilibria;
- we further analyze the problem of distributed network formation under the same framework, which is of independent interest;
- we introduce reward functions for distributed network formation that allow for multiple design criteria met in engineering applications such as sensor networks;
- we illustrate the applicability of dynamic reinforcement in equilibrium selection in distributed network formation.

1.3 Thesis outline

In Chapter 2, we introduce the basic framework that we will consider in the remainder of the dissertation. In particular, we introduce the elements of a strategic interaction (or game) among multiple agents in a coordination problem. We characterize the equilibria of coordination problems and we classify them based on payoff and risk. We further discuss several forms of learning dynamics that can be used for both describing social or economic interactions and designing engineering systems. When

agents apply learning dynamics to learn how to interact with other agents, multiple outcomes might emerge. Therefore, we present prior work on equilibrium selection and discuss the advantages and disadvantages of the existing approaches.

In Chapter 3, we introduce a distributed learning model that belongs to the general class of *learning automata*. The specific approach is well known in psychology, computer science and adaptive control, since it is characterized by its simplicity and minimal information requirements. It allows the possibility to design engineering systems, and furthermore can explain social and economic phenomena. The learning model has several similarities with a larger class of learning algorithms, usually called *reinforcement learning*, which have been extensively discussed in machine and robotic systems literature. We analyze the asymptotic properties of this reinforcement scheme that are useful for equilibrium selection in coordination problems.

In Chapter 4, we specialize the stability properties of the reinforcement scheme introduced in Chapter 3 for coordination games. Our main goal is to ensure convergence to a *desired* coordination in a *distributed* and *adaptive* fashion. The exploitation of transient (off-equilibrium) phenomena opens up the possibility of reinforcing a more desirable equilibrium. We are particularly interested in reinforcing the efficient equilibrium under *dynamic reinforcement*, a special form of selecting actions that is inspired by *feedback control* techniques. Unlike traditional reinforcement learning, agents using dynamic reinforcement use a combination of long term rewards and recent rewards to construct myopically forward looking action selection probabilities.

This form of *feedback* in agent's decisions can also be viewed as a more complex life-like behavior. We will show that dynamic reinforcement can be used as an equilibrium selection scheme, since only a *single* agent is able to destabilize the non-desirable equilibria. In fact, dynamic reinforcement, when applied by one or some agents, can make the probability of converging to non-desirable equilibria equal to zero.

We will illustrate the results in coordination games. In these games, we will show

that the dynamic processing presented here can destabilize a non-desirable equilibrium even if the *risk* associated with it is less than the risk of any other equilibrium (i.e., when it *risk-dominates*). In general, the resulting equilibrium selection under the presented dynamic reinforcement need not be determined by either payoff- or risk-dominance or both.

In Chapter 5, the problem of distributed network formation will also be treated within the same framework, where nodes strategically interact by establishing links with other nodes. Agents can form and sever unidirectional links and derive direct and indirect benefits from these links. Also, each agent's decisions depend on its own previous links and past benefits, and link selections are subject to random perturbations. We proceed by characterizing the stability properties of the proposed model. We illustrate the flexibility of the model to incorporate various design criteria, including dynamic cost functions that reflect link establishment and maintenance, and distance-dependent benefit functions. We show that the learning process assigns positive probability to the emergence of multiple stable configurations (called strict Nash networks), which need not emerge under alternative processes such as best-reply dynamics. We analyze the specific case of so-called frictionless benefit flow, and show that a single agent can reinforce the emergence of an efficient network through the aforementioned dynamic reinforcement. Finally, we illustrate how such a distributed reinforcement scheme can be used as a design method for topology control in sensor networks.

Finally, Chapter 6 presents concluding remarks and possible future directions of interest.

CHAPTER 2

Setup and Prior Work

2.1 Introduction

In this chapter, we present the basic framework that we will consider in the remainder of the dissertation. In particular, we introduce the elements of a strategic interaction (or game) among multiple agents in a coordination problem. We characterize the equilibria of coordination problems and we classify them based on payoff and risk. We further discuss several forms of learning dynamics that can be used for both describing social or economic interactions and designing engineering systems. When agents apply learning dynamics to learn how to interact with other agents, multiple outcomes might emerge. Therefore, we present prior work on equilibrium selection and discuss the advantages and disadvantages of the existing approaches.

2.2 Setup

2.2.1 Game

A game involves a finite number of agents, say n. Let $\mathcal{I} \triangleq \{1, 2, ..., n\}$ be the set of agents. Each agent $i \in \mathcal{I}$ has a *finite* set of available *choices* (or *actions*) that will be denoted by $\mathcal{A}_i$. Let $\alpha_i \in \mathcal{A}_i$ denote an action of agent i, and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ the combination of actions of all agents. We will also define $\mathcal{A}$ to be the cartesian product of the action spaces of all agents, i.e., $\mathcal{A} \triangleq \times_{i \in \mathcal{I}} \mathcal{A}_i$.

The combination of actions of all agents, α, produces a *payoff* (or *utility*) for each

agent. The utility of agent i, which will be denoted by R_i, maps the n-tuple of actions (or *action profile*) α to a payoff $R_i(\alpha) \in \mathbb{R}$. It constitutes a measure of the desirability of the action profile α, where a high-payoff action profile is more preferable than a low-payoff action profile. Let also denote by $R : \mathcal{A} \to \mathbb{R}^n$ the combination of payoffs (or *payoff profile*) of all agents, i.e., $R(\cdot) \triangleq (R_1(\cdot), R_2(\cdot), ..., R_n(\cdot))$.

Definition 2.2.1 (Game) *A (strategic-form) game* Γ *is a triple* $\{\mathcal{I}, \mathcal{A}, R\}$.

Since each agent selects actions independently, we generally assume that each agent's action is a realization of an independent discrete random variable. Let $x_{ij} \in [0, 1]$ denote the probability that agent i selects action $\alpha_i = j \in \mathcal{A}_i$. If $\sum_{j \in \mathcal{A}_i} x_{ij} = 1$, then $x_i \triangleq (x_{i1}, x_{i2}, ..., x_{i|\mathcal{A}_i|})$ is a probability distribution over the set of actions $\mathcal{A}_i$ (or *strategy* of agent i), where $|\mathcal{A}_i|$ denote the cardinality of the set $\mathcal{A}_i$.

Let $\Delta(|\mathcal{A}_i|)$ denote the set of probability distributions (or *probability simplex*) over the set of actions $\mathcal{A}_i$, i.e.,

$$\Delta(|\mathcal{A}_i|) \triangleq \{x \in \mathbb{R}^{|\mathcal{A}_i|} : x \geq 0, \mathbf{1}^{\mathrm{T}} x = 1\},$$

where $\mathbf{1}$ is the vector of ones of size $|\mathcal{A}_i|$. Then $x_i \in \Delta(|\mathcal{A}_i|)$. We will also use the term *strategy profile* to denote the combination of strategies of all agents $x = (x_1, x_2, ..., x_n) \in \times_{i \in \mathcal{I}} \Delta(|\mathcal{A}_i|)$. For brevity we will use the notation $\mathcal{X} \triangleq \times_{i \in \mathcal{I}} \Delta(|\mathcal{A}_i|)$.

Note that if x_i is a *unit vector* (or a vertex of $\Delta(|\mathcal{A}_i|)$), say e_j, then agent i selects action j with probability one. This strategy will be called *pure strategy*. Accordingly, a *pure strategy profile* is a profile of pure strategies. We will also use the term *mixed strategy* to denote a strategy that is *not* pure.

Let $E[X]$ denote the *expected value* of a random variable X. We define a *Nash equilibrium* as follows.

Definition 2.2.2 (Nash equilibrium) *A strategy profile* $x^* = (x_1^*, x_2^*, ..., x_n^*)$ *is a Nash equilibrium if and only if, for each agent* $i \in \mathcal{I}$*,*

$$E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(x_i^*, x_{-i}^*)] \geq E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(x_i, x_{-i}^*)] \tag{2.1}$$

for all $x_i \in \Delta(|\mathcal{A}_i|)$ *and* $x_i \neq x_i^*$*, where* x_{-i}^* *denote the equilibrium strategy profile of all agents but* i*.*[1]

In the special case where for all $i \in \mathcal{I}$, x_i^* is a pure strategy, then the Nash equilibrium is called *pure Nash equilibrium*. Also, in case the inequality in (2.1) is strict the Nash equilibrium is a *proper* or *strict equilibrium* and it will be called a *strict Nash equilibrium*.

2.2.2 Coordination problems

Based on the definition of a Nash equilibrium, a *coordination equilibrium* is defined as follows.

Definition 2.2.3 (Coordination equilibrium) *A coordination equilibrium is a pure Nash equilibrium* $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_n^*)$ *in which, for any agent* $i \in \mathcal{I}$ *and any agent* $s \in \mathcal{I}$*,*

$$E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha_s^*, \alpha_{-s}^*)] \geq E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha_s, \alpha_{-s}^*)] \tag{2.2}$$

for all $\alpha_s \in \mathcal{A}_s$ *and* $\alpha_s \neq \alpha_s^*$*.*

Similarly to the definition of a strict Nash equilibrium, a *strict coordination equilibrium* is a coordination equilibrium that satisfies (2.2) with strict inequality.

Definition 2.2.4 (Game with aligned interests) *A game with aligned interests is a strategic interaction of two or more agents in which, for any combination of*

[1]The notation $-i$ denotes the complementary set $\mathcal{I}\backslash\{i\}$. We will often split the argument of a function in this way, e.g., $F(\alpha) = F(\alpha_i, \alpha_{-i})$ or $F(x) = F(x_i, x_{-i})$.

actions $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$*, if there exist an agent* $i \in \mathcal{I}$ *and action* $\alpha'_i \in \mathcal{A}_i$ *such that* $\alpha'_i \neq \alpha_i$ *and*

$$E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha'_i, \alpha_{-i})] \geq E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha_i, \alpha_{-i})],$$

then $E[R_s(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha'_i, \alpha_{-i})] \geq E[R_s(\alpha_1, \alpha_2, ..., \alpha_n)|(\alpha_i, \alpha_{-i})]$*, for all* $s \neq i$*.*

As we have already stated in Definition 1.1.2, a coordination problem is defined as follows.

Definition 2.2.5 (Coordination problem) *A coordination problem is a game with aligned interests played by two or more agents in which there are two or more strict coordination equilibria.*

Then, a coordination game is defined as follows.

Definition 2.2.6 (Coordination game) *A coordination game is a coordination problem such that* $|\mathcal{A}_1| = |\mathcal{A}_2| = ... = |\mathcal{A}_n|$ *and the strict coordination equilibria are* $\{x \in \mathcal{X} : x = (e_j, e_j, ..., e_j), j \in \mathcal{A}_i\}$*.*

2.2.3 Payoff versus risk dominance

One measure for comparing coordination equilibria in a coordination problem is the importance that each agent assigns to it, which is reflected in the payoff level. Hence, we say that a coordination equilibrium *payoff-dominates* another coordination equilibrium, if and only if each agent's payoff is greater in the former equilibrium than in the latter one. For example, in the Stag-Hunt game of Table 1.2, the coordination equilibrium (A, A) payoff-dominates the coordination equilibrium (B, B). A coordination equilibrium will be called *payoff-dominant* (or *efficient*) if it payoff-dominates any other coordination equilibrium. For example, the coordination equilibrium (A, A) in Table 1.2 is the payoff-dominant equilibrium.

A different measure for comparing coordination equilibria is related to the *risk* associated with each equilibrium [HS88]. Consider, for example, the generic coordination game of Table 2.1. For such a game to be a coordination game, we need

	$2.A$	$2.B$
$1.A$	a_{11}, b_{11}	a_{12}, b_{12}
$1.B$	a_{21}, b_{21}	a_{22}, b_{22}

Table 2.1: A generic game.

to assume that $a_{11} > a_{21}$, $b_{11} > b_{12}$, $a_{22} > a_{12}$, and $b_{22} > b_{21}$. We will define the *risk factor* of equilibrium (A, A) (or (B, B)) as the smallest probability, say p, such that if one agent believes that the other agent is going to play action A (or B) with probability $> p$, then A (or B) is the unique optimal action to take.

For example, if agent 1 believes that agent 2 is playing action A with probability p, then agent 1 will prefer to play action A if and only if

$$a_{11}p + a_{12}(1-p) \geq a_{21}p + a_{22}(1-p)$$

or, equivalently, if

$$p \geq \frac{a_{22} - a_{12}}{a_{11} - a_{21} + a_{12} - a_{22}} \triangleq \alpha.$$

Similarly, agent 2 will prefer to play action A if agent 1 plays action A with probability

$$p \geq \frac{b_{22} - b_{21}}{b_{11} - b_{12} + b_{21} - b_{22}} \triangleq \beta.$$

Then, the *risk factor* of equilibrium (A, A) will be $r_A \triangleq \min\{\alpha, \beta\}$.

Accordingly, we can show that the risk factor of the coordination equilibrium (B, B) will be $r_B \triangleq \min\{1-\alpha, 1-\beta\}$. Then, we define the *risk-dominant* equilibrium as *the equilibrium with the smallest risk factor*. For example, if $r_A \leq r_B$, then the equilibrium (A, A) is the risk-dominant equilibrium.

In the simpler case of the symmetric game of Table 2.2, assume that $a > c$, $d > b$.

Under these conditions the symmetric game of Table 2.2 is a coordination game. The

	$2.A$	$2.B$
$1.A$	a, a	b, c
$1.B$	c, b	d, d

Table 2.2: A symmetric game.

combination (A, A) is risk-dominant if

$$r_A \leq \frac{1}{2} \Rightarrow a - c \geq d - b.$$

Note that the last condition corresponds to equilibrium (A, A) having the largest *deviation cost*[2].

For example, in the Stag-Hunt game of Table 1.2, if either agent deviates from the equilibrium (B, B), then the deviation cost is 2. Accordingly, when either agent deviates from A, the deviation cost is 1. Therefore, (B, B) is the risk-dominant equilibrium. However, in the coordination game of Table 1.1, the coordination equilibrium (A, A) is the risk-dominant equilibrium.

2.2.4 Repeated games and learning dynamics

If a game Γ is repeated for several periods, then it is called a *repeated game*. In this dissertation, we will only consider games that are played repeatedly. When a game is played repeatedly agents are able to learn through their previous experience *how* to play the game. In social sciences and economics, repeated games were introduced as a way to justify the emergence of a social or economic situations.

Since it was not clear enough which *learning dynamics* is *reasonable* for interactions, multiple learning dynamics have been introduced and analyzed. Some of these dynamics include:

[2]The deviation cost is associated with a deviation from an equilibrium and corresponds to the payoff loss that a agent experiences through this deviation.

– *Replicator dynamics:* A population of agents is considered, where each agent is assigned a strategy. An agent reproduces a number of identical agents (i.e., who play the same strategy) that is proportional to the payoff received at each period (see Chapter 2 in [Sam97]). Consequently, agents with high-payoff strategies are at a reproductive advantage compared to agents with low-payoff strategies.

– *Imitation dynamics:* Agents copy the strategies of others, especially strategies that are popular or appear to yield high payoffs [Ale00]. In contrast to replicator dynamics, the payoffs describe how agents select actions, and not on how fast they multiply. This behavior seems more reasonable for learning dynamics. On the other hand, it is assumed that agents observe the strategies of other agents, an assumption that is not always satisfied.

– *Reinforcement learning:* Agents tend to adopt actions that yielded a high payoff in the past, and to avoid actions that yielded a low payoff. More specifically, the probability of taking an action in the present increases with the payoff that resulted from taking that action in the past [NT89, SB98]. Similarly to the imitative models, agents make decisions based on payoffs, but it is an agent's *own* past payoffs that matter, not the payoffs of others.

– *Best-reply dynamics:* Agents adopt actions that optimize their expected payoff given what they expect others to do. In the simplest such models, agents select best replies to the empirical frequency distribution of their opponents' previous actions (*fictitious play*) [FL98, You98, SA05]. Contrary to reinforcement learning, agents make decisions based on observations of other agents' actions, an assumption that is not always satisfied.

The above dynamics can be considered "reasonable" for interactions depending on the underlying assumptions. In social interactions, both imitation and reinforcement learning are appropriate, since people usually observe other people's choices or learn

through their own experience. Instead in economic interactions a more sophisticated model might be more appropriate, such as best reply. From an engineer's point of view, the goal is to design an application where multiple agents are interacting *locally* with each other to accomplish a desirable *global* outcome. In this framework, a learning model that assumes limited information, such as reinforcement learning or imitation dynamics, would be more reasonable.

2.3 Equilibrium selection in coordination games

Given a model of learning dynamics, one of the questions arising in coordination problems is "*what outcome can be considered reasonable?*" Many different disciplines have tried to answer this question, including philosophers, game theorists, sociologists. The same problem has been treated in a different way by computer scientists and engineers, who try to answer the question "*can we achieve a desirable outcome?*" Of course, to answer such a question, it is necessary to know exactly the framework and the rules of interactions among agents.

We would like to answer the second question, i.e., how is it possible to achieve a desirable outcome in a coordination problem, when agents are not necessarily fully strategic, i.e., their behavior is myopic. A useful way for trying to answer such a question is to consider first the simplest coordination problem, that is a coordination game. In these simple games, we are going to investigate first which equilibrium selection techniques have been used in game-theoretic literature, and then discuss open problems and questions emerging. In particular, we will restrict our attention in the following information frameworks:

1. uniform interactions;

2. local interactions with fixed neighborhood;

3. local interactions with migration;

4. local interactions with evolving neighborhood.

In parallel, we will discuss the effect of communication among agents as well as different forms of dynamics in equilibrium selection.

2.3.1 Uniform interactions

The model of [KMR93] focuses on two-player and two-action symmetric coordination games. The model evolves over time, and at each period, each agent is randomly matched to play the game with each of the remaining agents exactly once (as in a tournament). It is assumed that agents select their strategies based on best-reply, i.e., they choose the action that provides the largest expected payoff given the current distribution of strategies in the population. Reference [KMR93] also assumes that agents experiment every once in a while with exogenously fixed probability. The dynamics define an underlying Markov chain, which has a unique stationary distribution. It is shown that when agents play a coordination game for which a risk-dominant equilibrium exists, then the process spends asymptotically most of its time at the risk-dominant equilibrium when the number of agents is sufficiently large.

Reference [You93] considers a similar model, where agents are drawn from a large, finite population of agents. Each agent who is selected to play the game, chooses a strategy that is a best-reply to a sample of the history of play in the past. It is also assumed that agents make mistakes or experiment with different strategies. When the dynamics are applied to a two-player and two-action coordination game for which a risk-dominant equilibrium exists, [You93] shows that this dynamic process spends asymptotically most of its time at the risk-dominant Nash equilibrium when the sample and history size are sufficiently large.

The results of [KMR93] and [You93] have a natural intuition, namely that *the*

basin of attraction of the risk-dominant equilibrium is larger than that of the non risk-dominant equilibrium. In the long run this leads to a higher probability that in any given period agents will be playing the risk-dominant equilibrium. This is the *stochastically stable* convention, in the sense defined by [FY90].

Reference [Blu03] investigates how robust the above results are to different models of noise effects, called *noise models*. The model considers a population of agents who are randomly paired to play a two-strategy symmetric coordination game. At randomly chosen moments, agents have an opportunity to revise their current strategy choice, according to the expected utility of each strategy when the distribution of strategies is assumed known (e.g., agents may best-respond to the current distribution of play). A stochastic alternative is to consider a noise model similar to [KMR93], or the "log-linear model" of [Blu93] where the log of the odds of choosing a given strategy is proportional to the payoff difference between the two strategies. Noise models map payoff differences into trembles. More specifically, the probability of playing a strategy would depend on the difference of its expected payoff from the expected payoff of the other strategies. It is shown that for a general class of noise models (including the log-linear model of [Blu93] and the mistakes model of [KMR93]), the same convergence results of [KMR93, You93] hold.

In the models of [KMR93] and [You93] the probability of a mistake is uniformly distributed in the population of agents and independent of the current state of the iteration process. A different noise model may reinforce different classes of equilibria (not only risk-dominant). More specifically, [BL96] has shown that it is possible to find small noise effects so that *any* long-run prediction is possible. This is very easy to see, if for example we define the noise rule such that there is no noise at either one of the equilibria. Then this is going to be the unique equilibrium that will survive through time. Although these specifically tailored perturbations are *state* dependent and *uniformly* distributed across population, this result suggests that the

noise process cannot be ignored.

Finally, we would like to point out that it is possible to get different convergence results compared to [KMR93, You93], when one allows agents to select more than one action at a time, as proposed by [GG97]. Although such an assumption is not easily justifiable, it is interesting to point out that under this assumption the learning process (which is based on best-reply with mistakes) converges to the payoff-dominant equilibrium in a two-player two-action coordination game.

2.3.2 Local interactions with fixed neighborhood

The problem of equilibrium selection in the presence of a fixed neighborhood structure has been considered by [Ell93]. More specifically, a framework similar to [KMR93] is adopted. In each period the agents are randomly matched and each pair plays a two-player and two-action coordination game. Agents are playing a best-reply to the current distribution of strategies in the population, while their decisions are subject to mistakes. The model of [Ell93] departs from the model of [KMR93] in that it allows for different matching processes within the population. In particular, a local matching rule is introduced in which agents interact with a small group of close friends, neighbors, or colleagues. It is shown that in the case of a local matching rule, the dynamic process converges to the risk-dominant equilibrium as in the model of [KMR93]. It is further shown that the relative probabilities of the time that the process spends on each equilibrium change for some *mutation rate*.[3] In particular, the probability that the process spends its time on the non-risk-dominant equilibrium for some given mutation rate is smaller than the corresponding one at the uniform matching model of [KMR93], which implies a change in the convergence rate as the noise effect approaches zero.

A different local interaction framework is considered by [Blu93]. It is assumed that

[3] *Mutation rate* is the probability of a mistake.

agents are located on a lattice, and each agent interacts directly with only a finite set of neighbors. Several stochastic methods for strategy revision are considered including best-reply and perturbed best-reply.[4] It is shown that in the case of two-player and two-action coordination games where a risk-dominant equilibrium exists, the log-linear strategy revision process is ergodic and converges to the risk-dominant equilibrium as the noise effect vanishes and the time goes to infinity.

2.3.3 Local interactions with migration

Several game theorists have investigated the effect of migration (or endogenous location interactions) in equilibrium selection. It is reasonable to assume, at least in sociological systems, that agents have discretion with regard to location choice, and hence some freedom in choosing their neighbors. In other words, rather than fixing exogenously the pattern of interaction, we would like to consider the case when the location and strategy co-evolve. It can be shown that migration can allow the payoff-dominant equilibrium to prevail.

More specifically, [Rob93] studies a model motivated by biological evolution. The population is partitioned into a finite set of sub-populations that grow independently at rates that reflect the payoffs earned by the agents within them. At exogenously fixed intervals, all populations become extinct and are re-populated by small groups randomly drawn from the preceding generation. When the time between extinction events is sufficiently long, the populations in which the efficient strategy is played grow arbitrarily large relative to other populations.

Reference [Oec99] studies an evolutionary model where agents initially are distributed over a given set of independent locations (or "cities"), and, over time, may *freely* adjust both their strategic and locational decisions. They are doing so by

[4] In particular, the *log-linear model* or *smooth best-reply* is considered, where the log of the probabilities of choosing a given strategy is proportional to the payoff difference between the two strategies.

playing a best-reply both in terms of their location and strategy. Assuming that all conventions are represented at the start of the process (i.e., are adopted by some city), [Oec99] shows that the payoff-dominant equilibrium will prevail throughout. The intuitive reason why this occurs is that any agent, when given the opportunity to adjust its location and strategy, will immediately prefer a city where the efficient convention is played.

A similar interaction model is considered by [Ely02]. Ely does not assume that all conventions are initially present at some location. Instead, agents' decisions are subject to mistakes with some small probability. As in the scenario considered in [Oec99], agents who are playing the non-efficient equilibrium will migrate to a location playing the efficient convention when the opportunity arises, and hence the efficient convention will prevail throughout.

A different model from the models of [Oec99, Ely02] is considered by [BV04]. Agents adjust both location and strategy as in the papers [Oec99, Ely02], however, these opportunities never arrive simultaneously. Hence, an agent who receives the opportunity to migrate will not be sure that it will be able to migrate to the appropriate location. This uncertainty has the consequence that the model no longer produces the efficiency conclusion of [Oec99, Ely02]. In particular, depending on the exact payoff structure, one may encounter (a) convention coexistence in the medium-run (when the possibility of mistakes in the decision process is absent), or (b) inefficiency in the long-run (when mutations are allowed).

Similar to the above frameworks is the context considered by [Hoj04]. In this paper, two-player and two-action coordination games are considered, when agents are allowed to change location, and when agents are interacting only with their neighbors. The paper studies best-reply dynamics in which agents choose an action from the underlying coordination game and a location from a finite set of locations. Each location has a circular topology. In contrast to [Ell93], which assumes the same

topology, agents endogenously choose their location which implies that the number of agents in each circle changes over time. At each period, an agent observes the distribution of play of its closest neighbors (not necessarily all the members of the location), and the average distribution of all the other locations. The agent is matched to a random subset of its closest neighbors, which introduces a *scale* effect, since the agent is interacting with more than one agent. This is the main difference with the previous models on mobility. Agents apply best-reply dynamics with a small probability of randomness. It is shown that in the long-run the process spends most of its time in states which are not necessarily efficient. The scale effect may have a significant impact on the emergence of non-efficient states, since a agent may prefer interacting with a large cluster of agents playing the non-efficient strategy than with a small cluster of agents playing the efficient strategy. Of course, this would also depend on the riskiness of each equilibrium.

2.3.4 Local interactions with evolving neighborhood

We have already discussed that in the context of two-player and two-action coordination game, [KMR93] and [You93] have shown that population of agents, who are subjected to small random perturbations in their strategy choices, tend in the long run to coordinate on risk-dominant strategies as defined by [HS88].

However not in every situation do agents face each other with equal probability (which corresponds to a uniform matching rule). Instead agents may interact through a specific interaction pattern. As we discussed above, the results of [KMR93, You93] continue to hold when agents interact according to a certain *fixed* neighborhood structures as shown by [Ell93] (see also [You98]). This may lead to the conclusion that the risk-dominant equilibrium is the only reasonable convention for a society, even if it is non-efficient and not in the society's common interest.

Different conclusions may be derived when an endogenous network structure is

considered, i.e., when the network structure also changes with time. As we saw in the discussion about the effect of mobility, e.g. [Ely02] and [MSS01], the network structure is endogenized through locational choices (agents have discretion over the neighborhood they want to locate themselves). Conditions can be derived under which the efficient equilibrium is the one that is reached by a society, even when it is not risk-dominant.

Such a model will require to sever all old ties, form new ties and switch strategies simultaneously. However, there are situations where agents can choose which links to form in a more discrete manner and without necessarily having to uproot all previous relationships. Reference [JW02b] considers a model where agents have the discretion over which links to form and which to sever, without changing their location. In particular, at each time a potential link is selected according to some probability distribution, and is formed or severed according to the myopic interest of both agents. In parallel, an agent is selected randomly to update its strategy according to a best-reply to the current network structure and the previous play. Agents are assumed to play a coordination game with the agents they are directed linked to.

Such a modification has a large impact on the way that play changes from one strategy to another, and thus leads to different results regarding the states that survive through time (*stochastically stable states*). In particular, [JW02b] shows that some of these states are neither efficient nor risk-dominant. Which of those states survive will depend on the specifics, such as the relative benefits of the play of different actions, the structure of the costs to links, and the number of agents in the society.

Reference [GV05] also derives similar conclusions when structure is evolving. The main difference with the model of [JW02b] is that links are bidirectional and can be formed by a single agent (i.e., it is a *noncooperative* link formation model). It is shown that if costs of forming links are below a certain threshold then agents coordinate on the risk-dominant action, while if costs are above this threshold then they coordinate

on the efficient action. Also, these findings are robust to modifications in the link formation process, different specifications of link formation costs, alternative models of mutations as well as the possibility of interaction among indirectly connected agents.

Reference [DGJ04] also considers a large population coordination game where agents are distributed spatially, and both the actions of the agents and the communication network between these agents evolve over time. The setting that [DGJ04] considers is similar to the setting of [JW02b]. The only differences are (a) agents are placed on a circle, although they can still create their own neighborhood by forming and severing links with other agents, and (b) the cost assigned to a link among two agents is proportional to the distance of the agents on the circle. Agents react myopically to their environment by deciding about (a) which strategy to play against their neighbors, and (b) which links to form. Links are formed using the link formation model of [JW02b], while strategies of play are adjusted by a myopic best-reply to the strategy of the neighborhood at the previous time instant, which is similar to the model of [Ell93], where a small possibility of mistakes is also considered. Reference [DGJ04] shows that the risk-dominant convention is the unique stochastically stable convention, meaning that it will be observed almost surely when the mistake probabilities are small.

2.3.5 The effect of communication

The effect of signals in equilibrium selection of communication games is investigated in several papers. Evolutionary dynamics with signals is found to have dramatically different dynamics from the same game without signals. Signals are able to change the stability properties of equilibria, the size of their basin of attraction, and create new polymorphic equilibria.

A nice example to see the effect of signals and communication on equilibrium selection is to consider a population of agents that are either programmed to play

strategy A or B in the game shown in Table 2.3, that is known as the *Prisoner's Dilemma*.

	$2.A$	$2.B$
$1.A$	$2, 2$	$0, 3$
$1.B$	$3, 0$	$1, 1$

Table 2.3: The Prisoner's Dilemma

The agents do not change their strategies, but instead replicate according to their "*success*" (or payoff) when they encounter other agents. This model has been extensively used in evolutionary game theory [Smi82]. It has also been shown that in this game, strategy A cannot spread throughout the whole population when there is a small possibility that a mutant behavior is generated.[5] The main reason for that is that an agent B that enters such a population can do better when it encounters agents playing A.

As [Rob90] points out, if there is a signal that is not used by the population, a mutant could invade by using this signal as a "*secret handshake*".[6] Mutants would defect against the natives and cooperate with each other. They would then do better than natives and would be able to invade. Without signals, a population of defectors in the Prisoner's Dilemma would be evolutionary stable. With signals this is no longer true. However, this does not mean that signals establish cooperation in the Prisoner's Dilemma.

Let us now consider the Aumann's Stag-Hunt (Assurance Game) of Table 2.4 in the context of evolutionary dynamics.

	A	B
A	$9, 9$	$0, 8$
B	$8, 0$	$7, 7$

Table 2.4: Aumann's Stag-Hunt

[5]Such a strategy is not an *evolutionary stable strategy* (ESS) [Smi82].
[6]This idea is also in accordance with the notion of *salient* equilibrium (Schelling's focal point).

In this game there are two evolutionary stable strategies (*hunt stug* (A) or *hunt hare* (B)). Suppose there are two available signals. In this case, a strategy specifies (a) which signal to send, (b) what act to do if signal 1 is received, and (c) what act to do if signal 2 is received. In this new game, there is an evolutionary stable strategy which is an entirely new equilibrium created by the signals. This is the state of the population in which: 50% sends signal 1, plays B when they receive 1, plays A when they receive A, and 50% sends signal 2, plays A when they receive signal 1, plays B when they receive signal 2 (see Chapter 5 of [Sky04], or [Sky02]). Note that these two strategies cooperate with each other producing outcome (A, A), but not with themselves. In a population that has only these two strategies, the replicator dynamics must drive them to the 50/50 equilibrium. This state will be evolutionary stable.

The question that arises is whether this new equilibrium plays a significant role in evolutionary dynamics. An interesting point is to consider how frequent such an equilibrium arises. References [Sky02, Sky04] performed a large number of simulations to measure the basin of attraction of this equilibrium. These papers further compared it with the basin of attraction of the equilibria *all hunt stag* at equilibrium and *all hunt hare* at equilibrium. The results were that the new polymorphic equilibrium has a non negligible basin of attraction.[7] On the other hand, the basin of attraction of *all hunt stag* at equilibrium was increased significantly. Without signaling, the basin of attraction of *all hunt stag* was smaller than the basin of attraction of *all hunt hare*.

Note that in an equilibrium where everyone is hunting stag or hare, signals carry no information. Still signals have something to do with the basin of attraction of these equilibria. In particular, [Sky02, Sky04] conclude that *transient information matters* since it is important in determining the eventual outcome of the evolutionary process.

[7]The size of its basin of attraction was as large as the basin of attraction of *all hunt hare* at equilibrium.

2.3.6 The effect of dynamics

Different forms of learning dynamics has been applied for strategy selection, including aspiration and imitation learning dynamics. A modification of the model of [KMR93] is considered by [RV96]. As mentioned above, the model of [KMR93] assumes that each agent plays a best-reply based on the payoff it would have received had it played against all other agents at once. Reference [RV96] considers instead the case where, at each iteration, an agent is randomly matched to play the game only once. Agents imitate the strategy that produced the largest average payoff among the agents that applied it at the previous step. The conclusions are quite different from those of the model of [KMR93]. In the class of two-player and two-action coordination games, the payoff-dominant equilibrium is selected even if it is not risk-dominant. Furthermore, convergence to the invariant distribution is relatively fast compared to the convergence in [KMR93].

Reference [BS97] examines the robustness of the results of [KMR93, You93] under a different learning process where mistakes are not necessarily negligible (as considered in the papers of [KMR93, You93] in the limiting behavior). In particular, each member of the population is characterized by a strategy (among two available strategies). An agent is selected to revise its strategy according to some probability which is independent across the population and across time. The strategy revision process assumes that pairs of agents are randomly drawn to play a two-action game, which implies that each agent has played the game infinite amount of times and with a distribution of agents that reflects the distribution of strategies in the population. Then, when the agent receives the opportunity to revise its strategy, it recalls its average realized payoff in the last period (learning period) and compares it with an *aspiration level*. A small probability of mistakes is also present. Depending on the specifics of the strategy revision process, it is possible for the payoff-dominant equilibrium or the risk-dominant equilibrium to be selected, when the size of the

population goes to infinity and the noise effect goes to zero. In fact, *the closer the revision process to best-reply dynamics the more likely it is to select the risk-dominant equilibrium.*

A similar approach that has been used to predict behavior in coordination games with aspiration learning algorithms is by [CM05]. In this model, agents base their decisions only on the rewards received (i.e., *payoff-based* algorithms). An agent continues playing an action if and only if the reward received is *strictly* greater than the agent's *aspiration level* (i.e., its average over all its previous payoffs). In case an efficient action profile has not been played in the past, the learning algorithm will get trapped in a non-efficient equilibrium. In [CM05], a small imperfection is added to agents' decisions. In particular, each agent experiments with different actions when the reward received is close enough to its aspiration level, hence avoiding the possibility that the algorithm converges to a non-efficient equilibrium. However, the analysis in [CM05] does not discuss the emergence of stable oscillations in agents' responses, which is the effect of the introduced imperfection in the decision rule. These oscillations do not die away unless agents average their payoffs over the whole history (i.e., when there is no discount).

References [SP00, Sky04, Sky07] investigate under which conditions local interaction can reinforce the payoff-dominant equilibrium when agents play a Stag-Hunt game. In particular, agents create links according to a reinforcement learning scheme that reinforces links with high rewards. The reward a agent receives is higher the more interactions it has (*reciprocal* benefits or symmetrized reinforcement). Also, as a model of interaction, the Stag-Hunt game is considered and the strategy is revised based on *imitate-the-best* strategy among neighbors. Simulations showed that when the imitation was "fast" most of the trials converted to "*all playing Hare*", while in the case of "slow" imitation most of the trials converted to "*all playing Stag*".

Reference [Sky04] questions whether or not such a conclusion is general or depends

on the specific choice of adaptive dynamics for both the structure and the strategy. For example, we may consider as well the case where strategy and structure is updated according to reinforcement learning, since the members of the population are quite naive to apply selection based on imitation. On the other hand, maybe the members of the population are more-strategic minded and they best-respond to their environment.[8] Reference [Sky04] gets to the conclusion that when we pair structure dynamics, which is based on either reinforcement or imitate-the-best or best-reply, with slow strategy dynamics based on imitate-the-best, we end up with all Stag Hunters (efficient equilibrium). On the other hand, when we change the strategy dynamics, we does not observe the same conclusions.

2.3.7 Discussion and open problems

Several issues are related to equilibrium selection in coordination games, as pointed out by the previous literature, including

- Multiplicity of long-term outcomes;
- Robustness of each equilibrium in noise;
- Effect of transient phenomena (such as learning dynamics, state-dependent noise and signals) on reinforcing the efficient outcomes.

We saw that long-term predictions in coordination games are not necessarily unique. Instead multiple equilibria may be observed, including equilibria that are not payoff-dominant. However, we need to note that the predictions of the learning processes depend on the specifics of the learning model (the off-equilibrium behavior).

For example, in the model of [You93], agents are matched to play the game only once and when each agent best-replies to a large sample of the history, then the risk-dominant equilibrium emerges as the prediction. On the other hand, in the model

[8] It has been applied to model network formation in the papers of [BG00] and [Wat01].

of [RV96], agents are matched to play the game only once, but imitate the most successful strategy of the previous time stage. In that case, the payoff-dominant equilibrium emerged. This remark is also supported by [BS97], where it was shown that the closer the learning process is to best-reply dynamics, the more likely it will select a risk-dominant equilibrium. Finally, similar arguments are stated by [Sky04], which observed that different forms of dynamics may lead to different conclusions.

Of course, this is not to say that certain learning dynamics can lead to certain outcomes. The importance of these observations are that under certain rules of local behavior, certain off-equilibria (transient) phenomena may be well exploited to produce desirable global behavior. This is very well illustrated by the effect of signaling as presented in Section 2.3.5, where it was shown that under the same learning dynamics, but with the presence of signals (with no meaning), long-term behavior was totally different than the behavior without signals.

2.4 Remarks

We conclude that learning dynamics in coordination games (and more generally in coordination problems) are characterized by an asymptotic behavior that is difficult to predict. The predictions derived by the existing literature depends highly on the underlying assumptions, that is the learning dynamics and the information structure. Different conclusions are derived when slight modifications are made in the underlying model, showing the *fragility* of models of multiagent simulation. Furthermore, the effect of transient phenomena seems to be underestimated in a large amount of the existing literature, since in several cases, such as in the presence of signals or when the communication structure changes with time, the predictions change dramatically. Our goal is to analyze coordination problems within a very simple framework that assumes minimal and local information to each agent, which is suitable for engineering applications. In parallel, we want to explore locally off-equilibrium behavior to

control the collective asymptotic behavior. Through this simple framework, we seek to produce design tools for distributed control of multiagent systems in coordination problems.

CHAPTER 3

Learning Automata

3.1 Introduction

In this chapter, we introduce the basic form of the learning dynamics that we will consider in the remainder of the dissertation. The learning dynamics considered belongs to the general class of reinforcement learning discussed in Section 2.2.4 and is a special form of the *learning automata* [NT89]. We selected this form of dynamics because of its distributed nature and simplicity, since it assumes minimal amount of information available to each agent, namely its *own* previous *actions* and previous *rewards*.

In this chapter, we will also describe the asymptotic behavior of the learning automata. In particular, we will consider two forms of algorithms, the ones where experience is averaged throughout history (*diminishing step-size* algorithms) and the ones where experience is discounted (*constant step-size* algorithms). We will present the advantages of each one of these algorithms in terms of their asymptotic behavior. Finally, we will introduce a new form of learning automata where decisions are exogenously perturbed, and we will illustrate its utility in equilibrium selection.

3.2 Variable structure stochastic automata

Variable-structure stochastic automata update the *strategy* or the action probabilities on the basis of an *input*. Reference [VV63] was the first to suggest automata that

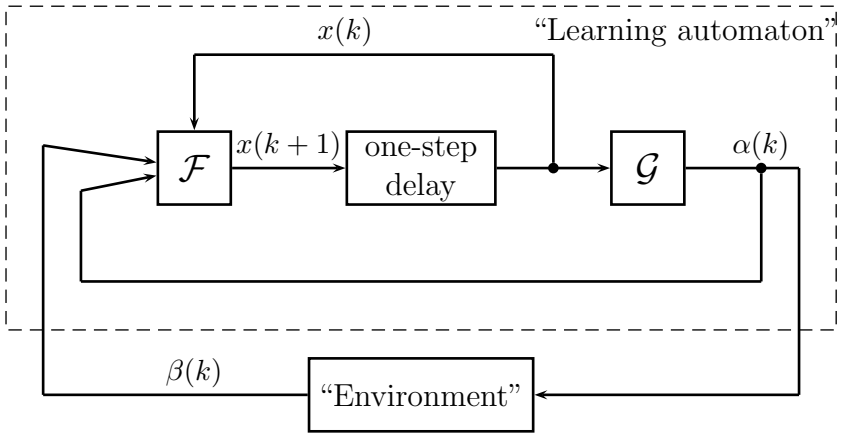


Figure 3.1: Learning automaton.

update transition probabilities. The automata are represented by the quintuple:

$$\{\mathcal{X}, \mathcal{A}, \mathcal{B}, \mathcal{F}, \mathcal{G}\}$$

where $\mathcal{X}$ is the set of internal states or strategies, $\mathcal{B}$ is the set of inputs, $\mathcal{A}$ is the set of outputs, $\mathcal{F} : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \rightarrow \mathcal{X}$ is a function that maps the current state and current input to the next state, i.e.,

$$x(k+1) = \mathcal{F}(x(k), \alpha(k), \beta(k))$$

and $\mathcal{G} : \mathcal{X} \rightarrow \mathcal{A}$ a function that maps the current state into the current output, i.e.,

$$\alpha(k) = \mathcal{G}(x(k)).$$

In such an automaton, the input and the current state together determine the next state as well as the current output. A graphical representation of a learning automaton is provided in Fig. 3.1.

3.3 Reinforcement schemes

We will consider automata with a finite set of outputs $\mathcal{A}$ that will be referred as *actions*. In general terms a reinforcement scheme can be represented by a mapping $\mathcal{F}$, where the state $x \in \Delta(|\mathcal{A}|)$ corresponds to the strategy, i.e., the probability distribution over the actions in $\mathcal{A}$. Therefore, if $x_i(k)$ corresponds to the probability of action $\alpha_i \in \mathcal{A}$ at time k, the automaton selects action α_i with probability $x_i(k)$.

The basic idea behind a reinforcement scheme is a rather simple one. If the automaton selects an action i at instant k and a favorable input[1] results, the action probability $x_i(k)$ is increased and all the other components of $x(k)$ are decreased. For an unfavorable input, $x_i(k)$ is decreased and all the other components are increased. These changes in $x_i(k)$ are known as *reward*[2] and *penalty*, respectively.

The precise manner in which $x(k)$ is changed depending on the action α_i performed at stage k and the response $\beta(k)$ of the environment, completely defines the reinforcement scheme. This, in turn, determines the resulting Markov process and hence the behavior of the overall system.

One of the reinforcement schemes that is extensively used is the *linear reward-inaction* scheme, whose description follows.

3.3.1 Linear Reward-Inaction (L_{R-I}) scheme

The basic idea of this scheme is not to change probabilities whenever an unfavorable response results from the environment. Following a favorable response, however, the probability of an action is increased. The L_{R-I} scheme was considered first in mathematical psychology by [Nor68] but was later independently conceived and introduced into the engineering literature by [SN69].

[1] Defined as output of the environment.
[2] A term derived from psychology.

Assume that there are only two actions, i.e., $\mathcal{A} = \{\alpha_1, \alpha_2\}$. Then, according to the L_{R-I} scheme the probability of selecting action α_1 is updated according to:

$$\begin{array}{rcll} x_1(k+1) & = & x_1(k) + \epsilon(1 - x_1(k)) & \alpha(k) = \alpha_1 \quad \beta(k) = 0 \\ x_1(k+1) & = & x_1(k) & \alpha(k) = \alpha_1 \quad \beta(k) = 1 \\ x_1(k+1) & = & (1-\epsilon)x_1(k) & \alpha(k) = \alpha_2 \quad \beta(k) = 0 \\ x_1(k+1) & = & x_1(k) & \alpha(k) = \alpha_2 \quad \beta(k) = 1 \end{array} \tag{3.1}$$

We may use the more compact form:

$$x(k+1) = x(k) + \epsilon R(\beta(k))(\alpha(k) - x(k)) \tag{3.2}$$

where $R : \mathcal{B} \to \{0, 1\}$ is the reward function. Moreover, $R(\beta(k)) = 1$ when $\beta(k) = 0$ (favorable response) and $R(\beta(k)) = 0$ when $\beta(k) = 1$ (unfavorable response). Also, here $\alpha(k) = e_1$ when action α_1 is performed, and $\alpha(k) = e_2$ when action α_2 is performed.

From equation (3.1) it follows that the probability $x_1(k)$ is increased if action α_1 is performed and results in a favorable response, is unchanged if an unfavorable response results when α_1 or α_2 is performed and is decreased only when the other action α_2 is performed and results in a favorable response.

Since we are going to use the recursive form (3.2) extensively, we will be using the same notation, $\alpha \in \mathcal{A}$, to refer to an element of $\mathcal{A}$ either in terms of an index over $\mathcal{A}$ or a vertex of $\Delta(|\mathcal{A}|)$.

3.3.2 Modified Linear Reward-Inaction ($\tilde{L}_{R-I}$) scheme

We consider a slightly modified linear reward-inaction scheme, according to which every action is successful with probability 1, however, the reward may vary depending

on the action profile.[3]

In the case of $|\mathcal{A}|$ actions, this scheme can be expressed recursively by the form:

$$x(k+1) = x(k) + \epsilon R(\alpha(k))[\alpha(k) - x(k)] \tag{3.3}$$

where $R : \mathcal{A} \to [0, \infty)$ is the reward function. Moreover, let $R(\alpha(k)) = d_i$ with probability one when action $\alpha(k) = \alpha_i$ is performed.

The performance of the automaton can be determined by the asymptotic behavior of $E[R(\alpha(k))|x(k)]$, which is given by

$$E[R(\alpha(k))|x(k)] = \sum_{i=1}^{|\mathcal{A}|} x_i(k) d_i = d^{\mathrm{T}} x(k) \tag{3.4}$$

We can show that:

Claim 3.3.1 *If the automaton uses the $\tilde{L}_{R-I}$ scheme, then*

$$\Delta x_i(k) \triangleq E[x_i(k+1) - x_i(k)|x(k)] = \epsilon x_i(k) \sum_{j=1}^{|\mathcal{A}|} x_j(k)(d_i - d_j)$$

Proof. We have

$$\begin{aligned}
\Delta x_i(k) &= E[x_i(k+1) - x_i(k)|x(k)] = E[\epsilon R(\alpha(k))(\alpha(k) - x_i(k))|x(k)] \\
&= \epsilon d_i (1 - x_i(k)) x_i(k) + \epsilon \sum_{j=1, j \neq i}^{|\mathcal{A}|} d_j (0 - x_i(k)) x_j(k) \\
&= \epsilon x_i(k) \sum_{j=1}^{|\mathcal{A}|} x_j(k)(d_i - d_j)
\end{aligned}$$

□

[3] In that case, a zero reward will implicitly correspond to an unfavorable action.

The conditional expectation of the change in payoff defined as

$$\Delta R(k) \triangleq E[R(\alpha(k+1)) - R(\alpha(k))|x(k)]$$

satisfies:

$$\Delta R(k) = d^{\mathrm{T}} \Delta x(k).$$

Claim 3.3.2 *$\Delta R(k) = \epsilon x^{\mathrm{T}} \tilde{D} x/2$, where the elements of the matrix $\tilde{D}$ are given by $\tilde{D}(i,i) = 0$, $\tilde{D}(i,j) = (d_i - d_j)^2$.*

Proof. We have

$$\begin{aligned}
\Delta R(k) &= d^{\mathrm{T}} \Delta x(k) \\
&= \epsilon \sum_{i=1}^{|\mathcal{A}|} d_i x_i \sum_{j=1, j\neq i}^{|\mathcal{A}|} x_j (d_i - d_j) \\
&= \epsilon \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1, j>i}^{|\mathcal{A}|} x_i x_j (d_i - d_j)^2 \\
&= \epsilon x^{\mathrm{T}} \tilde{D} x/2
\end{aligned}$$

where the elements of the matrix $\tilde{D}$ are given by $\tilde{D}(i,i) = 0$, $\tilde{D}(i,j) = (d_i - d_j)^2$. □

Note that the matrix $\tilde{D}$ is positive semidefinite.

3.4 Convergence results in a stationary environment for $\tilde{L}_{R-I}$

The convergence properties of $\tilde{L}_{R-I}$ can be described by as follows.

Proposition 3.4.1 (Constant step size: Convergence) *Assume that d_i, for $i = 1, 2, ..., |\mathcal{A}|$, are distinct and nonnegative. The Markov process $\{x(k)\}_k$ that corresponds to the $\tilde{L}_{R-I}$ reinforcement scheme converges w.p.1 to the set of unit $|\mathcal{A}|$-vectors.*

Proof. Let $R_{\max} \triangleq \max_{\alpha \in \mathcal{A}} R(\alpha)$. The stochastic process $\{R_{\max} - R(\alpha(k))\}_{k \in \mathbb{N}}$, is a nonnegative supermartingale, since

$$E[[R_{\max} - R(\alpha(k+1))] - [R_{\max} - R(\alpha(k))]|x(k)] = -\Delta R(k) \leq 0.$$

Hence by the martingale convergence theorem A.1.1, $\{R(\alpha(k))\}_k$ converges to a random variable w.p.1. It also follows from Corollary A.1.1 that

$$\lim_{k \to \infty} \Delta R(k) = 0 \quad \text{w.p.1.} \tag{3.5}$$

This is possible only if either

$$\left.\begin{array}{rll} & (i) & \epsilon \to 0 \\ \text{or} & & \\ & (ii) & x(k)^{\mathrm{T}} \tilde{D} x(k) \to 0 \end{array}\right\} \quad \text{w.p.1.} \tag{3.6}$$

Since we are considering a constant step size, condition (ii) is satisfied. Thus,

$$x(k)^{\mathrm{T}} \tilde{D} x(k) = 2 \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1, j>i}^{|\mathcal{A}|} x_i x_j (d_i - d_j)^2 = 0.$$

All the terms on the r.h.s. are of the same sign and further the coefficients $(d_i - d_j)^2$ are nonzero when $i \neq j$ by assumption. Hence (ii) is satisfied only when $x_i(k)x_j(k) \to 0$ for all $i \neq j$. This in turn means that $x(k)$ converges to a unit vector w.p.1. Thus, when (ii) is satisfied,

$$\lim_{k \to \infty} x(k) \in \{e_i : i = 1, 2, ..., |\mathcal{A}|\}, \tag{3.7}$$

which concludes the proof. □

In general it does not appear possible to prove convergence with probability one when there are roots of the martingale equation other than those corresponding to absorbing states.

Further conclusions regarding the convergence of the scheme can be derived when the specific form of the step size sequence ϵ is known. This has been considered in [Art93] for the $\tilde{L}_{R-I}$ where the step size sequence is decreasing instead of constant considered here. In particular, we assume the general form of the step size sequence:

$$\epsilon(k) = \frac{1}{ck^{\nu} + 1}, \tag{3.8}$$

for some $c > 1$.

When the step size sequence approaches zero, we may be able to exclude convergence to an interior point of the probability space $\Delta(|\mathcal{A}|)$ depending on how fast the step size approaches zero. This can be shown by applying Theorem B.1.2 of [NH76] on martingales.

Lemma 3.4.1 (Diminishing step size: Nonconvergence) *Consider the reinforcement scheme $\tilde{L}_{R-I}$ with step size sequence that is given by (3.8) with $\nu \in [0, 1]$. Let h be any interior point of the probability simplex $\Delta(|\mathcal{A}|)$. If $\mathcal{B}_{\varepsilon}(h)$ is an open ε-neighborhood of h, then*

$$P[\lim_{k\to\infty} x(k) = h] = 0.$$

Proof. For any $x(k) \in \mathcal{B}_{\varepsilon}(h)$, there exists $\kappa > 0$ such that $\min_{x\in\mathcal{B}_{\varepsilon}(h)} x(k)^{\mathrm{T}} \tilde{D} x(k)/2 = \kappa$. Hence,

$$\Delta R(k) \geq \kappa\epsilon(k).$$

Note also that $R(k) = \sum_{i=1}^{|\mathcal{A}|} d_i x_i(k) \leq R_{\max}$ for any $x(k) \in \mathcal{B}_{\varepsilon}(h)$ and for some $R_{\max} > 0$.

Let us define the function $V(k) = R_{\max} - R(\alpha(k))$, which is a nonnegative function in $\mathcal{B}_\varepsilon(h)$. The function V satisfies:

$$E[V(k+1) - V(k)|x(k)] = -E[R(\alpha(k+1)) - R(\alpha(k))|x(k)] = -\Delta R(k) \leq -\kappa\epsilon(k),$$

where $\kappa\epsilon(k) > 0$ and $\sum_{k=1}^{\infty} \kappa\epsilon(k) = \infty$. Therefore, by Theorem B.1.2 of [NH76] (Appendix A), we conclude that the process $x(k)$ must exit $\mathcal{B}_\varepsilon(h)$ in finite time with probability one. $\square$

The following theorem is a direct consequence of the proof of Proposition 3.4.1 and Lemma 3.4.1.

Theorem 3.4.1 (Diminishing step size: Convergence w.p.1) *Consider the reinforcement scheme $\tilde{L}_{R-I}$ and assume that d_i, $i = 1, 2, ..., |\mathcal{A}|$, are distinct and nonnegative. Let the step size given by (3.8) with $\nu \in [0, 1]$. Then, the Markov process $\{x(k)\}_k$ converges to the set of vertices, $\{e_i : i = 1, 2, ..., |\mathcal{A}|\}$, with probability one.*

3.5 Mathematical formulation of automata games

Let n automata be assumed to take part in a game. The automaton $i \in \mathcal{I} \triangleq \{1, 2, ..., n\}$ can be described by a quintuple:

$$\{\mathcal{X}_i, \mathcal{A}_i, \mathcal{B}_i, \mathcal{F}_i, \mathcal{G}_i\},$$

where $\mathcal{X}_i$ is the set of internal states of automaton $i \in \mathcal{I}$, $\mathcal{B}_i$ is the set of inputs of automaton i, $\mathcal{A}_i$ is the set of outputs of automaton i, $\mathcal{F}_i : \mathcal{X}_i \times \mathcal{A}_i \times \mathcal{B}_i \to \mathcal{X}_i$ is a function that maps the current state and the current input into the next state, i.e.,

$$x_i(k+1) = \mathcal{F}_i(x_i(k), \alpha_i(k), \beta_i(k))$$

and $\mathcal{G}_i : \mathcal{X}_i \to \mathcal{A}_i$ a function that maps the current state into the current output, i.e.,

$$\alpha_i(k) = \mathcal{G}_i(x_i(k)).$$

Theoretically, it is possible for automaton i to correspond to different reinforcement schemes for different values of i. For purposes of analysis, it is found more convenient to use identical learning algorithms for all the automata.

3.5.1 Games of $\tilde{L}_{R-I}$ automata

Let all the automata taking part in the game use the $\tilde{L}_{R-I}$ scheme. The automata update their probability distributions over their action sets at every instant based on the reward received from the environment.

When the action profile is $\alpha(k) = (\alpha_1, \alpha_2, ..., \alpha_n)$ let the reward of agent i be $R_i(\alpha(k))$. Also $x_i(k)$ denotes the probability distribution governing the choice of actions of automaton i at the kth stage. Then the expected reward of the automaton i at each stage is given by

$$E[R_i(\alpha)|x] = \sum_{\alpha_1,\alpha_2,...,\alpha_n} x_{1\alpha_1} x_{2\alpha_2} \cdots x_{n\alpha_n} R_i(\alpha_1, \alpha_2, ..., \alpha_n). \tag{3.9}$$

where here $x_{i\alpha_i}$ denote the α_ith entry of the vector x_i. Using the definitions above, the game can now be described by a Markov process whose state space is the product simplex space, that is $\mathcal{X} = \times_{i\in\mathcal{I}} \Delta(|\mathcal{A}_i|)$.

At every stage k, based on the probability distributions $x_1(k), x_2(k), ..., x_n(k)$, the automata choose a play $\alpha(k)$ and based on the response of the environment as well as the learning schemes used $x(k) \triangleq (x_1(k), x_2(k), ..., x_n(k))$ evolves in $\mathcal{X}$ according

to the recursion:

$$x_i(k+1) = x_i(k) + \epsilon(k) R_i(\alpha(k))[\alpha_i(k) - x_i(k)]. \tag{3.10}$$

3.6 Games with identical interests for $\tilde{L}_{R-I}$

Consider n automata, each operating independently and in total ignorance of the other automata. The automata are involved in a game Γ. The game Γ is of identical interests if, for any action profile $\alpha \in \mathcal{A}$, the payoffs are the same for all agents, i.e.,

$$R_i(\alpha) \equiv R(\alpha) \quad \text{for all } i \in \mathcal{I}.$$

We assume that each automaton is using the $\tilde{L}_{R-I}$ reinforcement scheme. The goal is to determine the asymptotic behavior of the sequential game.

3.6.1 Two-player case

Assume that both agents use the $\tilde{L}_{R-I}$ scheme and are involved in a game Γ with action sets $\mathcal{A}_1 = \{1, 2\}$ and $\mathcal{A}_2 = \{1, 2\}$. The game can be represented by a 2×2 matrix D whose (i, j) element d_{ij} is defined as

$$d_{ij} \triangleq R(\alpha = (i, j)), \quad i, j \in \{1, 2\}.$$

Our goal is to derive asymptotic properties of the processes $\{x(k)\}_k$. We will demonstrate that $\{R(\alpha(k))\}_k$ is a submartingale. This in turn will allow us for deriving conclusions for $\{x(k)\}_k$.

The conditional expected payoff for either one of the agents at each stage is given

by

$$E[R(\alpha)|x] = \sum_{i=1}^{|\mathcal{A}_1|} \sum_{j=1}^{|\mathcal{A}_2|} x_{1i} x_{2j} d_{ij} = x_1^{\mathrm{T}} D x_2.$$

Define

$$\begin{aligned}
\delta x_1(k) &\triangleq x_1(k+1) - x_1(k) \\
\delta x_2(k) &\triangleq x_2(k+1) - x_2(k) \\
\Delta x_1(k) &\triangleq E[\delta x_1(k) | x_1(k), x_2(k)] \\
\Delta x_2(k) &\triangleq E[\delta x_2(k) | x_1(k), x_2(k)].
\end{aligned}$$

Claim 3.6.1 *The conditional expectation of the change in payoff for each agent,*

$$\Delta R(k) \triangleq E[R(\alpha(k+1)) - R(\alpha(k)) | x_1(k), x_2(k)],$$

satisfies

$$\Delta R(k) = \Delta x_1(k)^{\mathrm{T}} D x_2(k) + x_1(k)^{\mathrm{T}} D \Delta x_2(k) + E[\delta x_1(k)^{\mathrm{T}} D \delta x_2(k) | x_1(k), x_2(k)].$$

Proof. See Appendix D.1.1. □

Omitting stage index k for conciseness of notation, we have for the expected incremental gain

$$\Delta R = \Delta {x_1}^{\mathrm{T}} D x_2 + x_1^{\mathrm{T}} D \Delta x_2 + E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2]. \tag{3.11}$$

The first two terms in (3.11) correspond to the incremental gain due to each agent when the action probabilities of the other agent are constant. These are equivalent to cases where each agent is operating in a stationary environment discussed earlier.

In particular, for given $x_1, x_2 \in \Delta(2)$, define

$$\overline{v}_1 \triangleq Dx_2 \quad \text{and} \quad \overline{v}_2 \triangleq D^{\mathrm{T}} x_1.$$

Proposition 3.6.1 *$\Delta {x_1}^{\mathrm{T}} D x_2 = \epsilon x_1^{\mathrm{T}} \tilde{D}_1 x_1/2$ and $x_1^{\mathrm{T}} D \Delta x_2 = \epsilon x_2^{\mathrm{T}} \tilde{D}_2 x_2/2$, where the elements of the matrix $\tilde{D}_s$, $s = 1, 2$, are given by $\tilde{D}_s(i,i) = 0$, $\tilde{D}_s(i,j) = (\overline{v}_{si} - \overline{v}_{sj})^2$, where $\overline{v}_{si}$, $s = 1, 2$, is the ith entry of the vector $\overline{v}_i$.*

Proof. We have

$$\Delta {x_1}^{\mathrm{T}} D x_2 = \Delta {x_1}^{\mathrm{T}} \overline{v}_1$$

where the jth entry of the vector Δx_1 is

$$\Delta x_{1j} = \epsilon \sum_{i=1}^{|\mathcal{A}_1|} x_{1j} x_{1i} (\overline{v}_{1j} - \overline{v}_{1i}).$$

Thus,

$$\Delta {x_1}^{\mathrm{T}} \overline{v}_1 = \epsilon \sum_{i=1}^{|\mathcal{A}_1|} \sum_{j=1, j>i}^{|\mathcal{A}_1|} x_{1i} x_{1j} (\overline{v}_{1i} - \overline{v}_{1j})^2 = \epsilon x_1^{\mathrm{T}} \tilde{D}_1 x_1/2$$

where the elements of the matrix $\tilde{D}_1$ are given by $\tilde{D}_1(i,i) = 0$, $\tilde{D}_1(i,j) = (\overline{v}_{1i} - \overline{v}_{1j})^2$. Similarly we can show that $\Delta {x_1}^{\mathrm{T}} \overline{v}_1 = \epsilon x_2^{\mathrm{T}} \tilde{D}_2 x_2/2$, where the elements of the matrix $\tilde{D}_2$ are given by $\tilde{D}_2(i,i) = 0$, $\tilde{D}_2(i,j) = (\overline{v}_{2i} - \overline{v}_{2j})^2$. $\square$

It therefore follows that the first two terms are of order ϵ, where ϵ is the algorithm step size, and are nonnegative. Hence, to show that $\{R(\alpha(k))\}_k$ is a submartingale it suffices to show that the third term is also nonnegative. The third term, $E[\delta {x_1}^{\mathrm{T}} D \delta x_2]$ can be interpreted as the interaction term, which includes the effect of the two automata acting simultaneously. It is evident that since the changes in action probabilities appear as a product in this term, it is of order ϵ^2. This term can be explicitly computed for a general identical interest game as follows:

Proposition 3.6.2 (2 automata with 2 actions) *For two automata and* $|\mathcal{A}_1| = |\mathcal{A}_2| = 2$ *actions, where both automata apply the* $\tilde{L}_{R-I}$ *scheme, we have*

$$E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2] = \epsilon^2 x_{11} x_{12} x_{21} x_{22} (d_{11} - d_{12} - d_{21} + d_{22})((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2).$$

Proof. See Appendix D.1.2. □

Note that with $\tilde{L}_{R-I}$ reinforcement scheme, the quantity $E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2]$ is nonnegative under certain conditions in the reward matrix D. For example, if $d_{11} > d_{21}$ and $d_{22} > d_{12}$, then $E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2] \geq 0$ for all x_1 and x_2 in $\Delta(2)$.

Similar to Proposition 3.6.2, for the case of multiple actions where both automata apply the $\tilde{L}_{R-I}$ scheme, we have:

Proposition 3.6.3 (2 automata with multiple actions) *For two automata and for any* $|\mathcal{A}_1|, |\mathcal{A}_2| \in \mathbb{N}$, $|\mathcal{A}_1|, |\mathcal{A}_2| \geq 2$, *where both automata apply the* $\tilde{L}_{R-I}$ *scheme, we have*

$$E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] = \epsilon^2 \sum_{i,j=1, i \neq j}^{|\mathcal{A}_1|, |\mathcal{A}_1|} \sum_{k,l=1, k \neq l}^{|\mathcal{A}_2|, |\mathcal{A}_2|} x_{1i} x_{1j} x_{2k} x_{2l} (d_{ik} - d_{jk} - d_{il} + d_{jl})((d_{ik})^2 - (d_{jk})^2 - (d_{il})^2 + (d_{jl})^2).$$

Proof. See Appendix D.1.3. □

We observe that for certain payoff structures, the sequence $\{R(k)\}_k$ will be a submartingale.

3.6.2 Example: pure coordination games

We consider here a pure coordination game, and characterize the sign of $\Delta R(k)$ for $k = 1, 2, ...$ The Typewriter game is the pure coordination game that is given in Table 3.1.

	$2.A$	$2.B$
$1.A$	$5, 5$	$1, 1$
$1.B$	$1, 1$	$2, 2$

Table 3.1: The Typewriter game.

In the Typewriter game of Table 3.1, we have $d_{11} = 5$, $d_{12} = 1$, $d_{21} = 1$ and $d_{22} = 2$. In this case, the payoff matrix for each agent is

$$D = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$$

Then, according to Proposition 3.6.2, the terms of order ϵ^2 are given by

$$E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2] = 25\epsilon^2 x_{11} x_{12} x_{21} x_{22} \geq 0.$$

Note that if $x_{11}, x_{12}, x_{21}, x_{22} \neq 0$, then $E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1, x_2] > 0$, which implies that the conditional expectation of the payoff change is positive as long as the probability of playing each of the two actions is non-zero.

3.6.3 Multiple player case

We may first consider the simple case of $n = 3$ agents with two actions each. Then, the conditional expectation of the change in payoff of either one of the three agents

will be

$$\Delta R = \sum_{s,l,m=1}^{2} E[x_{1s}^{+} x_{2l}^{+} x_{3m}^{+} - x_{1s} x_{2l} x_{3m} | x_1, x_2, x_3] R(\alpha_1 = s, \alpha_2 = l, \alpha_3 = m).$$

where for simplicity we use the superscript "+" to denote "next time stage." ΔR can be shown to have terms of the form, $\epsilon x_{1s} x_{2l} \delta x_{3m}$ (a single variational term), $\epsilon^2 x_{1s} \delta x_{2l} \delta x_{3m}$ (two variational terms), and $\epsilon^3 \delta x_{1s} \delta x_{2l} \delta x_{3m}$ (three variational terms).

Regarding the single variational terms of ΔR, denoted by $[\Delta R]_1$, we can show the following.

Proposition 3.6.4 $[\Delta R]_1 \geq 0$.

Proof. Note that

$$[\Delta R]_1 = \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{A}_i|} \Delta x_{ij} E[R(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_i = j, x]$$

where Δx_{ij} denote the jth entry of the vector Δx_i. Define the expected payoff of agent i when it selects action j as

$$\overline{v}_i(j, x) \triangleq E[R(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_i = j, x] \in \mathbb{R}_{+}^{|\mathcal{A}_i|}. \tag{3.12}$$

Then

$$[\Delta R]_1 = \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{A}_i|} \Delta x_{ij} \overline{v}_i(j, x).$$

We also have that the jth entry of the vector Δx_i is

$$\Delta x_{ij} = \epsilon \sum_{a \neq j} x_{ij} x_{ia} [\overline{v}_i(j, x) - \overline{v}_i(a, x)].$$

Thus,

$$\begin{aligned}
[\Delta R]_1 &= \epsilon \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{A}_i|} \sum_{a \neq j} x_{ij} x_{i\alpha} [\overline{v}_i(j,x) - \overline{v}_i(a,x)] \overline{v}_i(j,x) \\
&= \epsilon \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{A}_i|} \sum_{a > j} x_{ij} x_{ia} [\overline{v}_i(j,x) - \overline{v}_i(a,x)]^2 \\
&= \epsilon \sum_{i=1}^{n} x_i^{\mathrm{T}} \tilde{D}_i x_i / 2
\end{aligned}$$

where the elements of the matrix $\tilde{D}_i$ are given by $\tilde{D}_i(s,s) = 0$, $\tilde{D}_i(s,l) = [\overline{v}_i(s,x) - \overline{v}_i(l,x)]^2$ for $l \neq s$. Therefore, $[\Delta R(k)]_1 \geq 0$ for all $k = 1, 2, ...$ □

Regarding the terms of more than one variational terms, we can show the following.

Proposition 3.6.5 *If n automata, with multiple actions each, apply the $\tilde{L}_{R-I}$ scheme, then the terms of ΔR of order ϵ^n, denoted by $[\Delta R]_n$, satisfy*

$$\begin{aligned}
&[\Delta R]_n = \\
&\epsilon^n \sum_{s_1,k_1=1,s_1 \neq k_1}^{|\mathcal{A}_1|,|\mathcal{A}_1|} \cdots \sum_{s_n,k_n=1,s_n \neq k_n}^{|\mathcal{A}_n|,|\mathcal{A}_n|} x_{1s_1} x_{1k_1} x_{2s_2} x_{2k_2} \cdots x_{ns_n} x_{nk_n} \\
&\left[\sum_{\alpha_1 \in \{s_1,k_1\}} \cdots \sum_{\alpha_n \in \{s_n,k_n\}} (-1)^{\alpha_1 + ... + \alpha_n} (-1)^n R(\alpha_1, \alpha_2, ..., \alpha_n) \right] \cdot \\
&\left[\sum_{\alpha_1 \in \{s_1,k_1\}} \cdots \sum_{\alpha_n \in \{s_n,k_n\}} (-1)^{\alpha_1 + ... + \alpha_n} (-1)^n [R(\alpha_1, \alpha_2, ..., \alpha_n)]^n \right].
\end{aligned}$$

Proof. The proof follows similar steps with the proof of Proposition 3.6.3. □

The sign of the terms $[\Delta R]_n$ would depend on the reward function.

3.6.4 Convergence results for 2×2 pure coordination games

Based on Propositions 3.6.4–3.6.5, we may derive some general convergence results for 2×2 coordination games.

Theorem 3.6.1 (Convergence of 2×2 pure coordination games) *Let two automata play an identical interest game Γ with $|\mathcal{A}_1| = |\mathcal{A}_2| = 2$ actions and two strict pure Nash equilibria. Let all automata apply the $\tilde{L}_{R-I}$ scheme with step size $\epsilon(k) > 0$ such that $\sum_k \epsilon(k) = \infty$. Then, the process $\{x_i(k)\}_k$ converges to the set of vertices $\{e_j, j \in \mathcal{A}_i\}$ with probability one for every $i \in \mathcal{I}$.*

Proof. According to Proposition 3.6.4, $[\Delta R(k)]_1 \geq 0$ for all $k = 1, 2, ...$ By assumption, the game has two two strict pure Nash equilibria. These equilibria will correspond to the action profiles $\{(1,1),(2,2)\}$ or $\{(1,2),(2,1)\}$. In the first case, $d_{11} \geq d_{21}$ and $d_{22} \geq d_{12}$, which implies that $[\Delta R]_2 \geq 0$. In the second case, $d_{21} \geq d_{11}$ and $d_{12} \geq d_{22}$, which implies that $[\Delta R]_2 \geq 0$. Thus, $\Delta R(k) \geq 0$ for all $k = \{0, 1, 2, ...\}$, which implies that the process $\{R_{i,\max} - R_i(k)\}_k$ is a nonnegative supermartingale. From the martingale convergence theorem A.1.1, we conclude that the process $\{R_{i,\max} - R_i(k)\}_k$ converges w.p.1. From the Corollary A.1.1, we also conclude that it converges to the set of zeros of ΔR.

If $\epsilon(k) = \epsilon > 0$, then the process $\{x_i(k)\}_k$ converges to the set of vertices $\{e_j, j \in \mathcal{A}_i\}$ with probability one for every $i \in \mathcal{I}$.

For the more general case of $\sum_k \epsilon(k) = \infty$, we can apply Lemma 3.4.1 to show that there is probability zero of convergence to any strategy in the interior of $\mathcal{X} = \times_{i \in \mathcal{I}} \Delta(|\mathcal{A}_i|)$. Therefore, we conclude that the process $\{x_i(k)\}_k$ converges to the set of vertices $\{e_j, j \in \mathcal{A}_i\}$ with probability one for every $i \in \mathcal{I}$. $\square$

3.7 Games with aligned interests for $\tilde{L}_{R-I}$

When n automata are involved in a nonidentical payoff game Γ, and all of them apply the $\tilde{L}_{R-I}$ reinforcement scheme, then it is not necessarily true that the process $\{R_i(\alpha(k))\}_k$ is a submartingale for all $i \in \mathcal{I}$. That is because $[\Delta R_i(k)]_1$ is not necessarily nonnegative. In particular, for nonidentical payoff games

$$[\Delta R_i]_1 = \sum_{s=1}^{n} \sum_{j=1}^{|\mathcal{A}_s|} \Delta x_{sj} E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x]. \tag{3.13}$$

We can rewrite the above expression as

$$\begin{aligned}[\Delta R_i]_1 = \Delta x_i^{\mathrm{T}} \cdot [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_i = j, x]]_{j \in \mathcal{A}_i} + \\ \sum_{s=1, s \neq i}^{n} \Delta x_s^{\mathrm{T}} \cdot [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x]]_{j \in \mathcal{A}_s}.\end{aligned} \tag{3.14}$$

In case agent i updates its strategy with the $\tilde{L}_{R-I}$ reinforcement scheme, the first term of the r.h.s. can be written as

$$\Delta x_i^{\mathrm{T}} \cdot [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_i = j, x]]_{j \in \mathcal{A}_i} = \epsilon x_i^{\mathrm{T}} \tilde{D}_i x_i / 2 \geq 0$$

where $\tilde{D}_i \geq 0$ was defined in the proof of Proposition 3.6.4. Note that in games with identical interests the payoff function R_i is the same for every agent, therefore the second term of the r.h.s. of (3.14) is also of a quadratic form. However, in the case of nonidentical payoffs this is not necessarily the case.

Because all automata apply the $\tilde{L}_{R-I}$ scheme, we know that Δx_s, $s \neq i$, will be positive in the directions of increase in the expected payoff of agent s. Therefore, if the vector

$$[E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x]]_{j \in \mathcal{A}_s}$$

takes its maximum value in one of the directions where Δx_s is positive, then

$$\Delta x_s^{\mathrm{T}} \cdot [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n)|\alpha_s = j, x]]_{j \in \mathcal{A}_s} \geq 0, \quad s \neq i. \tag{3.15}$$

If this is true for every agent $s \neq i$, then we also have that $[\Delta R]_1 \geq 0$. In this case, the change in strategy of any agent s that applies the $\tilde{L}_{R-I}$ scheme, it does not make agent i worse off. Note that this situation corresponds to a subset of the set of games with *aligned interests*.[4]

Remark 3.7.1 *Propositions 3.6.3, 3.6.4 and 3.6.5 proved for identical interest games continue to hold for nonidentical interest games where ΔR_i is computed separately for each agent $i \in \mathcal{I}$.*

Based on this observation, we can characterize the stability properties of $\tilde{L}_{R-I}$ scheme when applied by 2 automata in the class of games that satisfy condition (3.15).

Theorem 3.7.1 (Convergence of 2×2 coordination games) *Let two automata play a game Γ with $|\mathcal{A}_1| = |\mathcal{A}_2| = 2$ actions and two pure strict Nash equilibria. Let both automata apply the $\tilde{L}_{R-I}$ scheme with step size $\epsilon(k) > 0$ such that $\sum_k \epsilon(k) = \infty$, and assume that condition (3.15) is also satisfied. Then, the process $\{x_i(k)\}_k$ converges to the set of vertices $\{e_j : j \in \mathcal{A}_i\}$ with probability one for every $i \in \mathcal{I}$.*

Proof. The proof follows the steps of the proof of Theorem 3.6.1. □

3.8 Perturbed learning automata

Here we consider a perturbation of the $\tilde{L}_{R-I}$ scheme, where the decisions of each agent are slightly perturbed. In particular, we consider that each agent i selects

[4] As defined by Definition 2.2.4.

action $\alpha_i \in \{1, ..., |\mathcal{A}_i|\}$ according to the perturbed policy

$$x_i^{\lambda}(\alpha_i) \triangleq (1-\lambda)x_i(\alpha_i) + \lambda/\left|\mathcal{A}_i\right|.$$

for some $\lambda \geq 0$, which is called *mutation rate*. We will denote this scheme by $\tilde{L}_{R-I}^{\lambda}$.[5]

3.8.1 Convergence analysis for constant step size

We will focus our analysis in two-player and two-action nonidentical payoff games where condition (3.15) is satisfied and each agent i has payoff matrix D_i. In order to characterize the asymptotic behavior of the stochastic process $\{x(k)\}_k$, we will use the nonnegative function

$$R_i(\alpha(k)) = \alpha_i(k)^{\mathrm{T}} D_i \alpha_{-i}(k).$$

By Claim 3.6.1, the conditional expected change in agent 1's payoff is

$$\Delta R_i = \Delta {x_i}^{\mathrm{T}} D_i x_{-i} + x_i^{\mathrm{T}} D_i \Delta x_{-i} + E[\delta {x_i}^{\mathrm{T}} D_i \delta x_{-i} | x].$$

In order to characterize the asymptotic properties of the game of learning automata, we will compute the sign of ΔR_i. When condition (3.15) is satisfied, we already know that if $\Delta {x_i}^{\mathrm{T}} D_i x_{-i} \geq 0$, then $x_i^{\mathrm{T}} D_i \Delta x_{-i} \geq 0$. Therefore, it suffices to compute the sign of the terms $\Delta {x_i}^{\mathrm{T}} D_i x_{-i}$ and $E[\delta {x_i}^{\mathrm{T}} D_i \delta x_{-i}]$.

Define $\overline{v}_i = D_i x_{-i}$ and $\tilde{D}_i \in \mathbb{R}^{2\times 2}$ such that $\tilde{D}_i(s,s) = 0$ and $\tilde{D}_i(s,j) = (\overline{v}_{is} - \overline{v}_{ij})^2$, where $\overline{v}_{is}$ is the sth entry of the vector $\overline{v}_i$. Let us also define the matrix $X_i \in \mathbb{R}^{2\times 2}$ such that $X_i(s,s) = 1 - x_{is}$ and $X_i(s,j) = -x_{is}$.

[5]Note that for $\lambda = 0$, the learning scheme $\tilde{L}_{R-I}^{0}$ coincides with $\tilde{L}_{R-I}$.

Claim 3.8.1 *When agent $i \in \{1,2\}$ applies the $\tilde{L}^{\lambda}_{R-I}$ scheme for some $\lambda > 0$, then*

$$\Delta {x_i}^{\mathrm{T}} D_i x_{-i} = \frac{1-\lambda}{2} x_i^{\mathrm{T}} \tilde{D}_i x_i + \frac{\lambda}{|\mathcal{A}_i|} \overline{v}_i^{\mathrm{T}} X_i \overline{v}_i. \tag{3.16}$$

Furthermore, there exists $M_1 \geq 0$ such that $\Delta {x_i}^{\mathrm{T}} D_i x_{-i} \geq -\lambda M_1$ for all $x \in \mathcal{X}$.

Proof. We have:

$$\begin{aligned}
\Delta {x_i}^{\mathrm{T}} D_i x_{-i} &= \Delta {x_i}^{\mathrm{T}} \overline{v}_i \\
&= (1-\lambda) \sum_{j \in \mathcal{A}_i} \left[\sum_{s \neq j} \overline{v}_{ij} x_{is} x_{ij} - \overline{v}_{is} x_{ij} x_{is} \right] \overline{v}_{ij} + \\
&\quad \frac{\lambda}{|\mathcal{A}_i|} \sum_{j \in \mathcal{A}_i} \left[\sum_{s \neq j} \overline{v}_{ij} x_{is} - \overline{v}_{is} x_{ij} \right] \overline{v}_{ij} \\
&= \frac{1-\lambda}{2} x_i^{\mathrm{T}} \tilde{D}_i x_i + \frac{\lambda}{|\mathcal{A}_i|} \overline{v}_i^{\mathrm{T}} X_i \overline{v}_i.
\end{aligned}$$

The first part of the r.h.s. has already been shown to be for the unperturbed case. The second part can be easily verified. Note also that since the first term of the r.h.s. is ≥ 0 for any x, and the second term is absolutely bounded, the conclusion follows. □

Note that $\tilde{D}_i$ is a positive semi-definite matrix, and therefore the first term of the r.h.s. of (3.16) will be a nonnegative quantity. The sign of the second term of the r.h.s. depends on x_i.

Claim 3.8.2 *When agent $i \in \{1,2\}$ applies the $\tilde{L}^{\lambda}_{R-I}$ scheme for some $\lambda > 0$, then*

$$E[\delta {x_i}^{\mathrm{T}} D^i \delta x_{-i}] =$$
$$\epsilon^2 x_{11} x_{12} x_{21} x_{22} (d_{11} - d_{12} - d_{21} + d_{22})((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2) + \lambda \psi(\lambda, x).$$

where $\psi : \mathcal{X} \to \mathbb{R}$ such that: (a) there exists $M > 0$ such that $|\psi(\lambda, x)| \leq M$, and (b) $\lim_{\lambda \to 0} \lambda \psi(\lambda, x) = 0$.

Proof. Note that

$$\begin{aligned}
&E[\delta {x_i}^{\mathrm{T}} D^i \delta x_{-i}] \\
&= \epsilon^2(d_{11}-d_{12}-d_{21}+d_{22})[x_{12}x_{22}x_{11}^{\lambda}x_{21}^{\lambda}(d_{11})^2 - x_{12}x_{21}x_{11}^{\lambda}x_{22}^{\lambda}(d_{12})^2 - \\
&\quad x_{11}x_{22}x_{12}^{\lambda}x_{21}^{\lambda}(d_{21})^2 + x_{11}x_{21}x_{12}^{\lambda}x_{22}^{\lambda}(d_{22})^2] \\
&\approx \epsilon^2 x_{11}x_{12}x_{21}x_{22}(d_{11}-d_{12}-d_{21}+d_{22})((d_{11})^2-(d_{12})^2-(d_{21})^2+(d_{22})^2) \\
&\quad +\frac{\lambda}{2}\epsilon^2(d_{11}-d_{12}-d_{21}+d_{22})[x_{12}x_{22}x_{11}(d_{11})^2 + x_{12}x_{22}x_{21}(d_{11})^2 \\
&\quad -x_{12}x_{21}x_{11}(d_{12})^2 - x_{12}x_{21}x_{22}(d_{12})^2 - x_{11}x_{22}x_{12}(d_{21})^2 - \\
&\quad x_{11}x_{22}x_{21}(d_{21})^2 + x_{11}x_{21}x_{12}(d_{22})^2 + x_{11}x_{21}x_{22}(d_{22})^2]
\end{aligned}$$

plus higher order terms of λ. Therefore, the conclusion follows. □

Let two automata apply the perturbed reinforcement scheme $\tilde{L}_{R-I}^{\lambda}$. Let x^* correspond to a vertex of the state space $\mathcal{X}$. Let us define the set $\mathcal{V}$ as the set of vertices of the domain. Let us also define an ε-neighborhood of this set, denoted by $\mathcal{B}_\varepsilon(\mathcal{V}) \subset \mathcal{X}$ for some $\varepsilon > 0$, by

$$\mathcal{B}_\varepsilon(\mathcal{V}) = \{x \in \mathcal{X} : \mathrm{dist}(x, \mathcal{V}) \leq \varepsilon\}$$

where $\mathrm{dist}(x, \mathcal{V}) \triangleq \inf_{y \in \mathcal{V}} |x - y|$. Finally, define

$$\mathcal{D}_\varepsilon(\mathcal{V}) \triangleq \mathcal{X} \backslash \mathcal{B}_\varepsilon(\mathcal{V}).$$

Proposition 3.8.1 *Let two agents play a game Γ with $|\mathcal{A}_1| = |\mathcal{A}_2| = 2$ actions and two pure strict Nash equilibria. Assume that each agent apply the perturbed reinforcement scheme $\tilde{L}_{R-I}^{\lambda}$ with some constant step size $\epsilon > 0$, and let condition (3.15) hold. For any $\varepsilon > 0$, there exists $\lambda_0 = \lambda_0(\varepsilon)$ such that the set $\mathcal{B}_\varepsilon(\mathcal{V})$ is recurrent for $\{x(k)\}_k$ for all $\lambda < \lambda_0$ in that $x(k) \in \mathcal{B}_\varepsilon(\mathcal{V})$ for infinitely many k w.p.1.*

Proof. Let us define the functions $V_i : \Delta(|\mathcal{A}_i|) \to \mathbb{R}_+$ such that $V_i(\alpha(k)) = R_{i,\max} -$

$R_i(\alpha(k))$ which is nonnegative. Let us also define

$$V(\alpha(k)) = \sum_{i \in \mathcal{I}} V_i(\alpha(k)), \tag{3.17}$$

which is also nonnegative. Then

$$\begin{aligned} \Delta V(x(k)) &\triangleq E[V(\alpha(k+1)) - V(\alpha(k))|x(k)] \\ &= -\sum_{i \in \mathcal{I}} [R_i(\alpha(k+1)) - R_i(\alpha(k))|x(k)] \\ &= -\sum_{i \in \mathcal{I}} \Delta R_i(x(k)). \end{aligned}$$

Note that for any two-player and two-action game with two pure strict Nash equilibria we have

$$(d_{11} - d_{12} - d_{21} + d_{22})((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2) = K$$

for some $K > 0$. According to Claim 3.8.2, for some $\lambda > 0$ and $\varepsilon > 0$,

$$E[\delta {x_i}^{\mathrm{T}} D^i \delta x_{-i}] \geq \epsilon^2 \varepsilon^2 (1-\varepsilon)^2 K + \lambda \psi(x)$$

for all $x \in \mathcal{D}_\varepsilon(\mathcal{V})$.

According to Claim 3.8.1 and the fact that condition (3.15) is satisfied, there are constants $M_1 > 0$ and $M_2 > 0$ such that $\Delta {x_i}^{\mathrm{T}} D^i x_{-i} \geq -\lambda M_1$ and $x_i^{\mathrm{T}} D^i \Delta x_{-i} \geq -\lambda M_2$ for all $x \in \mathcal{D}_\varepsilon(\mathcal{V})$.

Therefore, for all $x \in \mathcal{D}_\varepsilon(\mathcal{V})$, we have

$$\Delta R(x(k)) \geq \epsilon^2 \varepsilon^2 (1-\varepsilon)^2 K + \lambda \psi(x) - \lambda(M_1 + M_2).$$

We conclude that for any given step size $\epsilon > 0$ and any $\varepsilon > 0$, there exists λ_0, such that $\Delta R_i(x(k)) > 0$ for all $\lambda < \lambda_0$, $i \in \mathcal{I}$ and $x \in \mathcal{D}_\varepsilon(\mathcal{V})$. Since this holds for

both agents, we conclude that $\Delta V(x(k)) < 0$ for all $\lambda < \lambda_0$ and $x \in \mathcal{D}_\varepsilon(\mathcal{V})$. Thus, by Corollary B.1.1, we conclude that for any $\varepsilon > 0$, the set $\mathcal{B}_\varepsilon(\mathcal{V})$ is recurrent for $\{x(k)\}_k$. $\square$

The above proposition shows that the a small neighborhood of the vertices of the probability simplex is a recurrent set. However, such an approach is not able to show where exactly the process will converge if it converges. It turns out that there cannot be convergence w.p.1. Instead the asymptotic properties can be characterized in a distributional sense.

3.8.2 The ODE approach

In order to give a more specific characterization of the asymptotic convergence of $\{x(k)\}_k$ within the set $\mathcal{B}_\varepsilon(\mathcal{V})$, which is recurrent for the induced Markov process, we need to use methods that rely on an ODE approach. According to these methods, the behavior of the process as the step size ϵ approaches zero can be described by the solution of a collection of ordinary differential equations (ODE's), which represents the *mean dynamics*. In fact, for our reinforcement scheme, the collection of ODE's will be

$$\frac{dx_i(t)}{dt} = E[R_i(\alpha) \cdot (\alpha_i - x_i)|x(t)] \triangleq \overline{g}(x), \quad i \in \mathcal{I}. \tag{3.18}$$

It can be shown (applying Theorem 8.2.1 in [KY97]) that the smaller the step size, the larger the amount of time that the stochastic process spends in a set of chain recurrent points of the ODE (3.18). Essentially, the convergence here is in distribution and the proof is based on weak convergence techniques.

Hence, in order to characterize the convergence of the stochastic process with constant step size, it is essential to compute the set of chain recurrent points of the ODE (3.18). In the case of two-player and two-action coordination games, and based

on Proposition 3.8.1, it suffices to compute the chain recurrent points within $\mathcal{B}_\varepsilon(\mathcal{V})$.

A similar approach is followed when the step size is diminishing satisfying the general form

$$\epsilon(k) = \frac{1}{ck^\nu + 1} \tag{3.19}$$

when $\nu \in (1/2, 1]$. In this case, the process converges to a set of chain recurrent points of the ODE (3.18) as Proposition C.1.1 shows. However, when the underlying game is a two-player and two-action coordination game, $\mathcal{B}_\varepsilon(\mathcal{V})$ is not necessarily the unique invariant set of the ODE (3.18). Recall that in the two-player and two-action coordination games, the stochastic process with $\lambda = 0$ has isolated stationary points other than the vertices of the probability simplex. Convergence to such a point can be excluded when $\lambda = 0$, based on the martingale convergence theorem, as has been already shown. However, when $\lambda > 0$, computing the exact sign of ΔR_i in the vicinity of that point is a quite complicated procedure. Other functions may be more appropriate, however it is still part of an ongoing work.

3.8.3 Convergence analysis for diminishing step size

Consider a step size sequence as defined by (3.19). In that case, we wish to give some more specific convergence properties based on the ODE methods for stochastic approximations. First, we observe the following.

Proposition 3.8.2 *Let two automata play a game Γ with $|\mathcal{A}_1| = |\mathcal{A}_2| = 2$ actions and two pure strict Nash equilibria. Let both automata apply the $\tilde{L}_{R-I}$ scheme with step size $\epsilon(k) > 0$ such that $\sum_k \epsilon(k) = \infty$, and assume that condition (3.15) is also satisfied. For any $\varepsilon > 0$ and for sufficiently small $\lambda > 0$, the set $\mathcal{B}_\varepsilon(\mathcal{V})$ is invariant for the ODE (3.18).*

Proof. Let us define the nonnegative function

$$V(x) = \sum_{i \in \mathcal{I}} R_{i,\max} - \overline{R}_i(x),$$

where $R_{i,\max}$ is the upper bound of the payoff function of agent i. Note that

$$\begin{aligned}
\frac{dV(x)}{dt} &= -\sum_{i \in \mathcal{I}} \frac{d\overline{R}_i(x)}{dt} \\
&= -\sum_{i \in \mathcal{I}} [\nabla_x \overline{R}_i(x)]^{\mathrm{T}} \cdot \frac{dx(t)}{dt} \\
&= -\sum_{i \in \mathcal{I}} \sum_{s \in \mathcal{I}} [\nabla_{x_s} \overline{R}_i(x)]^{\mathrm{T}} \cdot \frac{dx_s(t)}{dt} \\
&= -\sum_{i \in \mathcal{I}} \sum_{s \in \mathcal{I}} [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x(t)]]_{j \in \mathcal{A}_s}{}^{\mathrm{T}} \cdot \frac{dx_s(t)}{dt}
\end{aligned}$$

where we use the fact that

$$\nabla_{x_s} \overline{R}_i(x) = [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x]]_{j \in \mathcal{A}_s}.$$

Also, by definition of the ODE (3.18), we have that

$$\frac{dx_s(t)}{dt} \equiv E[R_s(\alpha) \cdot (\alpha_s - x_s) | x], \quad s \in \mathcal{I}.$$

According to the definition of the first-order variational terms $[\Delta R_i(k)]_1$ in equation (3.13), we have

$$\begin{aligned}
&[\Delta R_i(k)]_1 \\
&= \sum_{s \in \mathcal{I}} \sum_{j \in \mathcal{A}_s} E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x(k)] \Delta x_{sj}(k) \\
&= \epsilon(k) \sum_{s \in \mathcal{I}} [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x(k)]]_{j \in \mathcal{A}_s} \cdot E[R_s(\alpha) \cdot (\alpha_s - x_s) | x(k)].
\end{aligned}$$

By condition (3.15) and according to Claim 3.8.1, we see that for any given $\varepsilon > 0$ and for sufficiently small $\lambda > 0$, $[\Delta R_i(k)]_1 \geq 0$ for all x on the boundary of $\mathcal{B}_\varepsilon(\mathcal{V})$. Since $\epsilon(k) > 0$ for $k = 0, 1, 2, ...$, we also have that

$$\sum_{s \in \mathcal{I}} [E[R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_s = j, x]]_{j \in \mathcal{A}_s} \cdot E[R_s(\alpha) \cdot (\alpha_s - x_s) | x] \geq 0$$

for all x on the boundary of $\mathcal{B}_\varepsilon(\mathcal{V})$. Thus,

$$\frac{dV(x)}{dt} \leq 0,$$

for all x on the boundary of $\mathcal{B}_\varepsilon(\mathcal{V})$, which concludes the proof. □

Note that the ODE (3.18) might have other invariant sets besides $\mathcal{B}_\varepsilon(\mathcal{V})$.

We can characterize the convergence properties of the learning automata game based on the ODE method for stochastic approximations as described in Appendix C. We first need to assume the following:

Assumption 3.8.1 *The function $\overline{g}(x)$ is continuously differentiable on $\mathcal{X}$.*

This assumption is not restrictive. For example, in the coordination problems which will be considered, the expected reward function will be continuously differentiable with respect to the strategy x.

Proposition 3.8.3 (Diminishing step size: Convergence) *Under Assumption 3.8.1, for bounded reward function and $\lambda > 0$, the sequence $\{x(k)\}$ converges to an invariant set of the ODE (3.18). Furthermore, let $A \subset \mathcal{X}$ be a locally asymptotically stable set in the sense of Lyapunov for (3.18). Then $P[\lim_{k \to \infty} x(k) \in A] > 0$.*

Proof. Note that the stochastic process satisfies Assumptions C.1.1, C.1.2, C.1.3 and C.1.4. Therefore, from Proposition C.1.1 follows that the stochastic process will

converge to an invariant set of the ODE (3.18). Furthermore, according to the same proposition, if $x(k)$ is in some compact set in the domain of attraction of a locally asymptotically stable set A infinitely often with probability $\geq \rho$, then $x(k) \to A$ with at least probability ρ. However, because of the randomization in the action selection that is induced by the mutation rate $\lambda > 0$, there is a positive probability that the process is in some compact set in the domain of attraction of any locally asymptotically stable set A of the vector field $\overline{g}(x)$. □

Convergence to linearly unstable points of the ODE (3.18) can be excluded as stated by the following proposition:

Proposition 3.8.4 (Diminishing step size: Nonconvergence) *Let x^* be any point in $\mathcal{X}$ such that x^* is a linearly unstable equilibrium point of the ODE (3.18). Then, $P[x(k) \to x^*] = 0$.*

Proof. The proposition is a direct consequence of Proposition C.2.1. □

3.9 Remarks

In this chapter, we presented the basic convergence properties of the reinforcement learning scheme that we will use in the remainder of the dissertation. The reinforcement scheme is a small modification of the reward-inaction scheme of learning automata. We distinguished among two different forms of the reinforcement scheme: the unperturbed and the perturbed scheme. The main reason we introduced the perturbed learning algorithm is the fact that it allows for equilibrium selection, as it will become obvious in the following chapter.

For the unperturbed scheme, it was shown that the learning algorithm (with either constant or diminishing step size sequence) converges to the set of vertices of

the probability simplex w.p.1 under certain conditions on the reward function. We showed that a special class of two-player and two-action coordination games always satisfy these conditions. For the perturbed scheme with constant step size, and for the same class of two-player and two-action coordination games, it was shown that any small neighborhood of the set of vertices is a recurrent set of the induced Markov process. If the step size sequence diminishes to zero, then this set is an invariant set of the corresponding mean dynamics. Further conclusions about local convergence properties within this invariant set were also derived based on existing results in stochastic approximations.

CHAPTER 4

Distributed Dynamic Reinforcement of Efficient Outcomes in Multiagent Coordination

4.1 Introduction

In this chapter, we analyze the asymptotic behavior of a class of the perturbed learning dynamics introduced in Section 3.8 under "*dynamic reinforcement*." Unlike traditional reinforcement learning, agents using dynamic reinforcement use a combination of long term rewards and recent rewards to construct myopically forward looking action selection probabilities.

We analyze the long term stability of the learning dynamics for general games with pure strategy Nash equilibria and specialize the results for coordination games and distributed network formation. Prior work [SA05] has shown how such dynamic reinforcement can enable convergence to mixed strategy equilibria. Unlike those results, the focus here is how dynamic reinforcement can influence equilibrium selection, particularly in coordination games. In this class of problems, more than one stable equilibrium (i.e., coordination configuration) can exist. We show that dynamic reinforcement can be used as an equilibrium selection scheme. Moreover, only a *single* agent is able to destabilize an equilibrium in favor of another by appropriately adjusting its dynamic reinforcement parameters. This sort of single agent sensitivity also has implications for agent based simulations.

We compare the equilibrium selection properties of the dynamic reinforcement

algorithm with existing results for coordination games. Prior work by [You93] and [KMR93] on coordination games with 2 agents and 2 actions when best-reply dynamics are applied has shown that the risk-dominant equilibrium[1] is the only robust equilibrium when agents' decisions are subject to small mistakes (*mutations*). That is, the risk-dominant equilibrium is the long-run prediction of the perturbed process when the mutation rate approaches zero. The resulting equilibrium selection under the present dynamic reinforcement need not be determined by either payoff or risk dominance or both. A related result [BL96] has shown that it is possible to find small mutation rates so that *any* long-run prediction is possible. However, these specifically tailored perturbations are state dependent and uniformly distributed across population. By contrast, dynamic reinforcement affects equilibrium selection without modifying the information available to each agent.

We also illustrate the results in distributed network formation with three nodes, where each node establishes recursively links with other nodes. For the case of three agents, we illustrate how a network formation game can be designed so that certain desirable configurations are efficient, while non-efficient equilibria can be destabilized by the proposed dynamic reinforcement. The more general case (of more that three agents) will be analyzed in Chapter 5.

4.2 Motivation

4.2.1 Coordination games

As discussed in Section 1.1.1, several coordination problems can be analyzed by coordination games, a branch of game theory problems that has been studied by [Sch06]. Recall that we defined coordination games as coordination problems were agents' interest is to achieve uniformity of actions by each doing whatever the others will

[1] See Section 2.2.3.

do. In coordination games uniform combinations of actions are *Nash equilibria*, i.e., situations in which no agent can be better off had it decided to act otherwise alone.

There are more restrictive classes of coordination problems, depending on the level of coincidence of interest among agents. For example, in the game of Table 4.1(a), there is perfect coincidence of interest (*pure coordination*) [Sch06], while in the game of Table 4.1(b) there is no perfect coincidence of interest. However, the latter game belongs to the class of games with *aligned interests* as defined in Section 3.7.

	$2.A$	$2.B$
$1.A$	$5,5$	$1,1$
$1.B$	$1,1$	$2,2$

(a)

	$2.A$	$2.B$
$1.A$	$5,5$	$1,3$
$1.B$	$3,1$	$3,3$

(b)

Table 4.1: (a) The Typewriter game, (b) The Stag-Hunt Game

Both coordination games of Table 4.1 has drawn a lot of attention in social sciences since several social phenomena can be modeled by them. The Typewriter game can model the adoption of new technologies, where it is of common interest to use a compatible technology even though there might be a better one. The Stag-Hunt game is a more general game which can model the adoption or modification of the social contract for mutual benefit [Sky02]. Intuitively, in this setting each agent has two options, (A) to devote energy to instituting the new social contract, or (B) not. If everyone takes the first action, the *social contract equilibrium* is achieved (A, A), and if everyone takes the second action, the *state of nature equilibrium* results (B, B). But the second course carries no *risk*, while the first does.

An important question in social sciences is how we can get from the risk-dominant equilibrium (B, B) to the payoff-dominant equilibrium (A, A).[2] Although games are usually analyzed in either a *static* setting (where the game is played only once) or in a *dynamic* setting, the first approach has little to say about how agent's beliefs about

[2]See definitions in Section 2.2.3.

what others will do change. As we discussed in Section 2.3, most of the current work in learning dynamics show that risk-dominant equilibrium is the most reasonable choice.

4.2.2 Distributed network formation

Although coordination problems have drawn a lot of attention in social sciences, they are also present in several practical settings. For example, in sensor networks literature, one of the greatest challenges is to design protocols that guarantee an *energy efficient* configuration, since the major part of energy is consumed in transmitting signals. The question that arises is: *how efficient networks can be reinforced in a distributed and adaptive fashion?*

The problem of network formation can be modeled as a strategic interaction among nodes. Here nodes can be thought of as decision makers who have discretion over establishing omnidirectional links with other nodes. In a general setting, we may assume that nodes are sources of benefits (or information) that can be tapped through directed or indirected links, while the establishment of a link is costly. Agents may establish as many links as they want, however, they would prefer (due to their *rational* nature) to have access to as many agents as possible but with the minimum number of established links.

Such a strategic interaction setup of the network formation problem corresponds to the *connections model* of [JW96] and has been used to describe several economic and social contexts such as the transmission of information. Such a model exhibits several Nash equilibria. For example, in case of three nodes, and assuming that the benefit and cost associated to each link is constant among agents, there are two Nash equilibria which are shown in Figure 4.1. Under these conditions, it will be preferable to be able to guarantee convergence to network (a), since every node has access to the benefits of every other node with the minimum possible number of links.

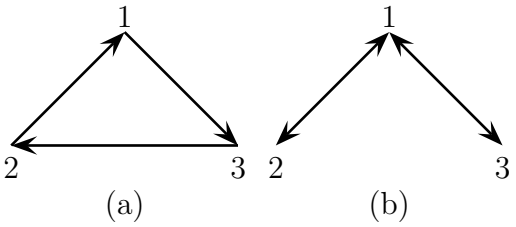


Figure 4.1: Nash equilibria in case of the *connections model* of [JW96].

4.3 The reinforcement learning algorithm

We will model an agent i in a set of agents $\mathcal{I} \triangleq \{1, 2, \ldots, n\}$ as a learning automaton which can take actions in the set $\mathcal{A}_i \triangleq \{1, 2, \ldots, |\mathcal{A}_i|\}$. At each time $k \in \{0, 1, 2, \ldots\}$, agent i interacts with the unknown environment (i.e., other agents) by selecting one action, $\alpha_i(k) \in \mathcal{A}_i$, and receives a reward, $R_i(\alpha(k))$, which depends on the actions of all agents, $\alpha(k) = (\alpha_1(k), ..., \alpha_n(k))$.

We assume that each agent i "learns" via the perturbed learning algorithm $\tilde{L}^{\lambda}_{R-I}$, introduced in Section 3.8. This algorithm is written recursively as

$$x_i(k+1) = x_i(k) + \epsilon(k) \cdot R_i(\alpha(k)) \cdot [\alpha_i(k) - x_i(k)], \tag{4.1}$$

where $x_i(k)$ is the *strategy* of agent i at time k. A strategy is a vector of probabilities $[x_{ij}(k)]_{j \in \mathcal{A}_i}$ that agent i assigns to his selecting actions 1 through $|\mathcal{A}_i|$. Accordingly, $x_i(k)$ belongs to the probability simplex $\Delta(|\mathcal{A}_i|)$.

Agent i chooses action j at time k with probability

$$(1 - \lambda)x_{ij}(k) + \lambda / |\mathcal{A}_i| ,$$

where $\lambda \geq 0$ models possible perturbations in the decision making process, also called *mutations* [KMR93, You93]. We will associate actions $\{1, 2, ..., |\mathcal{A}_i|\}$ with vertices of the simplex, $\left\{e_1, ..., e_{|\mathcal{A}_i|}\right\}$. If agent i chose action j at time k, then $\alpha_i(k) = e_j$ in

equation (4.1).

We assume that the step size sequence satisfies

$$\epsilon(k) \triangleq \frac{1}{k+1}. \tag{4.2}$$

Such model is similar to the so-called "*anticipated utility*" model: "each period the agent makes decisions based on his beliefs, treating his beliefs as constant, and then updates the beliefs upon observing outcomes."

In all applications considered here, rewards are strictly positive and bounded. Even if rewards are nonpositive, they can be always normalized to the positive axis. Therefore, we assume:

Assumption 4.3.1 (Strictly positive rewards) *For every $i \in \mathcal{I}$, the reward function $R_i(\cdot)$ satisfies $0 < R_i(\alpha(k)) < R_{i,\max}$ for any action profile $\alpha(k)$ and some $R_{i,\max} > 0$.*

4.4 Analysis

The asymptotic convergence properties of the perturbed reinforcement scheme $\tilde{L}^{\lambda}_{R-I}$ with diminishing step size was described in Section 3.8.3. In this framework, we showed by Proposition 3.8.3 that the reinforcement scheme converges to an invariant set of the set of ordinary differential equations:

$$\dot{x}_i = \overline{g}_i(x) \triangleq \overline{r}_i(x) - \overline{R}_i(x) \cdot x_i,$$

where

$$\overline{r}_i(x) \triangleq E[R_i(\alpha(k))\alpha_i(k)|x(k) = x] \in \mathbb{R}_+^{|\mathcal{A}_i|},$$

$$\overline{R}_i(x) \triangleq E[R_i(\alpha(k))|x(k) = x] \in \mathbb{R}_+.$$

The above set of ODE's can be written more compactly as

$$\dot{x} = \overline{g}(x) \triangleq \text{col}\{\overline{g}_i(x)\}_{i \in \mathcal{I}}. \tag{4.3}$$

where col$\{A\}$ denote the column vector of the elements of a finite set A.

Moreover, again according to Proposition 3.8.3, there is a positive probability that the reinforcement scheme converges to a locally stable set (in the sense of Lyapunov) of the ODE (4.3). It was also shown by Proposition 3.8.4 that there is probability zero that the reinforcement scheme will converge to a linearly unstable point of the ODE (4.3).

In this section, we are going to analyze the local stability properties of the stationary points of the ODE (4.3), since stationary points are invariant sets. This way, we can derive conclusions regarding convergence in several classes of coordination problems.

4.4.1 Characterization of the stationary points

Let $x^* \in \mathcal{S}$ be a candidate stationary point of the ODE (4.3). In order to characterize the stationary points, we define $\overline{v}_i(j, x^*)$ as the expected reward of agent i given that it selects action $\alpha_i = j \in \mathcal{A}_i$ and every other agent follows $x_{-i} = x^*_{-i}$, where $-i$ denotes the complementary set $\mathcal{I} \backslash i$, i.e.,

$$\overline{v}_i(j, x^*) \triangleq E\{R_i(\alpha_1, \alpha_2, ..., \alpha_n) | \alpha_i = j, x_{-i} = x^*_{-i}\}. \tag{4.4}$$

The stationary points are characterized by the following proposition.

Proposition 4.4.1 (Stationary points) *A strategy profile $x^* = (x^*_1, \ldots, x^*_n)$ is a stationary point of the ODE (4.3) if and only if, for every agent $i \in \mathcal{I}$, there exists a constant $c_i > 0$, such that for any action $j \in \mathcal{A}_i$, $x^*_{ij} > 0$ implies $\overline{v}_i(j, x^*) = c_i$.*

Proof. For some agent $i \in \mathcal{I}$, let x_i^* be a (possibly mixed) strategy, such that $x_{ij}^* > 0$ for some $j \in \mathcal{A}_i$. The strategy profile x^* will be a stationary point of the ODE (4.3) if and only if for any agent $i \in \mathcal{I}$ the identity

$$E[R_i(\alpha)(\alpha_i - x_i)|x_i = x_i^*] = 0$$

holds, which, componentwise, can be written as,

$$\overline{v}_i(q, x^*)x_{iq}^* - \sum_{j \in \mathcal{A}_i} \overline{v}_i(j, x^*)x_{ij}^* x_{iq}^* = 0 \quad \text{for all } q \in \mathcal{A}_i$$

or, equivalently, for all $q \in \mathcal{A}_i$ such that $x_{iq} > 0$,

$$\overline{v}_i(q, x^*) - \sum_{j \in \mathcal{A}_i} \overline{v}_i(j, x^*)x_{iq}^* = 0 \quad \text{for all } q \in \mathcal{A}_i.$$

The above linear system of equations has a solution if and only if $\overline{v}_i(q, x^*) = c_i$ for some positive number c_i for all $q \in \mathcal{A}_i$. □

An immediate consequence of the above proposition is that for $\lambda = 0$, any pure strategy profile is a stationary point of the stochastic iteration.

Proposition 4.4.2 (Pure Strategies) *For $\lambda = 0$, any pure strategy profile $x^* = (x_1^*, x_2^*, ..., x_n^*)$, such that x_i^* is a vertex of the probability simplex for all $i \in \mathcal{I}$, is a stationary point of the ODE (4.3).*

Proof. According to Proposition 4.4.1 and for $\lambda = 0$, any strategy profile $x^* = (x_1^*, \cdots, x_n^*)$, such that x_i^* is a vertex of the probability simplex (pure strategy), is a stationary point of the ODE (4.3), since the support of a pure strategy is a single action. □

Vertices cease to be equilibria for $\lambda > 0$. The following proposition provides the sensitivity of pure strategy equilibria to small values of λ.

Proposition 4.4.3 (Sensitivity of pure strategies) *For any pure strategy profile x^*, and for sufficiently small $\lambda > 0$, there exists a unique continuously differentiable function $w^* : \mathbb{R}_+ \to \mathbb{R}^{|\mathcal{A}|}$, such that $\lim_{\lambda \to 0} \lambda w^*(\lambda) = 0$, and*

$$\tilde{x} = x^* + \lambda w^*(\lambda)$$

is a stationary point of the ODE (4.3).

Proof. Let $\overline{v}_i(j, \tilde{x})$ be the conditional reward (4.4) of agent i evaluated at a point $\tilde{x}$ when it selects action $\alpha_i = j \in \mathcal{A}_i$. For any agent $i \in \mathcal{I}$ and any action $s \in \mathcal{A}_i$, the corresponding entry of the vector field is

$$\overline{g}_{is}(\tilde{x}) = \overline{v}_i(s, \tilde{x})[(1-\lambda)\tilde{x}_{is} + \lambda/|\mathcal{A}_i|] - \sum_{q \in \mathcal{A}_i} \overline{v}_i(q, \tilde{x})[(1-\lambda)\tilde{x}_{iq} + \lambda/|\mathcal{A}_i|]\tilde{x}_{is}. \quad (4.5)$$

Consider any pure strategy profile x^*, and take $\tilde{x} = x^* + \nu$, for some $\nu \in \times_{i \in \mathcal{I}} \mathbb{R}^{|\mathcal{A}_i|}$. Substituting $\tilde{x}$ into (4.5), yields

$$\begin{aligned}\overline{g}_{is}(\nu, \lambda) = {} & \overline{v}_i(s, \tilde{x}) \left[(1-\lambda)(x^*_{is} + \nu_{is}) + \lambda/|\mathcal{A}_i|\right] \\ & - \sum_{q \in \mathcal{A}_i} \overline{v}_i(q, \tilde{x}) \left[(1-\lambda)(x^*_{iq} + \nu_{iq}) + \lambda/|\mathcal{A}_i|\right](x^*_{is} + \nu_{is}).\end{aligned}$$

Note that $\overline{g}_{is}(0,0) = 0$, since x^* is a stationary point of the unperturbed dynamics. Moreover, the partial derivatives of $\overline{g}_{is}$ evaluated at $(0,0)$ are:

$$\left.\frac{\partial \overline{g}_{is}(\nu, \lambda)}{\partial \nu_{is}}\right|_{(0,0)} = \overline{v}_i(s, \tilde{x})(1 - x^*_{is}) - \sum_{q \in \mathcal{A}_i} \overline{v}_i(q, \tilde{x}) x^*_{iq},$$

$$\left.\frac{\partial \overline{g}_{is}(\nu,\lambda)}{\partial \nu_{iq}}\right|_{(0,0)} = -\overline{v}_i(q,\tilde{x})x^*_{is}, \quad \text{for all } q \in \mathcal{A}_i \backslash s.$$

The Jacobian matrix with respect to ν has full rank if we exclude the trivial case that every action has the same expected reward. Then, by implicit function theorem, there exists a neighborhood D of $\lambda = 0$ and a unique differentiable function $\nu^* : D \to \mathbb{R}^{|\mathcal{A}|}$ such that $\nu^*(0) = 0$ and $\overline{g}(\nu^*(\lambda), \lambda) = 0$, for any $\lambda \in D$.

Also, since $\nu^*(\cdot)$ is continuously differentiable and $\nu^*(0) = 0$, we may write $\nu^*(\lambda) = \lambda w^*(\lambda)$, for some unique continuously differentiable function $w^*(\cdot)$ of λ in D. Moreover, for sufficiently small $\lambda > 0$, we can show that $\lambda w^*(\lambda) \approx \lambda w^*(0)$ and $\mathbf{1}^{\mathrm{T}} w^*_i(0) = 0$ for all $i \in \mathcal{I}$, which implies that $\tilde{x}_i \approx x^*_i + \lambda w^*_i(0) \in \Delta(|\mathcal{A}_i|)$. $\square$

In other words, Proposition 4.4.3 states that a sufficiently small perturbation λ in the decision process moves the stationary point away from the pure strategy by an order of λ.

4.4.2 Example: Stationary points in a symmetric game

Table 4.2 presents a symmetric game of two agents ($n = 2$) and two actions ($m = 2$). In this matrix, the rewards of agent 1 are given by the first entry of each block, and the rewards of agent 2 are given by the second entry.

	$2.A$	$2.B$
$1.A$	a, a	b, c
$1.B$	c, b	d, d

Table 4.2: The symmetric game

We are particularly interested in the case where $a > c > 0$ and $d > b > 0$. Then the game is a *coordination game*, since agents receive the maximum possible reward if and only if they coordinate. In particular, the action profiles (A, A) and (B, B) provide higher reward than (A, B) or (B, A). If $a > d$ (or $d > b$), then the action

profile (A, A) (or (B, B)) is the payoff-dominant equilibrium. Moreover, according to the definition of the risk-dominant equilibrium in Section 4.2.1, $a - c > d - b$ (or $a - c < d - b$) implies that (A, A) (or (B, B)) is the risk-dominant equilibrium.

According to Proposition 4.4.2 for $\lambda = 0$, any pure strategy profile is a stationary point of the ODE (4.3). The four pure strategy stationary points in the above coordination game are (A, A), (A, B), (B, A) and (B, B). Moreover, according to Proposition 4.4.1 there is also a mixed strategy stationary point.

4.4.3 Local asymptotic stability (LAS)

Having characterized the stationary points of the ODE (4.3), we will describe locally the stability properties of these points. We will first focus on the case $\lambda = 0$, which will give a clear picture about the local stability of the stationary points in the coordination games.

Proposition 4.4.4 (LAS - unperturbed system) *For $\lambda = 0$, let $x^* = (x_1^*, \ldots, x_n^*)$ be a stationary point of the ODE (4.3), such that for each $i \in \mathcal{I}$ there exists $j^* = j^*(i) \in \mathcal{A}_i$ for which $x_i^* = e_{j^*}$, i.e., x_i^* is a vertex of the probability simplex. Let $\overline{v}_i(\cdot, x^*)$ be the conditional reward (4.4) of agent i evaluated at x^*. Under Assumption 4.3.1, the stationary point x^* is a locally asymptotically stable point of the ODE (4.3) if, for each $i \in \mathcal{I}$, $\overline{v}_i(j^*, x^*) > \overline{v}_i(s, x^*)$ for all $s \in \mathcal{A}_i \backslash j^*$.*

Proof. Let $x^* = (x_1^*, \ldots, x_n^*)$ be a stationary point such that for each $i \in \mathcal{I}$ there exists $j^* = j^*(i) \in \mathcal{A}_i$. Define the Lyapunov function

$$V(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}}(x - x^*).$$

Differentiating $V(x(t))$ along solutions of (4.3) results in

$$\dot{V}(x) = (x - x^*)^{\mathrm{T}} \overline{g}(x) = \sum_{i \in \mathcal{I}} (x_i - x_i^*)^{\mathrm{T}} \overline{g}_i(x).$$

Note that in order to characterize the behavior of $\dot{V}(\cdot)$ about x^*, it suffices to analyze its behavior separately on each agent's strategy x_i, $i \in \mathcal{I}$, while $x_{-i} \approx x^*_{-i}$.

For some $\tilde{x}_i \in \Delta(m)$ and $\varepsilon > 0$, let

$$x_i = (1 - \varepsilon) e_{j^*} + \varepsilon \tilde{x}_i$$

be a perturbation of the agent i's strategy. Finally, let $\overline{v}_i(\cdot, x^*)$ be the conditional rewards vector (4.4) evaluated at x^*.

For sufficiently small ε,

$$\begin{aligned} \dot{V}(x) &= \sum_{i \in \mathcal{I}} \sum_{q \in \mathcal{A}_i} (x_{iq} - x^*_{iq})(\overline{r}_{iq}(x) - \overline{R}_i(x) x_i(q)) \\ &\approx \sum_{i \in \mathcal{I}} \sum_{q \in \mathcal{A}_i \setminus j^*} (\varepsilon \tilde{x}_{iq})^2 [\overline{v}_i(q, x^*) - \overline{v}_i(j^*, x^*)] \\ &\quad + \varepsilon^2 (1 - \tilde{x}_{ij^*}) \sum_{s \in \mathcal{A}_i} [\overline{v}_i(s, x^*) - \overline{v}_i(j^*, x^*)] \tilde{x}_{is} \end{aligned}$$

plus higher order terms in ε. Therefore, $\overline{v}_i(j^*, x^*) > \overline{v}_i(s, x^*)$ for all $s \neq j^*$ implies that $\dot{V} < 0$ for sufficiently small ε. $\square$

It turns out that the condition $\overline{v}_i(j^*, x^*) > \overline{v}_i(s, x^*), s \neq j^*$ implies that x^* is a strict Nash equilibrium[3]. Therefore, any strict Nash equilbrium is a locally asymptotically stable point of the ODE (4.3).

[3]See Definition 2.2.2.

If we let $\lambda > 0$, then the stationary points of the ODE (4.3) move slightly away from the pure strategy profiles as Proposition 4.4.3 indicates. In this case, we can assess stability as follows.

Proposition 4.4.5 (LAS - perturbed system) *For sufficiently small $\lambda > 0$, let $\tilde{x}$ be a stationary point of the ODE (4.3), where $\tilde{x}_i = x_i^* + \lambda w_i^*$ for all $i \in \mathcal{I}$, with $x_i^* = e_{j^*}$ for some $j^* = j^*(i) \in \mathcal{A}_i$ and w_i^* defined according to Proposition 4.4.3. Let $\overline{v}_i(\cdot, \tilde{x})$ be the conditional rewards vector (4.4) evaluated at $\tilde{x}$. Under Assumption 4.3.1, the stationary point $\tilde{x}$ is a locally asymptotically stable point of the ODE (4.3) if and only if, for each $i \in \mathcal{I}$, $\overline{v}_i(j^*, \tilde{x}) > \overline{v}_i(s, \tilde{x})$ for all $s \in \mathcal{A}_i \backslash j^*$.*

Proof. This theorem will follow directly as a special case of the forthcoming Theorem 4.5.1. □

Proposition 4.4.6 *In the framework of Proposition 4.4.5, $P[\lim_{k\to\infty} x(k) = \tilde{x}] > 0$ if, for each $i \in \mathcal{I}$, $\overline{v}_i(j^*, \tilde{x}) > \overline{v}_i(s, \tilde{x})$ for all $s \in \mathcal{A}_i \backslash j^*$, i.e., $\tilde{x}$ is a strict Nash equilibrium. Instead, if there exist $i \in \mathcal{I}$ and $s \in \mathcal{A}_i \backslash j^*$ such that $\overline{v}_i(j^*, \tilde{x}) < \overline{v}_i(s, \tilde{x})$, then $P[\lim_{k\to\infty} x(k) = \tilde{x}] = 0$.*

Proof. This is a direct consequence of Propositions 3.8.3, 3.8.4, and 4.4.5. □

For example, if we consider the symmetric game of Table 4.2 with $a > d > b > 0$ and $b = c$, then the game is a coordination game and there are two locally stable stationary points (A, A) and (B, B) which correspond to the two strict Nash equilibria of the game. For the perturbed ($\lambda > 0$) system, we will see that the equilibria associated with the vertices (A, B) and (B, A) are linearly unstable. This coordination game also has a mixed strategy stationary point which is unstable as well.

To illustrate these conclusions, consider the case $a = 4$, $b = c = 1$ and $d = 2$. Assume that both agents 1 and 2 update their strategies based on (4.1). The solution of the ODE (4.3) for some randomly selected initial condition in the domain of attraction of (B, B) is shown in Fig. 4.2 for $\lambda = 0.01$. In this figure, we can see that convergence to the locally stable stationary point (B, B) is established.

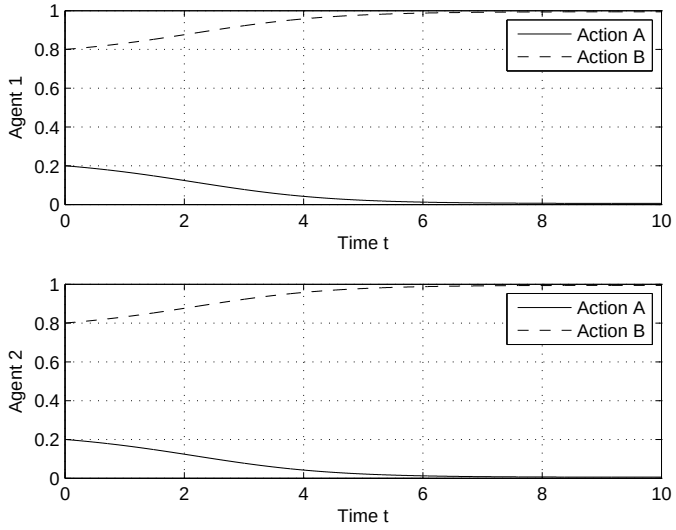


Figure 4.2: Solution of the ODE (4.3) for initial conditions $x_1(0) = (0.2, 0.8)$, $x_2(0) = (0.2, 0.8)$, when the reward function is defined by Table 4.2 for $a = 4$, $b = c = 1$, $d = 2$, and $\lambda = 0.01$.

4.5 Dynamic reinforcement

4.5.1 Approximate derivative action

So far, the decision (action) vector of agent i at time k, $a_i(k)$, depends only on the probability distribution $x_i(k)$. We would like to explore the case where the decision of each agent i is also affected by a *dynamic* processing of the probability distribution $x_i(k)$.

Assume that a control input $u_i(k)$ also affects the decisions of agent i. Since the

algorithms presented here are decentralized, the control input $u_i(k)$ should *not* make use of the histories of other agents. In particular, suppose that the beliefs are updated by (4.1), while the action vector of agent i depends not only on the state vector, $x_i(k)$, but also on a control input $u_i(k)$, such that:

$$E[\alpha_i(k)|x_i(k)] = \Pi_\Delta\{(1-\lambda)(x_i(k)+u_i(k)) + \lambda/m\},$$

for some $\lambda > 0$.

One possible dynamic reinforcement scheme which is akin to derivative action in classical control is to make use of the *changes* in $x_i(k)$. A standard controls interpretation is that agents use derivative action to "predict" more rewarding outcomes. A similar approach was investigated by [SA05] as an approach to enable stabilization of mixed equilibria in learning in games. The intention here is different. Instead of using derivative action to stabilize a mixed equilibrium, our goal is to use derivative action to enforce convergence to an efficient pure equilibrium.

In particular, we consider *approximate derivative action* (ADA) defined as follows. For each agent $i \in \mathcal{I}$, we introduce two additional state vectors $y_i(k) \in \Delta(|\mathcal{A}_i|)$ and $\rho_i(k) \in \mathbb{R}_+$, which are updated according to the recursions

$$y_i(k+1) = y_i(k) + \epsilon(k) \cdot (x_i(k) - y_i(k)), \tag{4.6}$$

and

$$\rho_i(k+1) = \rho_i(k) + \epsilon(k) \cdot (R_i(\alpha(k)) - \rho_i(k)), \tag{4.7}$$

respectively. In words, for $\epsilon(k) = 1/(k+1)$, $y_i(k)$ is the running average of the strategy vector $x_i(k)$ and $\rho_i(k)$ is the running average of the reward $R_i(\alpha(k))$. Note also that since $x_i(k) \in \Delta(|\mathcal{A}_i|)$ for all large k, then also $y_i(k) \in \Delta(|\mathcal{A}_i|)$ for all large

k.

Now let

$$u_i(k) = \gamma_i(\rho_i(k)) \cdot (x_i(k) - y_i(k)),$$

for a function $\gamma_i : \mathbb{R}_+ \to \mathbb{R}_+$. This additional control term is reinforcing recent changes as reflected by the current $x_i(k)$ and its running average $y_i(k)$, scaled by a feedback gain that in general will be assumed to be reward-dependent. This importance of this dependence will become clear later on when we will discuss the asymptotic properties of the stochastic iteration.

The action vector for each agent $i \in \mathcal{I}$ is then selected according to the rule

$$E[\alpha_i(k)|x_i(k)] = \Pi_\Delta\{(1-\lambda)[x_i(k) + \gamma_i(\rho_i(k)) \cdot (x_i(k) - y_i(k))] + \lambda/\,|\mathcal{A}_i|\}. \quad (4.8)$$

This selection rule, in combination with the update recursion of (4.1), constitutes a reinforcement scheme which we will call "*dynamic reinforcement with approximate derivative action.*" This reinforcement scheme has an extended state vector

$$z(k) \triangleq \begin{pmatrix} x(k) \\ y(k) \\ \rho(k) \end{pmatrix},$$

where $x(k)$ is updated by (4.1), $y(k)$ is updated by (4.6), $\rho(k)$ is updated by (4.7) and the action vector $\alpha(k)$ is selected according to (4.8).

4.5.2 Asymptotic stability of approximate derivative action

The asymptotic stability analysis of approximate derivative action (4.8) will be based on the ODE method for stochastic iterations. These results will follow by applying

suitably modified versions of Propositions 3.8.3–3.8.4. The relevant ODE is now

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{\rho} \end{pmatrix} = \begin{pmatrix} \overline{g}(z) \\ x - y \\ \overline{R}_i(z) - \rho \end{pmatrix}, \tag{4.9}$$

where $\overline{g}(z)$ is defined as in (4.3) except that expectations over actions is now taken according to the selection rule (4.8).

It is straightforward to check that the stationary points of the ODE (4.9) coincide with the stationary points of ODE (4.3), since at the equilibrium $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\rho})$, we have $\tilde{x} = \tilde{y}$.

We need to check whether the asymptotic stability results of the recursion (4.1) can change by appropriately selecting the feedback gain $\gamma_i(\cdot)$. The linearized dynamics of the ODE (4.9) about a stationary point $\tilde{z} = (\tilde{x}, \tilde{x}, \tilde{\rho})$ is:

$$\frac{d}{dt} \begin{pmatrix} \delta x(t) \\ \delta y(t) \\ \delta \rho(t) \end{pmatrix} = \tilde{A}^{\lambda,\gamma} \cdot \begin{pmatrix} \delta x(t) \\ \delta y(t) \\ \delta \rho(t) \end{pmatrix} \tag{4.10}$$

where the perturbation $\delta x(t) = (\delta x_1(t), \ldots, \delta x_n(t))$ is such that

$$x_i(t) = \tilde{x}_i + N \delta x_i(t), \quad i \in \mathcal{I}, \tag{4.11}$$

for some $|\mathcal{A}_i| \times (|\mathcal{A}_i| - 1)$ orthonormal matrix N which spans the null space of the row vector $\mathbf{1}^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{A}_i|}$, i.e.,

$$\mathbf{1}^{\mathrm{T}} N = 0 \quad \text{and} \quad N^{\mathrm{T}} N = I. \tag{4.12}$$

Similarly, $\delta y(t) = (\delta y_1(t), \ldots, \delta y_n(t))$, where

$$y_i(t) = \tilde{y}_i + N\delta y_i(t), \quad i \in \mathcal{I}.$$

Finally, $\delta\rho(t) \triangleq \rho(t) - \tilde{\rho}$.

The matrix N is introduced to reflect that solutions evolve over the probability simplex, and so have a restricted degree of freedom.

Proposition 4.5.1 (LAS of ADA) *For sufficiently small $\lambda > 0$, let $\tilde{z} = (\tilde{x}, \tilde{x}, \tilde{\rho})$ be a stationary point of the ODE (4.9), where $\tilde{x}_i = x_i^* + \lambda w_i^*$ for all $i \in \mathcal{I}$, with $x_i^* = e_{j^*}$ for some $j^* = j^*(i) \in \mathcal{A}_i$ and w_i^* defined according to Proposition 4.4.3. Let Assumption 4.3.1 hold. There exist λ-independent matrices A_{ii}^γ, B_{ii}^γ, $i \in \mathcal{I}$, and matrices $W^{\lambda,\gamma}$, $V^{\lambda,\gamma}$, with*

$$\lim_{\lambda\to 0} \lambda W^{\lambda,\gamma} = 0, \quad \lim_{\lambda\to 0} \lambda V^{\lambda,\gamma} = 0,$$

such that the linearization (4.10) of the ODE (4.9) about $(\tilde{x}, \tilde{x}, \tilde{\rho})$ has system matrix of the form[4]

$$\tilde{A}^{\lambda,\gamma} = \begin{pmatrix} \mathcal{N} & O \\ O & \mathcal{N} \\ I & I \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} A^{\lambda,\gamma} & B^{\lambda,\gamma} & O \\ I & -I & O \\ \times & \times & -I \end{pmatrix} \begin{pmatrix} \mathcal{N} & O \\ O & \mathcal{N} \\ I & I \end{pmatrix}, \tag{4.13}$$

with

$$\begin{aligned} A^{\lambda,\gamma} &= \operatorname{diag}\{A_{ii}^\gamma\}_{i\in\mathcal{I}} + \lambda W^{\lambda,\gamma}, \\ B^{\lambda,\gamma} &= \operatorname{diag}\{B_{ii}^\gamma\}_{i\in\mathcal{I}} + \lambda V^{\lambda,\gamma}, \end{aligned}$$

and $\mathcal{N} = \operatorname{diag}\{N, ..., N\}$ where N is of appropriate dimension. Furthermore, there

[4]The symbol $\times$ corresponds to terms that do not affect the analysis.

exists an $|\mathcal{A}_i| \times |\mathcal{A}_i|$ *unitary matrix* U_i *such that*

$$A_{ii}^{\gamma} = U_i \begin{pmatrix} -\overline{v}_i(j^*, \tilde{x}) & \text{row}\{-(1+\gamma_i(\tilde{\rho}_i))\overline{v}_i(s, \tilde{x})\}_{s \neq j^*} \\ 0 & \text{diag}\{-\overline{v}_i(j^*, \tilde{x}) + (1+\gamma_i(\tilde{\rho}_i))\overline{v}_i(s, \tilde{x})\}_{s \neq j^*} \end{pmatrix} U_i^{\mathrm{T}}, \quad (4.14)$$

and

$$B_{ii}^{\gamma} = U_i \begin{pmatrix} 0 & \text{row}\{\gamma_i(\tilde{\rho}_i)\overline{v}_i(s, \tilde{x})\}_{s \neq j^*} \\ 0 & \text{diag}\{-\gamma_i(\tilde{\rho}_i)\overline{v}_i(s, \tilde{x})\}_{s \neq j^*} \end{pmatrix} U_i^{\mathrm{T}}. \quad (4.15)$$

Proof. See Appendix D.2.1. □

Based on Proposition 4.5.1 we can compute the range of values of the parameter γ_i that guarantees stability of the pure strategy profiles for sufficiently small $\lambda > 0$.

Theorem 4.5.1 (Stability Range) *Assume the hypotheses of Proposition 4.5.1 and let* $\overline{v}_i$ *be the conditional rewards vector (4.4) evaluated at the equilibrium* $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\rho})$ *with* $\overline{v}_i(j^*, \tilde{x}) > \overline{v}_i(s, \tilde{x})$ *for every* $i \in \mathcal{I}$ *and* $s \neq j^*$*. For sufficiently small* $\lambda > 0$*, the equilibrium* $\tilde{z}$ *will be a locally asymptotically stable stationary point of the linearization (4.10) if and only if, for each agent* $i \in \mathcal{I}$*, the derivative feedback gain satisfies*

$$0 < \gamma_i(\tilde{\rho}_i) < \frac{\overline{v}_i(j^*, \tilde{x}) + 1}{\overline{v}_i(s, \tilde{x})} - 1, \quad \forall s \neq j^*. \quad (4.16)$$

Proof. According to the definition of the linearization (4.10), the perturbed state vector δx_i of each agent $i \in \mathcal{I}$ satisfies

$$x_i - \tilde{x}_i = N\delta x_i \in \text{null}\{\mathbf{1}^{\mathrm{T}}\},$$

which implies that

$$x_{ij^*} - \tilde{x}_{ij^*} = -\mathbf{1}^{\mathrm{T}}\mathrm{col}\{x_{is} - \tilde{x}_{is}\}_{s\neq j^*}.$$

Accordingly, we have

$$y_{ij^*} - \tilde{y}_{ij^*} = -\mathbf{1}^{\mathrm{T}}\mathrm{col}\{y_{is} - \tilde{y}_{is}\}_{s\neq j^*}.$$

Thus, the linearized dynamics (4.10) of agent i about $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\rho})$ can be described by the reduced state vector

$$\delta\hat{z} = (\delta\hat{x}, \delta\hat{y}, \delta\rho) \triangleq (\mathrm{col}\{\delta\hat{x}_i\}_{i\in\mathcal{I}}, \mathrm{col}\{\delta\hat{y}_i\}_{i\in\mathcal{I}}, \mathrm{col}\{\delta\rho_i\}_{i\in\mathcal{I}})$$

where $\delta\hat{x}_i \triangleq \mathrm{col}\{x_i(s) - \tilde{x}_i(s)\}_{s\neq j^*}$ and $\delta\hat{y}_i \triangleq \mathrm{col}\{y_i(s) - \tilde{y}_i(s)\}_{s\neq j^*}$. Therefore, according to Proposition 4.5.1, the evolution of the reduced state vector $\delta\hat{z}_i$ takes on the form

$$\frac{d}{dt}\delta\hat{z} = \tilde{A}^{\lambda,\gamma}\delta\hat{z}_i \triangleq \begin{pmatrix} \hat{A}^{\lambda,\gamma} & \hat{B}^{\lambda,\gamma} & O \\ I & -I & O \\ \times & \times & -1 \end{pmatrix}\delta\hat{z}_i$$

with

$$\begin{aligned} \hat{A}^{\lambda,\gamma} &= \mathrm{diag}\left\{\hat{A}^{\gamma}_{ii}\right\}_{i\in\mathcal{I}} + \lambda\hat{W}^{\lambda,\gamma}, \\ \hat{B}^{\lambda,\gamma} &= \mathrm{diag}\left\{\hat{B}^{\gamma}_{ii}\right\}_{i\in\mathcal{I}} + \lambda\hat{V}^{\lambda,\gamma}, \end{aligned}$$

where

$$\hat{A}^{\gamma}_{ii} \triangleq \mathrm{diag}\{-\overline{v}_i(j^*, \tilde{x}) + (1 + \gamma_i(\tilde{\rho}_i))\overline{v}_i(s, \tilde{x})\}_{s\neq j^*}, \quad i \in \mathcal{I}$$

$$\hat{B}_{ii}^{\gamma} = \mathrm{diag}\{-\gamma_i(\tilde{\rho}_i)\overline{v}_i(s,\tilde{x})\}_{s\neq j^*}, \quad i \in \mathcal{I}$$

and $\hat{W}^{\lambda,\gamma}$, $\hat{V}^{\lambda,\gamma}$ are matrices such that

$$\lim_{\lambda\to 0} \lambda \hat{W}^{\lambda,\gamma} = 0, \quad \lim_{\lambda\to 0} \lambda \hat{V}^{\lambda,\gamma} = 0.$$

Therefore, the spectrum of the linearization matrix $\tilde{A}^{\lambda,\gamma}$, when $\lambda = 0$, is given by

$$\mathrm{eig}\tilde{A}^{0,\gamma} = \bigcup_{i\in\mathcal{I}} \mathrm{eig} \begin{pmatrix} \hat{A}_{ii}^{\gamma} & B_{ii}^{\gamma} \\ I & -I \end{pmatrix} \cup \{-1\}.$$

It is straightforward to show that for $\gamma_i(\tilde{\rho}_i) > 0$ such that $1 + \overline{v}_i(j^*,\tilde{x}) - [1 + \gamma_i(\tilde{\rho}_i)]\overline{v}_i(s,\tilde{x})$ is small, the eigenvalues of $\tilde{A}^{0,\gamma}$ are complex numbers with real part

$$\mathrm{Re}\left\{\mathrm{eig}\tilde{A}^{0,\gamma}\right\} = -\frac{1}{2}\left\{1 + \overline{v}_i(j^*,\tilde{x}) - [1+\gamma_i(\tilde{\rho}_i)]\overline{v}_i(s,\tilde{x})\right\}_{i,s\neq j^*}.$$

This implies that for sufficiently small $\lambda > 0$, the strategy profile $\tilde{z}$ will be locally asymptotically stable if and only if condition (4.16) holds. □

Proposition 4.5.2 (ADA convergence) *In the framework of Theorem 4.5.1, if the derivative feedback gains satisfy (4.16) for all $i \in \mathcal{I}$, then $P\{\lim_{k\to\infty} x(k) = \tilde{x}\} > 0$. Instead, if there exist $i \in \mathcal{I}$ and $s \in \mathcal{A}_i \backslash j^*$ such that $\gamma_i(\tilde{\rho}_i) > (\overline{v}_i(j^*,\tilde{x}) - \overline{v}_i(s,\tilde{x}) + 1)/\overline{v}_i(s,\tilde{x})$, then $P[\lim_{k\to\infty} x(k) = \tilde{x}] = 0$.*

Proof. This is a direct consequence of Propositions 3.8.3–3.8.4 and Theorem 4.5.1. □

4.6 Applications

4.6.1 Equilibrium selection in identical interest coordination games

According to Proposition 4.5.2, each agent i, by applying approximate derivative action, can selectively alter the stability properties of the Nash equilibria of a coordination game. In particular, one agent, by appropriately defining the feedback gain, may destabilize certain non-desirable equilibria. For example, in the case of the Typewriter game of Table 4.1(a), the non-efficient equilibrium of (B, B) can be destabilized.

Consider, for example, the payoff matrix of the coordination game of Table 4.2 with $ab + b > dc + c$. In that case, payoff- and risk-dominant equilibrium coincide, which corresponds to the Typewriter game. According to Proposition 4.5.1, when λ is sufficiently small, the only stable equilibria of the learning dynamics are small variations of (A, A) and (B, B).

Since agents may not be aware of the current payoff matrix, varying the derivative feedback gain allows them to experiment. In particular, suppose that both agents run the stochastic iteration of (4.1) and that agent 1 applies approximate derivative action with $\gamma_1(\tilde{\rho}_1) \equiv \gamma > 0$. According to Theorem 4.5.1 the coordination states (A, A) and (B, B) are locally stable if and only if $\gamma < (a - c + 1)/c$ and $\gamma < (d - b + 1)/b$, respectively. Agent 1 is able to destabilize the non-efficient equilibrium (B, B) by gradually increasing γ, since there is a range of gains that destabilizes (B, B), while (A, A) remains stable. In particular, any γ such that $(d-b+1)/b < \gamma < (a-c+1)/c$ accomplishes that. We summarize this conclusion in the following claim:

Claim 4.6.1 (Dynamic Reinforcement in the Typewriter Game) *Consider the learning dynamics (4.1) with step size sequence (4.2) with two agents and two actions. Assume the payoff matrix of Table 4.2 with $a > c > 0$, $d > b > 0$, $a > d$ and $ab+b > dc+c$. For sufficiently small $\lambda > 0$, the learning dynamics exhibit stationary*

points $\tilde{x}^1$ *and* $\tilde{x}^2$ *which are small variations of the two pure strategy profiles* (A, A) *and* (B, B)*, respectively. If agent* $i \in \{1, 2\}$ *applies approximate derivative action (4.8) with* $\gamma_i(\rho_i) \equiv \gamma$ *for all* $\rho_i \in \mathbb{R}_+$*, such that* $(d - b + 1)/b < \gamma < (a - c + 1)/c$*, then* $P\{\lim_{k\to\infty} x(k) = \tilde{x}^1\} > 0$ *and* $P\{\lim_{k\to\infty} x(k) = \tilde{x}^2\} = 0$*.*

Proof. The proof is a direct application of Proposition 4.5.2. The condition $ab + b > dc + c$ guarantees that there are feasible feedback gains that can destabilize (B, B) while (A, A) is stable. □

Note that when $b = c$ the condition $ab + b > dc + c \Leftrightarrow a + b > d + c$ which implies that the equilibrium profile (A, A) risk-dominates (B, B). In general, however, this is not the case.

Fig. 4.3 shows the solution of the ODE (4.9) when $a = 5$, $b = c = 1$ and $d = 2$. We also assume that agent 1 applies approximate derivative action with $\gamma_1 = 3.5$. According to Claim 4.6.1, there is zero probability that the stochastic process converges to (B, B) when $2 < \gamma_1 < 5$. For an initial condition that is very close to the non-efficient equilibrium (B, B), Fig. 4.3 shows that the solution escapes the non-efficient equilibrium, despite being initiated very close to it, and convergence to (A, A) is attained. Also, in Fig. 4.4, a typical response of the stochastic iteration (4.1) is shown, which illustrates that the process does not converge to the non-efficient equilibrium.

Since in Theorem 4.5.1 there is no constraint on the number of agents or actions, Claim 4.6.1 can be easily extended to the multiplayer case. Consider, for example, the Typewriter game of three agents and two actions of Table 4.3.

It is straightforward to see that the conclusions of Claim 4.6.1 continue to hold, where each agent $i \in \{1, 2, 3\}$ is able to destabilize the non-efficient equilibrium (B, B, B) by applying approximate derivative action (4.8) with $\gamma_i(\rho_i) \equiv \gamma$ for all $\rho_i \in \mathbb{R}_+$, such that $(d - b + 1)/b < \gamma < (a - c + 1)/c$. Note also that in the example

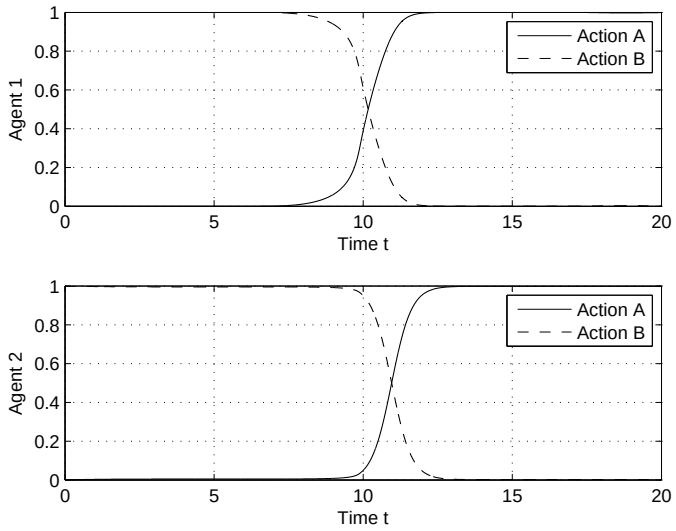


Figure 4.3: The solution of ODE (4.9) with initial conditions $x_1(0) = x_2(0) = (0, 1)$ and $y_1(0) = (1, 0)$, when the reward function is defined by Table 4.2 for $a = 5$, $b = c = 1$ and $d = 2$, while the decisions are taken according to (4.8) with $\lambda = 0.01$, $\gamma_1 = 3.5$ and $\gamma_2 = 0$.

	$2.A$	$2.B$
$1.A$	$5, 5, 5$	$1, 1, 1$
$1.B$	$1, 1, 1$	$1, 1, 1$

$3.A$

	$2.A$	$2.B$
$1.A$	$1, 1, 1$	$1, 1, 1$
$1.B$	$1, 1, 1$	$2, 2, 2$

$3.B$

Table 4.3: The Typewriter game of 3 players and 2 actions

of Table 4.3, $b = c = 1$. In that case, the solution of the ODE, when agent 1 applies approximate derivative action with $\gamma = 3.5$ is shown in Fig. 4.5.

4.6.2 Equilibrium selection in aligned interest coordination games

Similarly to the equilibrium selection in the Typewriter game, in the case of the Stag-Hunt game of Table 4.1(b), the non-efficient (risk-dominant) equilibrium (B, B) can also be destabilized by appropriately defining the feedback gain.

Note that in the case of the Typewriter game, a single agent is able to destabilize the non-efficient equilibrium (B, B) by gradually increasing the feedback gain. Intu-

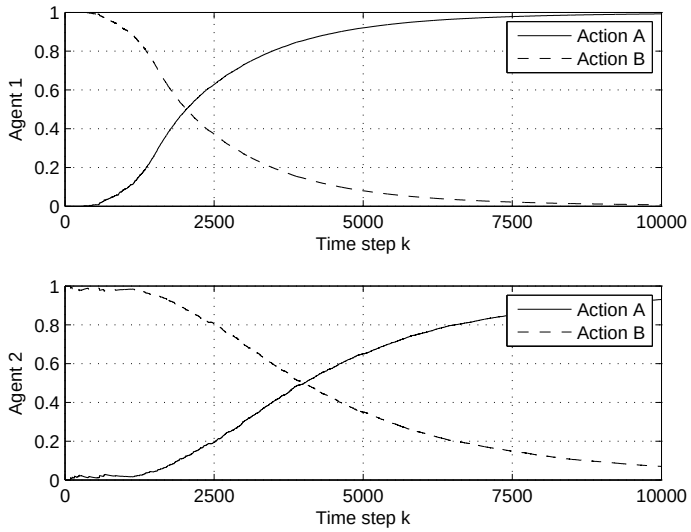


Figure 4.4: A typical response of the stochastic iteration (4.1) with initial conditions $x_1(0) = x_2(0) = y_1(0) = (0, 1)$, when the reward function is defined by Table 4.2 for $a = 4$, $b = c = 1$ and $d = 2$, while the decisions are taken according to (4.8) with $\lambda = 0.01$, $\gamma_1 = 3.5$ and $\gamma_2 = 0$.

itively, this is possible because the deviation cost from (B, B), $d - b$, is less than the deviation cost from (A, A), $a - c$. However, in the case of the Stag-Hunt game this is not the case. Instead, the deviation cost from (B, B) is larger than the deviation cost from (A, A). Therefore, a feedback gain $\gamma_i(\rho_i)$ that is constant for all ρ_i cannot destabilize (B, B).

Let us consider, instead, a payoff-dependent feedback gain of the form

$$\gamma_i(\rho_i) \triangleq \frac{\gamma}{\rho_i^\kappa}, \tag{4.17}$$

where $\gamma > 0$ and $\kappa > 1$ are constants. Such a feedback gain is large when ρ_i is small, and vice versa. Since the approximate derivative action can be thought of as a way of exploiting more rewarding actions, it is natural to consider a small feedback gain when the current payoff is large, and vice versa.

Claim 4.6.2 (Dynamic Reinforcement in the Stag-Hunt Game) *Consider the*

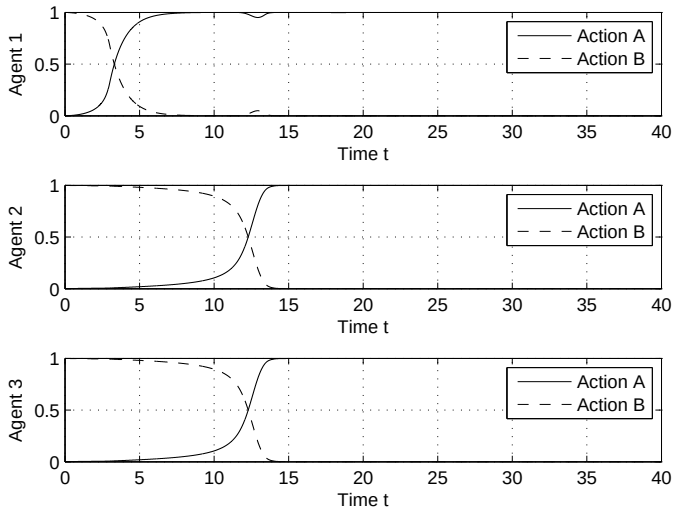


Figure 4.5: The solution of ODE (4.9) with initial conditions $x_1(0) = x_2(0) = x_3(0) = (0, 1)$ and $y_1(0) = (1, 0)$, when the reward function is defined by Table 4.3, while the decisions are taken according to (4.8) with $\lambda = 0.01$, $\gamma_1 = 3.5$ and $\gamma_2 = \gamma_3 = 0$.

learning dynamics (4.1) with step size sequence (4.2) with two agents and two actions. Assume the payoff matrix of Table 4.2 with $a > c > 0$, $d > b > 0$, $a > d$ *and* $a + b < d + c$. *For sufficiently small* $\lambda > 0$, *the learning dynamics exhibit stationary points* $\tilde{x}^1$ *and* $\tilde{x}^2$ *which are small variations of the two pure strategy profiles* (A, A) *and* (B, B), *respectively. If agent* $i \in \{1, 2\}$ *applies approximate derivative action (4.8) with feedback gain (4.17), such that*

$$\kappa > \frac{\log\left(\frac{a-c+1}{d-b+1} \cdot \frac{b}{c}\right)}{\log\left(\frac{d}{a}\right)} \tag{4.18}$$

and

$$d^{\kappa}\frac{d-b+1}{b} < \gamma < a^{\kappa}\frac{a-c+1}{c}, \tag{4.19}$$

then $P\{\lim_{k\to\infty} x(k) = \tilde{x}^1\} > 0$ *and* $P\{\lim_{k\to\infty} x(k) = \tilde{x}^2\} = 0$.

Proof. For sufficiently small $\lambda > 0$, equilibrium profile (A, A) is linearly stable if and only if

$$\gamma_i(a) < (a - c + 1)/c \Leftrightarrow \gamma < a^\kappa \frac{a - c + 1}{c},$$

while equilibrium profile (B, B) is unstable if and only if

$$\gamma_i(d) > (d - b + 1)/b \Leftrightarrow \gamma > d^\kappa \frac{d - b + 1}{b}.$$

Therefore, the conclusion follows if γ satisfies condition (4.19) and

$$d^\kappa \frac{d - b + 1}{b} < a^\kappa \frac{a - c + 1}{c}$$

which is equivalent to condition (4.18). □

Fig. 4.6 shows the solution of the ODE (4.9) when $a = 5$, $b = 1$, $c = 4$ and $d = 3$. We also assume that agent 1 applies approximate derivative action according to (4.17). According to Claim 4.6.2, when $\kappa > 2.151$ there exists γ for which the stochastic process does not converge to (B, B). For example, if $\kappa = 5$, then for any $729 < \gamma < 3125$, the equilibrium profile (B, B) is unstable, while (A, A) is stable.

For an initial condition that is very close to the non-efficient equilibrium (B, B), Fig. 4.6 shows that the solution escapes the non-efficient equilibrium, despite being initiated very close to it, and convergence to (A, A) is attained. Also, in Fig. 4.7, a typical response of the stochastic iteration (4.1) is shown, which illustrates that the process does not converge to the non-efficient equilibrium.

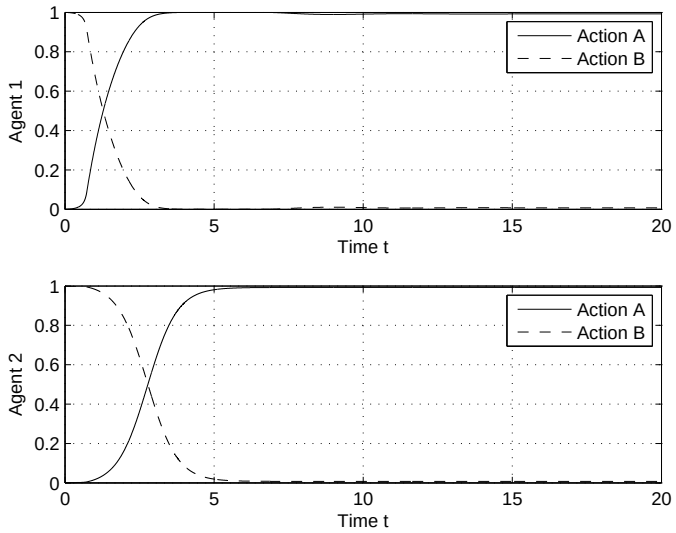


Figure 4.6: The solution of ODE (4.9) with initial conditions $x_1(0) = x_2(0) = y_1(0) = (0, 1)$, when the reward function is defined by Table 4.2 for $a = 5$, $b = 1$, $c = 3$ and $d = 3$, and agent 1 applies approximate derivative action (4.8) with $\lambda = 0.01$ and γ_1 defined by (4.17) with $\gamma = 2000$ and $\kappa = 5$.

4.6.3 Equilibrium selection in distributed network formation

We consider the problem of distributed network formation as introduced in Section 4.2.2. We assume that nodes apply the learning algorithm of (4.1), where the action set $\mathcal{A}_i$ for each agent includes all the possible combinations of neighboring nodes (equivalently, links), i.e., $\mathcal{A}_i \triangleq 2^{\mathcal{N}_i}$, where $\mathcal{N}_i$ is the set of neighboring nodes of node i. For example, in the case of $n = 3$ agents of Figure 4.1, the set of actions of agent 1 will be $\mathcal{A}_1 = \{\{1\}, \{2\}, \{3\}, \{2, 3\}\}$, where for example, action $\{1\}$ implies that agent 1 does not establish any link, and action $\{2, 3\}$ implies that agent 1 creates a link with both agents 2 and 3. Similarly to the previous analysis, we may define an enumeration $\{1, 2, ..., |\mathcal{A}_i|\}$ of the actions in $\mathcal{A}_i$. To minimize notation, the set $\mathcal{A}_i$ will also denote the corresponding set of vertices $\{e_1, ..., e_{|\mathcal{A}_i|}\}$ in $\Delta(|\mathcal{A}_i|)$.

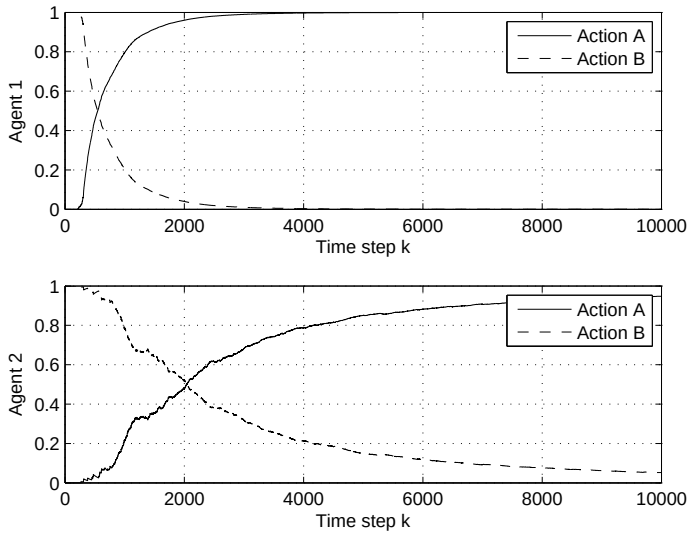


Figure 4.7: A typical response of the stochastic iteration (4.1) with initial conditions $x_1(0) = x_2(0) = y_1(0) = (0, 1)$, when the reward function is defined by Table 4.2 for $a = 5$, $b = 1$, $c = 3$ and $d = 3$, and agent 1 applies approximate derivative action (4.8) with $\lambda = 0.01$ and γ_1 defined by (4.17) with $\gamma = 2000$ and $\kappa = 5$.

We assume that agents apply the following variation of (4.1):

$$x_i(k+1) = x_i(k) + \epsilon(k) \cdot [R_i(\alpha(k)) - C_i(\alpha_i(k))] \cdot [\alpha_i(k) - x_i(k)], \tag{4.20}$$

where $R_i : \mathcal{A} \to \mathbb{R}_+$ is the reward of agent i that depends on the action profile $\alpha = (\alpha_1, ..., \alpha_n) \in \mathcal{A} \triangleq \times_{i \in \mathcal{I}} \mathcal{A}_i$ and $C_i : \mathcal{A}_i \to \mathbb{R}_+$ denotes the cost associated with the links of agent i. We assume that the cost of establishing a link is always strictly less than the benefits of that link, so that the empty network cannot be a Nash equilibrium.

Here, we define the benefits of agent i, $R_i(\cdot)$, to be equal to the number of nodes that are accessible to i through direct and indirect connections following the orientation of the graph. The cost function of agent i, $C_i(\cdot)$, is simply defined as constant c, such that $0 < c < 1$, for each link that is established by agent i.

It is straightforward to check that the asymptotic analysis of Section 4.4 can be

applied here. Furthermore, the dynamic process exhibits multiple Nash equilibria. For example, in the case of $n = 3$ agents, the *wheel* and the *star* network of Fig. 4.1 are Nash equilibria. In particular, the wheel network of Fig. 4.1(a) is a strict Nash equilibrium (all eigenvalues of the linearized ODE about this equilibrium are strictly negative), while the star network of Fig. 4.1(b) is a Nash equilibrium (all eigenvalues of the linearized ODE about this equilibrium are non-positive).

According to Proposition Proposition 3.8.3, convergence to the star network is *not* excluded. This problem fits to the Typewriter problem of Section 4.6.2. The star network can be destabilized if either agent 2 or 3 in Fig. 4.1(b) applies the approximate derivative action (4.8) with a small feedback gain, due to the fact that the deviation cost for those agents is zero.

In particular, for $i = 2$ or $i = 3$ in the star network of Fig. 4.1(b), we have $v_{ij^*} = 2 - c$, where $j^* = j^*(i)$ corresponds to the current action, and $\max_{s \neq j^*} v_{is} = 2 - c$. Therefore, according to Theorem 4.5.1, if agent $i \in \{1, 2, 3\}$ applies dynamic reinforcement with $\gamma_i \geq (2 - c + 1)/(2 - c) - 1 \equiv 1/(2 - c)$, then the star network is linearly unstable. Accordingly, in the wheel network of Fig. 4.1(a), we have $v_{ij^*} = 2-c$ and $\max_{s \neq j^*} v_{is} = 2-2c$, which corresponds to the case of establishing two links. Thus, if agent $i \in \{1, 2, 3\}$ applies dynamic reinforcement with $\gamma_i < (2-c+1)/(2-2c)-1 \equiv (1+c)/(2-2c)$, then the wheel network is locally asymptotically stable. We conclude that:

Claim 4.6.3 (Dynamic Reinforcement in Network Formation) *Consider the learning dynamics (4.20) with step size sequence (4.2) with the network formation framework presented here with $n = 3$ agents. For sufficiently small $\lambda > 0$, the learning dynamics exhibit two stationary points $\tilde{x}^1$ and $\tilde{x}^2$, which are small variations of the wheel and star network of Figs. 4.1(a)-(b), respectively. If agent $i \in \{1, 2, 3\}$ applies approximate derivative action (4.8) with $\gamma_i(\rho_i) \equiv \gamma$ for all $\rho_i \in \mathbb{R}_+$, such that $1/(2 - c) \leq \gamma < (1 + c)/(2 - 2c)$, then $P\{\lim_{k\to\infty} x(k) = \tilde{x}^1\} > 0$ and $P\{\lim_{k\to\infty} x(k) =$*

$\tilde{x}^2\} = 0$.

This result will be extended to the multiplayer case in Chapter 5.

4.7 Remarks

We considered the problem of distributed convergence to efficient outcomes where agents have access only to their own prior actions and received rewards. We showed that convergence to an efficient coordination structure (i.e., efficient global outcome) can be reinforced by dynamic processing of only local information. This illustrates how (unilateral) local decisions can affect the aggregate outcome of an evolutionary process.

We used a simple form of dynamic processing, which reinforces recent changes of the state from its running average. We showed that each agent by applying this dynamic reinforcement scheme is able to destabilize non-efficient equilibria by appropriately adjusting the derivative feedback gain. We specialized our results in coordination games, where risk- and payoff-dominant equilibria might not coincide, and we showed that destabilization of the non-efficient or risk-dominant equilibrium is possible. We also illustrated the utility of such reinforcement scheme in a network formation process, which will be analyzed in further detail in the next chapter.

CHAPTER 5

Efficient Network Formation by Distributed Reinforcement

5.1 Introduction

Recent research on social networks has shown how structure affects norm and attitude formation in a society [Fri01]. Moreover, the efficiency of flow of information through a social network has an important relationship with one of the most popular studies of social networks, which is the study of centrality, social power and influence as a function of the structure of, and positions in, social networks. Likewise, a challenge in sensor networks is to design protocols that guarantee energy efficient network formation, where the energy of transmitting signals is the major part of the energy consumption [SGA00, ASS02]. In this paper, we wish to provide a dynamic framework that will serve both as a design procedure for distributed convergence to a desirable network and as a justification for the emergence of certain networks.

In particular, according to [Jac03], some questions addressed by the problem of network formation are:

1. *Which networks are likely to form* when agents have the discretion to choose their connections?

2. If there are several networks that are likely to emerge, *how the likelihood of emergence of each of these networks is related to the mechanism of interpersonal interaction*?

3. *How are such network relationships important in determining the outcome of a process*, such as information flow, attitude or norm formation?

4. *How efficient are the networks that form* and how does that depend on the way agents interact locally?

In this chapter, we consider the problem of efficient network formation in a distributed fashion. To this end, we formulate network formation as a game of learning automata, where each node corresponds to a different learning automaton. According to this formulation, agents can form and sever unidirectional links and derive direct and indirect benefits from these links. Also, each agent's choices depend on its own previous links and past benefits, and link selections are subject to random perturbations.

In the following sections, we first present a small review of different approaches on formulating the problem of network formation in a distributed fashion. These methods model network formation as a game where each agent has discretion over its own links. Then, we proceed on characterizing the stability properties of the proposed model. We illustrate the flexibility of the model to incorporate various design criteria, including dynamic cost functions that reflect link establishment and maintenance, and distance-dependent benefit functions. We show that the learning process assigns positive probability to the emergence of multiple stable configurations (i.e., strict Nash networks), which need not emerge under alternative processes such as best-reply dynamics. We analyze the specific case of so-called frictionless benefit flow, and show that a single agent can reinforce the emergence of an efficient network through an enhanced evolutionary process known as dynamic reinforcement.

5.2 Network formation as a game

Our objective in this section is to analyze and compare different theoretical approaches that have been proposed to address the problem of distributed network formation. Based on the techniques used these approaches can be classified into the following categories:

- *Game-theoretic static models*, where the problem of network formation is usually modelled as the strategic interaction of several agents in an one-stage game.
- *Game-theoretic dynamic models*, where the problem of network formation is modelled as a game which is played repeatedly.
- *Social evolutionary models*, where agents react adaptively to the circumstances facing them.

Moreover, both static and dynamic models belong to the area of *conventional game theory* and as we will see they can be distinguished into two subcategories:

- *cooperative models*, where mutual consent is needed to form a link. Because of that, either some sort of coalitional equilibrium concept is required, or the game needs to be an extensive form with a protocol for proposing and accepting links in some sequence.
- *noncooperative models*, where agents are considered as opponents trying to establish those links that will maximize their own utility. In this case, mutual consent is not required to form a link.

At a first sight, it seems that dynamic and evolutionary models are more useful than static models, since they describe the mechanism under which a graph emerges. However, some of the most important concepts in network analysis, such as *pairwise stability* or *efficiency* of a graph were introduced in the framework of static models.

5.2.1 Game-theoretic static models

Reference [Mye77] is probably one of the first important contributions to this literature. Myerson analyzes a cooperative game that is enriched by a network structure describing the possibilities for communication or cooperation among different agents. Agents can act as a coalition if and only if they are connected through links in the network. While this idea constitutes an important step forward, it leaves several issues unsolved. In particular, because the value (or reward) function is still defined on coalitions and not on the network directly, the theory does not distinguish between different networks that connect the same agents but differ in the way these agents are connected. As a consequence, many interesting details of the network formation process, for example costs and benefits of particular links, cannot be analyzed by the model of [Mye77].

Reference [AM88] were the first to take an explicit look at network formation in a strategic context, where agents had discretion over their connections. In particular, reference [AM88] were the first to model network formation explicitly as a game, and did so by describing an extensive form game for the formation of a network in the context of cooperative games with communication structures. In their game, agents sequentially propose links which are then accepted or rejected. The extensive form game begins with an ordering over possible links. The game is such that each pair of agents decide on whether or not to form a link knowing the decisions of all pairs coming before them, and forecasting the play that will come after them. A decision to form a link is binding and cannot be undone. In terms of its usefulness as an approach to modeling network formation, this game has some nice features to it. However, the extensive form makes it difficult to analyze beyond very simple examples and the ordering of links can have a non-trivial impact on which networks emerge. Moreover, one of the most important theoretical debate stemming from [AM88] is about the potential conflict between efficient and stable networks.

Reference [Mye91] suggests a different game for modeling network formation. It is in a way the simplest one that one could come up with, and as such is a natural one. It can be described as follows:

- The strategy space of each agent is the list of other agents.
- Agents (simultaneously) announce which other agents they wish to be connected to.
- A link is formed if and only if both agents select each other.

This game has the advantage of being very simple and directly capturing the idea of forming links. Unfortunately, it has a large multiplicity of Nash equilibria. For example, *the empty network is always a Nash equilibrium*, regardless of what the payoffs are. The idea is that no agents suggests any links under the correct expectation that no agents will reciprocate. This is especially unnatural in situations where links result in some positive payoff. That means that in order to make use of this game, *one must really use some refinement of Nash equilibrium*. Reference [DM97] discuss some refinements in detail and the relationship of the equilibria to the concept of pairwise stability.

Closer to the work of cooperative games, e.g., [Mye77], is the model of [JW96]. Its main contribution lies in the introduction of an *allocation rule* that assigns a *reward* to each agent that is not necessarily equal to its *production*. Based on the rewards, first we can analyze the stability of the networks, and second we can distinguish between the set of networks which are productively efficient, and those which are stable. This work differs from the literature of cooperative games in some important respects: (a) The value of the network can depend on exactly how agents are interconnected, not just who they are directly or indirectly connected to; (b) This work focuses on network stability and formation and its relationship to efficiency.

According to this model, agents directly communicate with those to whom they are linked. Through these links, they also benefit from indirect communication from those to whom their adjacent nodes are linked, and so on (*connections model*). Also, this model uses two quantities, the *value of the network* (which is the sum of all productions) and *the allocation rule* (that assigns rewards to each agent). Reference [JW96] tries to answer the question of whether there are strongly efficient graphs that are pairwise stable. In other words, if we are free to structure the allocation rule in any way we like, is it possible to find one such that there is always at least one strongly efficient graph which is pairwise stable?

References [DM97] and [DJ00] are extensions of the work of [JW96]. In particular, reference [DM97] deals with the problem of constructing allocation rules for which efficient networks are pairwise stable. Similarly, reference [DJ00] derives the results of [JW96] for the case of *directed* communication networks. In particular, it is investigated under which conditions (i.e., allocation rules) an efficient network is individually stable in the context of directed network models. The reason of examining directed networks is that the set of applications for the directed and non-directed models is quite different. Examples of directed network models include the production and transmission of gossip and jokes, to information about job opportunities, securities, consumer products, and even information regarding the returns to crime.

The above methods of modeling network formation are such that the network formation process and the allocation of value among agents in a network are separated. Reference [CM00] provides an interesting approach where the allocation of value (or reward) among agents takes place simultaneously with the link formation, i.e., agents may bargain over their shares of value as they negotiate whether or not to add a link.[1] The simultaneous bargaining over allocations and network formation can make an important difference in conclusions about the efficiency of the networks that are

[1]See also [SN00] for similar approach.

formed. The main difficulty with this approach is the specification of the bargaining game, whose fine details (such as how the game ends) can be very important in determining what networks form and how value is distributed.

5.2.2 Game theoretic dynamic models

Dynamic models analyze how networks emerge and how the decisions of agents contribute to network formation. Some of the most important works in this area include the models of [Wat01], [JW02a] and [BG00].

The majority of the static models discussed so far assume that links can be formed if and only if there is mutual consent from both agents. This is primarily the reason for which these models were called *cooperative models*. The models of [Wat01] and [JW02a] belong to the same category of cooperative models. Instead, the model of [BG00] introduces a noncooperative framework, where mutual consent is not necessary for establishing a link.

Reference [Wat01] was the first to model network formation as a dynamic process where networks are formed over time. The process begins with an empty network. At each time a link is randomly identified. The current network is altered if and only if the addition or deletion of the link would defeat the current network.[2] Thus, agents add or delete links through myopic considerations of whether this would increase their payoffs.[3] A network has reached a *stable state* if there is some time after which no links would ever be added or deleted.

A difficulty with the idea of a stable state is that in some situations one can get stuck at the empty network because any single link results in a negative value, even though it might be that larger networks are valuable. If one can start at any

[2]Agents decide whether or not they add or severe a link by considering the corresponding reward. Essentially, they play a best reply.

[3]Both agents need to agree but their decisions are myopic (i.e., an agent agrees if and only if it benefits from establishing or severing the link.)

network, then any stable network could be reached by an improving path. But without specifying the process more fully, it is not clear what the right starting conditions are. Introducing some stochastics into the picture solves this quite naturally.

The introduction of random perturbations to the formation process was first studied in [JW02a]. In this work, the framework of the model of [JW96] is altered, so that the emergence of a network is the result of a dynamic process. In particular, a dynamic model is introduced in which agents form and severe links based on the improvement that the resulting network offers them relative to the current network. Then, each agents receives a payoff based on the network configuration that is in place. In other words, at each iteration agents *myopically* decide whether to form a new link or severe an existing link. This decision is based on the improvement this action will cause in their payoff.[4] However, this process may result in cycles, where a number of agents are repeatedly visited.

Thus, the main difference of the model of [JW02a] from the dynamic model of [Wat01] is that the evolution of the network is now more natural. In [Wat01], at each iteration, a link is randomly identified to be updated with uniform probability. If the link already existed, then both parties can decide whether or not to sever this link. If the link is not currently existing, then both parties can form the link (if it in their interest to do so) and simultaneously sever any of their other links if both parties agree. Instead, in [JW02a], any link can be updated at any iteration (i.e., there is no central authority that decides which link will be updated). That is the main reason for which cycles might occur.

Reference [JW02a] also examines the effect of several stochastic changes to a network. For example such stochastic changes model situations where two agents will add a link that they normally would not add, or a single agent will severe a link that it normally would not sever. This random element (mutation) in the process will

[4]Note that the knowledge of the payoffs corresponding to each action is necessary in order to make a decision.

allow the dynamic formation process to deviate from an improving path.

These stochastic mutations in the formation process have several different interpretations or justifications. They might represent errors made by the agents. They might also represent a lack of knowledge on the part of the agents and be a form of experimentation. Such mutations might also be due to exogenous factors that are beyond the agents' control.

This stochastic process defines a Markov chain and a set of evolutionary stable networks can be derived as the effect of mutations goes to zero.[5] The resulting evolutionary stable networks depend on the problem's setup, the value function and the allocation rule. Reference [JW02a] also examines under which conditions stable configurations are efficient.

Somewhat parallel to [JW02a], reference [BG00] develop models of network formation that use tools from noncooperative game theory. Rather than considering pairwise stability, reference [BG00] assumes that *agents can form and sever links unilaterally*, i.e., no mutual consent is needed to form a link between two agents. Clearly, this assumption changes the incentives of the agents, hence the analysis in [BG00] differs substantially from the analysis in the models mentioned above. A central implication of unilateral link formation is that it leads to the concept of Nash equilibrium.

The main idea of the network model in [BG00] is similar to the connections model of [JW96]. In particular, a finite set of agents is considered, where each agent is a source of benefits that other can tap via the formation of costly pairwise links. A link with another agent allows access to the benefits available to the latter via its own links. The costs of link formation are incurred only by the agent who initiates the link. This allows for modeling the network formation process as a *noncooperative game*, where an agent's strategy is a specification of the set of agents with whom it

[5] A network that is in the support of the limiting (as the probability of mutation goes to 0) stationary distribution of the above-described Markov process is evolutionary stable.

forms links. The links formed by agents define a social network.

An important result of this work is that *Nash networks are either connected or empty.* With one-way flows a society with 6 agents can have upwards of $20,000$ Nash networks representing more than 30 different architectures. This multiplicity of Nash equilibria motivates an examination of a stronger equilibrium concept. Agents have to choose among these different equilibria. This leads to study the process by which agents learn about the network and revise their decisions on link formation, over time.

[BG00] use a version of the best-reply dynamic to study this issue. In particular, *the best-reply dynamic* has the following features:

- *Repeated game.* The network formation game is played repeatedly, with agents making investments in link formation in every period.
- *Decision making.* When making its decision *an agent chooses a set of links that maximizes its payoffs given the network of the previous period.*

  Two assumptions are made:

  1. *Agents exhibit inertia.* There is some probability that an agent chooses the same strategy as in the previous period. This ensures that agents do not perpetually miscoordinate.
  2. *Randomization.* If more than one strategy is optimal for an agent, then it *randomizes* across the optimal strategies. This requirement implies, in particular, that a non-strict Nash network can never be a steady state of the dynamics.

The rules on agent behavior define a Markov chain on the state space of all networks; moreover, the set of absorbing states of the Markov chain coincides with the set of strict Nash networks of the one-stage game. In the case of *one-way* and *frictionless*[6] flow of benefits, the only strict Nash architectures are the empty network

[6]Indirect rewards are not discounted.

and the wheel, while in *two-way flow of benefits* the Nash architectures are the empty network and the center-sponsored star.

5.2.3 Social evolutionary models

The models of network formation discussed in the previous sections (static or dynamic) make use of a game theoretic approach, where agents try to maximize their reward by selecting appropriately their pairs in either a cooperative or noncooperative manner.

As we have already mentioned, static models help us understand the notions of pairwise stability and efficiency of a network, while dynamic models are more important since they provide also the mechanism under which a network emerges. However, as we commented in the model of [JW02a], the inclusion of uncertainties (mutations) in the process of decision making was necessary, for several reasons, such as to model either

1. a lack of knowledge on the part of the agents, or
2. experimentation, or
3. some errors that agents might make.

The third interpretation is more closely related to the opposition to the notion of Nash equilibrium. Nash equilibrium play assumes that every agent is rational, and this rationality is common knowledge. However, one could easily ask: "*what happens if a agent who is sure to play its equilibrium strategy does not do so?*," [BS92].

Evolutionary models incorporate both experimentation and decision errors in such a way that the strong assumptions of Nash equilibrium play are not necessary. In these models, agents will learn how to play the game through time by adaptively reacting to the circumstances facing them. Moreover, these models can model situations of

incomplete information, where agents can realize only their own reward, but not the actions played by the other agents.

Some characteristic examples of evolutionary models of network formation are [Zeg94], [Zeg95] and [SP00]. The model presented in [Zeg94, Zeg95] is a dynamic model that transforms the choices of agents in a closed group (initially mutual strangers with different characteristics) into the resulting structures of a friendship network. Although it does not follow a game theoretic approach (as most of the models do in economic networks), it does point out some of the important aspects of the dynamic models discussed before, such as the incompatibility between stable and efficient networks. This shows that the problem of network formation in either an economic or sociological framework can be dealt in a unified way.

Returning to the model of [Zeg94, Zeg95], it is stressed out that having a model that can predict future structure from an initial situation of mutual strangers is important. Up to that time, the main literature in social networks was dealing only with the effects of dyadic or triadic substructures in the final structure. However, there was no theoretical model that explained how a dyadic or triadic substructure was created.[7] The goal of [Zeg94, Zeg95] was to model the process of friendship formation between two agents in the surroundings of more people, in order to grasp the dynamic process of the evolving friendship network. The behavioral rules of agents are based on *tension minimization* with respect to the so-called *issues*. An *issue* is any kind of dimension with respect to friendship relations on which the agents have values, e.g. the need for social contact (the desired number of friends). The *tension function* with respect to an *issue* measures the discrepancy between the value of this issue from its desired value. So, a small tension corresponds to large utility. The actions performed by the agents depend on the tension values attached to the issues.

[7] In the framework of balance theory, [Joh86] attempts to specify in more detail the link between the micro and macro level, but defines the triad level as the micro level. However, decisions about establishing or dissolving relations are not made at the triad level, but are the aggregate level of such decisions.

Let us consider only one *issue*, the number of friends that each agent has, [Zeg94]. A friendship (or link) between two agents can be established if and only if it is mutual (reciprocated). In order to model the dynamic process, actions are taken in a discrete-time manner and the following assumptions are made:

1. *Every agent tries to optimally reduce its tension* by *randomly* extending as many choices as its unreciprocated links. If some of its choices are unreciprocated, then it adds choices so that the total number of friends matches the desired one.

2. *Information is not complete.* An agent is not informed about any structural characteristics of the network but does know the number of agents in the closed set. An agent observes only actions and choices that concern itself.

3. The network is in equilibrium if all agents have tension 0, or those agents which do not have the desired number of friends cannot lower their tension.

It is important to point out that this model assumes that *agents do not know the current state of the network*, instead of the dynamic models presented in the previous section, where agents try to improve their reward given the current state of the network. For example in the model of [BG00] agents choose those actions that will maximize their utility assuming that the other agents will play their previous actions, which are known. Instead in the model of [Zeg94, Zeg95], *each agent observes only actions and choices that concern himself.*

One of the characteristic results of [Zeg94, Zeg95] is that *macro-level outcomes are often not globally optimal as a result of the local optimization of the agents.* In particular, if we define the *network tension* as the summation of the tensions of all agents, then the network tension can be viewed as a measure of global satisfaction. A global optimum situation (network tension is zero) in which every agent would be best off and have tension zero occurs only with small probability. This result is

somehow similar to the results of other dynamic models, where efficient networks (which corresponds to a global optimum) are not necessarily stable networks.

Reference [SP00] introduced a learning algorithm where agents play repeated games in pairings determined by a stochastically evolving social network. The rewards received from each agent determine which interactions will be reinforced, and the network structure emerges as a consequence of the dynamics of the agents' learning behavior. The learning process belongs to the general class of *Polya urn models*, as defined in [BST04], and it constitutes a *learning automaton*. A general description of the dynamic process is the following:

- At each time interval, each agent chooses an agent based on some probability distribution over all possible agents.
- When a link is created between two agents, an interaction takes place. Their interaction can be modelled as a strategic form game, where agents try to maximize their own utility function over all possible actions.
- Based on their payoff, agents update their probability distribution for choosing pairings according to some adaptive rule. For example, an agent who obtains unsatisfactory results may choose either to change strategies or to change associates.

Since the model of [SP00] is a learning algorithm,

1. agents are experimenting over their possible pairings and they reinforce the most prosperous choices,
2. agents do not necessarily maintain a link for a large amount of time, which agrees with the dissolution of friendship relationships in the real world,
3. the notion of Nash equilibrium is not necessary to describe the equilibrium network structure.

Something that it is not examined in [SP00] is *how such an adaptive scheme of updating friendships can help us understand better the difference between stable and efficient networks*, or *if it is possible to reinforce certain desirable global outcomes.*

5.3 Our approach

Our work is motivated by the current research on social network formation, and, more specifically, on how the emergence of specific forms of networks is associated with the strategic framework of local interactions [Jac03].

Our approach is concerned with dynamic or evolutionary models, and is mostly related to [DJ00, BG00, SP00, BL03]. In particular, we consider self-interested agents that have the discretion of establishing or severing unidirectional links with neighboring agents based on myopic considerations. However, we drop the typical assumption that agents are aware of the current network structure. Accordingly, agents are not able to employ processes such as best-reply to an existing network configuration. Rather, our model is *payoff based*. Agents can only measure derived benefits from past decisions of forming or severing links. Agents will reinforce a link if it was beneficial in the past and suppress it otherwise. These dynamics belong to the general class of learning automata [NT89, NP94] and are motivated by related models of human-like decision making [Art93].

The main difference with both [SP00] and [BL03] is in the reward function. In [SP00, BL03] the reward function is based on the principle of reciprocity which models social relationships such as friendship. Here instead we use the *connections model* of [JW96]. According to this model, the benefits received from each agent can be viewed as the information available from its direct and indirect links. In other words, agents are rewarded for being connected to other agents, either directly or indirectly. Additional features of our model are a *state-dependent* cost function for the establishment

and maintenance of links and a *distance-dependent* reward function for information benefits. This framework can model various economic and social contexts, such as the production and transmission of information, consumer products, etc. Models of this form (that also assume the consent of both parties) include the static model of [JW96] and the dynamic models of [Wat01, JW02a].

We will show that our model (i.e., the combined evolutionary process, reward functions, and cost functions) assigns positive probability to the emergence of multiple stable configurations (Nash networks). When the aforementioned *state-dependent* cost function is considered, we show that the set of strict Nash networks emerging may be larger than the one arising from best-reply dynamics considered in [BG00].

A specific case of our reward functions is "frictionless information flow", i.e., where benefits are derived from being connected to other agents and are not distance dependent. For this special case, we demonstrate the utility of an enhanced evolutionary process known as *dynamic reinforcement*. In particular, we will show how a single agent can reinforce the emergence of an efficient network through a simple "dynamic" processing of its own available information that uses the *rate* of observed reward changes [CS07]. This has the effect of reinforcing efficient networks while destabilizing the non-efficient networks.

5.4 The model

5.4.1 One-way benefit flow

Let $\mathcal{I} = \{1, ..., n\}$ denote a finite set of agents. The network relations among agents are represented by a graph, whose nodes are identified with the agents and whose edges capture the pairwise relations.

We will consider a *one-way* (directed) flow model, where a *network* $\mathcal{G}$ is defined as a collection of pairwise directed links, (i, s), $i, s \in \mathcal{I}$. More precisely, $\mathcal{G} \subseteq \{(i, s) :$

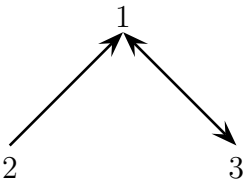


Figure 5.1: A network of three agents and one-way flow of benefits.

$i, s \in \mathcal{I}\}$. For example, the network $\mathcal{G} = \{(1,2),(1,3),(3,1)\}$ is illustrated in Fig. 5.1. In terms of the illustration, a link starts at s with the arrowhead pointing at i. This represents flow of benefits/information from s to i.

Define a *path* from s to i in $\mathcal{G}$, as $(i \leftarrow s) = \bigcup_{k=0}^{m-1}(s_{k+1}, s_k)$ for some positive integer, m, where $\{s_k\}_{k=0}^{m}$ is a sequence in $\mathcal{I}$ that satisfies $s_0 = s$, $s_m = i$, $s_k \neq s_{k+1}$ and $(s_{k+1}, s_k) \in \mathcal{G}$ for any $k = 0, 1, ..., m-1$.

Definition 5.4.1 (Connectivity) *A node $i \in \mathcal{I}$ is connected to a node $s \in \mathcal{I}\backslash i$ if there is a path from s to i. A network is connected if any $i \in \mathcal{I}$ is connected to any $s \in \mathcal{I}\backslash i$.*

We further assume that each agent is able to establish links only with "neighboring agents". The set of neighbors of agent i is denoted as $\mathcal{N}_i$ with cardinality $|\mathcal{N}_i|$. In the unconstrained neighbors case, $\mathcal{N}_i = \mathcal{I}\backslash i$.

5.4.2 The network formation model

We will model network formation as an evolutionary process, where at each stage agents decide which links to form. Based on agents' decisions, a graph is being formed, and a reward is assigned to each agent based on the information it receives through its links and its neighbors' links. In detail, the network formation model is described as follows.

5.4.2.1 Action space

The set of actions of agent i, denoted $\mathcal{A}_i$, contains all the possible combinations of neighbors with which a link can be established including the case of not establishing any link, i.e., $\mathcal{A}_i = 2^{\mathcal{N}_i}$.[8]

By enumerating the elements in $\mathcal{A}_i$, we can associate the j^{th} element of $\mathcal{A}_i$ with a vertex, e_j, of the probability simplex of dimension $|\mathcal{A}_i|$, i.e., $\Delta(|\mathcal{A}_i|)$. Accordingly, we will use the same notation, $\alpha_i \in \mathcal{A}_i$, to refer to an element of $\mathcal{A}_i$ either in terms of an index over $\mathcal{A}_i$, a vertex of $\Delta(|\mathcal{A}_i|)$, or an clement of $2^{\mathcal{N}_i}$. Finally, let $|\alpha_i|$ denote the cardinality of α_i viewed as an element of $2^{\mathcal{N}_i}$.

5.4.2.2 Learning algorithm

At each stage $k \in \mathbb{N}$, each agent i selects an action $\alpha_i(k) \in \mathcal{A}_i$ according to the probability distribution over $\mathcal{A}_i$

$$(1-\lambda)x_i(k) + \frac{\lambda}{|\mathcal{A}_i|}\mathbf{1},$$

where i) $x_i(k) \in \Delta(|\mathcal{A}_i|)$ is the *strategy* of agent i at stage k; ii) $\mathbf{1}$ is a vector of appropriate dimension with each element equal to 1; and iii) $\lambda \geq 0$ is a parameter used to model possible perturbations in the decision making process, also called *mutations* [KMR93, You93].

We assume that each agent i "learns" via a modified version of the perturbed learning algorithm $\tilde{L}^{\lambda}_{R-I}$, introduced in Section 3.8. This algorithm is written recursively as

$$x_i(k+1) = x_i(k) + \epsilon(k) \cdot (R_i(\alpha(k)) - C_i(\alpha_i(k), x_i(k))) \cdot (\alpha_i(k) - x_i(k)), \qquad (5.1)$$

[8]Note that $\varnothing \in 2^{\mathcal{N}_i}$, which corresponds to the case of establishing no link.

which is a stochastic recursion with step size sequence $\epsilon(k) \triangleq 1/(k+1)$.[9]

In the above recursion, $R_i : \mathcal{A} \to \mathbb{R}_+$ denotes the reward of agent i, which generally depends on the choices of all agents $\alpha = (\alpha_1, ..., \alpha_n)$ (i.e., the current network) defined in the product set $\mathcal{A} \triangleq \times_{i \in \mathcal{I}} \mathcal{A}_i$.

We will assume that the rewards are bounded and nonnegative, i.e., $0 \leq R_i(\cdot) < R_{\max} < \infty$, for some $R_{\max}$.

We also assume that the establishment and maintenance of a link is costly. In recursion (5.1), $C_i : \mathcal{A}_i \times \Delta(|\mathcal{A}_i|) \to \mathbb{R}_+$ denotes the cost of establishing and maintaining a link. This cost is assumed to depend on both the current and previously established links. Dependence of previously established links is implicit through the strategy $x_i(k)$.

It is important to recall that the dynamics of the strategy, x_i, is *payoff based*. That is, its evolution is determined by realized reward and cost function values.

5.4.2.3 Learning algorithm with dynamic reinforcement

Further insights into the possible emergence of efficient network structures can be derived by considering a dynamic processing of the local information available to each agent. In particular, agents might "value" a link's significance by also considering the recent reward changes provided by this link. That is, agents might be more satisfied with links that increased their available information in the *recent* history than with links that have provided large amounts of information throughout the *whole* history.

Based on similar reasoning, we will utilize a modified action selection probability distribution of the form

$$\Pi_\Delta\{(1-\lambda)[x_i(k) + \gamma_i \cdot (x_i(k) - y_i(k))] + \frac{\lambda}{|\mathcal{A}_i|}\mathbf{1}\}, \tag{5.2}$$

[9]The foregoing analysis holds for any step size of the form $\epsilon(k) = 1/(k^\nu + 1)$, where $\nu \in (1/2, 1]$.

where a new state variable, y_i, is updated according to the recursion

$$y_i(k+1) = y_i(k) + \epsilon(k) \cdot (x_i(k) - y_i(k)).$$

This reinforcement scheme is a special case of the more general dynamic reinforcement scheme introduced in Section 4.5.

5.4.3 Reward and cost function

The reward function will be defined as in the network formation models of [JW96, BG00, Wat01]. In those models, each agent is a source of benefits that others can tap via the formation of pairwise links. We suppose that a link with another agent allows access to the benefits available to the latter via its own links. In particular, we define:

$$R_i(\alpha) \triangleq \sum_{s \in \mathcal{I}, s \neq i} \delta^{d_{is}(\alpha)} \tag{5.3}$$

where i) $\delta \in (0, 1]$ measures the level of information decay and ii) $d_{ij} : \mathcal{A} \to \mathbb{N}$ is defined as the minimum distance from j to i given the current action profile $\alpha \in \mathcal{A}$. We adopt the convention that $d_{ij}(\cdot) = \infty$, when $(i, j) \notin \mathcal{G}$.

For each agent $i \in \mathcal{I}$, we define the cost function $C_i : \mathcal{A}_i \times \Delta(|\mathcal{A}_i|) \to \mathbb{R}_+$ to be:

$$C_i(\alpha_i, x_i) \triangleq \kappa_0 \, |\alpha_i| + \kappa_1 \varphi_i(\alpha_i)^{\mathrm{T}} (\mathbf{1} - \varphi_i(x_i)), \tag{5.4}$$

for some $\kappa_0, \kappa_1 \geq 0$. The parameter κ_0 corresponds to the cost of maintaining an existing link, while κ_1 corresponds to the cost of establishing a new link. The function $\varphi_i : \Delta(|\mathcal{A}_i|) \to \mathbb{R}^{|\mathcal{N}_i|}$ is defined by

$$[\varphi_i(x_i)]_j = \sum_{\{a \in \mathcal{A}_i : j \in a\}} x_{ia}.$$

By abuse of notation, we are using a as both an index, as in x_{ia}, and a set, as in $j \in a \in 2^{\mathcal{N}_i}$. In words, the $[\varphi_i(x_i)]_j$ denotes the probability that agent i will form a link to neighbor j based on the distribution x_i. The term $(\mathbf{1}-\varphi_i(x_i))^{\mathrm{T}}\varphi_i(\alpha_i)$ penalizes misalignment of the action α_i with the distribution x_i. In the perfectly aligned case, for any $\alpha_i \in \mathcal{A}_i$ (viewed as a vertex of $\Delta(|\mathcal{A}_i|)$),

$$\varphi_i(\alpha_i)^{\mathrm{T}}(\mathbf{1} - \varphi_i(\alpha_i)) = 0$$

whereas in the worst case,

$$\max_{x_i} \varphi_i(\alpha_i)^{\mathrm{T}}(\mathbf{1} - \varphi_i(x_i)) = |\alpha_i| \,.$$

We make the following assumptions for the ***remainder of the chapter***:

Assumption 5.4.1 $0 \leq \kappa_0 + \kappa_1 < \delta$.

This assumption assures that

$$\max_{\alpha_i \in \mathcal{A}_i} R_i(\alpha_i, \alpha_{-i}) - C_i(\alpha_i, x_i) > 0$$

for all $\alpha_{-i} \in \times_{s \neq i} \Delta(|\mathcal{A}_s|))$ and $x_i \in \Delta(|\mathcal{A}_i|)$. In particular, agents always have an incentive to form at least one link.

Assumption 5.4.2 *The neighbor sets* $\{\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_n\}$ *are such that a connected network is feasible.*

5.4.4 Efficiency

Having stated the general properties of the reward and cost functions, we also need to characterize the *efficiency* of a network structure. To this end, we borrow the definition of the *value* of a network from [JW96].

First, define the agent utility function $v_i : \mathcal{A} \times \Delta(\mathcal{A}_i) \to \mathbb{R}_+$ as

$$v_i(\alpha, x_i) = R_i(\alpha) - C_i(\alpha_i, x_i), \tag{5.5}$$

i.e., the combined reward minus cost in the update equation (5.1). Note that unlike typical utility functions in network formation games, this utility function depends explicitly on both collective actions, α, and an agent's strategy, x_i. In the special case where $x_i = \alpha_i$, the cost term only reflects maintenance costs (i.e., the κ_0 term), whereas establishment costs (the κ_1 term) are zero.

Definition 5.4.2 (Network value) *The value of the network $V : \mathcal{A} \to \mathbb{R}_+$, is the sum of agent rewards minus maintenance costs at an action profile, $\alpha \in \mathcal{A}$, i.e.,*

$$V(\alpha) = \sum_{i \in \mathcal{I}} v_i(\alpha, \alpha_i). \tag{5.6}$$

Definition 5.4.3 (Efficient network) *An efficient network is a joint action profile $\alpha \in \mathcal{A}$ with the maximum value.*

The following is a direct consequence of the Definition 5.4.3:

Claim 5.4.1 *An efficient network is connected. In the special case of $\delta = 1$, an efficient network is a connected network with a minimal number of links.*

Proof. Assume that an efficient network, say $\mathcal{G}$, is not connected. Then there exist $i, j \in \mathcal{I}$ such that $j \in \mathcal{N}_i$ and $(i \leftarrow j) \notin \mathcal{G}$. Adding (i, j) to $\mathcal{G}$ increases the reward of agent i, since $\delta > \kappa_0 + \kappa_1$, and therefore it increases the value of the network. Thus, the network $\mathcal{G}$ is not efficient, which contradicts our initial assumption. Thus, any efficient network is a connected network.

In the special case of $\delta = 1$, the value of any connected network is

$$V(\alpha) = n(n-1) - \kappa_0 \sum_{i=1}^{n} |\alpha_i|$$

If $\mathcal{A}_c$ denote the set of connected networks, any efficient network, say α^*, satisfies

$$\alpha^* \in \arg\max_{\alpha \in \mathcal{A}_c} V(\alpha)$$

or, equivalently,

$$\alpha^* \in \arg\min_{\alpha \in \mathcal{A}_c} \sum_{i=1}^{n} |\alpha_i|,$$

which implies that an efficient network has a minimal number of links. □

5.5 Stability analysis

5.5.1 Asymptotic stability analysis

The asymptotic convergence properties of the stochastic recursion (5.1) with diminishing step size was described in Section 3.8.3. In this framework, we showed by Proposition 3.8.3 that the reinforcement scheme converges to an invariant set of the set of ordinary differential equations:

$$\dot{x}_i = \overline{g}_i(x) \triangleq \overline{r}_i(x) - \overline{R}_i(x) \cdot x_i,$$

where

$$\overline{r}_i(x) \triangleq E[R_i(\alpha(k))\alpha_i(k)|x(k) = x] \in \mathbb{R}_+^{|\mathcal{A}_i|},$$

$$\overline{R}_i(x) \triangleq E[R_i(\alpha(k))|x(k) = x] \in \mathbb{R}_+.$$

The above set of ODE's can be written more compactly as

$$\dot{x} = \overline{g}(x) \triangleq \text{col}\{\overline{g}_i(x)\}_{i\in\mathcal{I}}, \tag{5.7}$$

where col$\{A\}$ denote the column vector of the elements of a finite set A.

Moreover, again according to Proposition 3.8.3, there is a positive probability that the reinforcement scheme converges to a locally stable set (in the sense of Lyapunov) of the ODE (5.7). It was also shown by Proposition 3.8.4 that there is probability zero that the reinforcement scheme will converge to a linearly unstable point of the ODE (5.7).

In this section, we are going to analyze the local stability properties of the stationary points of the ODE (5.7), since stationary points are invariant sets. This way, we can derive conclusions regarding convergence of the stochastic recursion (5.1).

5.5.2 Stationary points

It has been shown by Proposition 4.4.1 that for $\lambda = 0$, any pure strategy profile $\alpha^* = (\alpha_1^*, ..., \alpha_n^*)$ is a stationary point of the stochastic recursion (5.1).

Moreover, by Proposition 4.4.3, for sufficiently small $\lambda > 0$, there exists a unique continuously differentiable function $w^* : \mathbb{R}_+ \to \times_i \mathbb{R}^{|\mathcal{A}_i|}$, such that $\lim_{\lambda\to 0} \lambda w^*(\lambda) = 0$, and

$$x^* = \alpha^* + \lambda w^*(\lambda) \tag{5.8}$$

is a stationary point of the ODE (5.7).

5.5.3 Local asymptotic stability (LAS)

Having characterized the stationary points of the stochastic recursion (5.1), we will describe locally the stability properties of these points. To this end, we first need to

define the *conditional expected utility* $\overline{v}_i(\alpha_i, x)$ as the expected utility of agent i given that it selects action α_i, incurs establishment costs according to x_i, and other agents select their actions according to the probability distributions x_{-i}, i.e.,

$$\overline{v}_i(\alpha_i, x) = E\{v_i(\alpha, x_i)\big|\alpha_i, x_{-i}\},$$

where $v_i(\cdot, \cdot)$ is defined in (5.5).

Proposition 5.5.1 (LAS of Standard Reinforcement) *For sufficiently small $\lambda > 0$, let x^* be a stationary point of the ODE (5.7) corresponding to some $\alpha^* \in \mathcal{A}$ according to (5.8). The stationary point x^* is a locally asymptotically stable point of the ODE (5.7) for sufficiently small $\lambda > 0$ if and only if, for each $i \in \mathcal{I}$,*

$$\overline{v}_i(\alpha_i^*, x^*) > \overline{v}_i(\alpha_i', x^*) \tag{5.9}$$

for all $\alpha_i' \in \mathcal{A}_i \backslash \alpha_i^*$.

Proof. The proof follows similar reasoning as Proposition 4.4.4. □

In the case of the dynamic reinforcement scheme of (5.2), the relevant ODE is now

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \overline{g}(x, y) \\ x - y \end{pmatrix}, \tag{5.10}$$

and the condition for stability takes the following form:

Proposition 5.5.2 (LAS of Dynamic Reinforcement) *Assume the hypotheses of Proposition 5.5.1 under stability condition (5.9). Assume that each agent i applies dynamic reinforcement (5.2) for some $\gamma_i > 0$. The strategy profile x^* is a locally asymptotically stable stationary point of the ODE (5.10) for sufficiently small $\lambda > 0$*

if and only if, for each agent $i \in \mathcal{I}$, the derivative feedback coefficient satisfies

$$0 \leq \gamma_i < \frac{\overline{v}_i(\alpha_i^*, x^*) - \overline{v}_i(\alpha_i', x^*) + 1}{\overline{v}_i(\alpha_i', x^*)} \tag{5.11}$$

for all $\alpha_i' \in \mathcal{A}_i \backslash \alpha_i^*$.

Proof. The proof follows similar reasoning as Theorem 4.5.1. □

5.6 Nash networks

In the literature of network formation, where each agent performs a best-reply at the current network structure, Nash equilibria are usually called *Nash networks*, [BG00], and correspond to the case where there is no agent that can unilaterally benefit from establishing a new link or severing an existing link. In the framework of our network formation model, where decisions are state-dependent, we define:

Definition 5.6.1 (Nash network) *An action profile $\alpha^* \in \mathcal{A}$ is a Nash network if and only if*

$$v_i((\alpha_i^*, \alpha_{-i}^*), \alpha_i^*) \geq v_i((\alpha_i', \alpha_{-i}^*), \alpha_i^*), \tag{5.12}$$

for all $\alpha_i' \in \mathcal{A}_i \backslash \alpha_i^*$ and $i \in \mathcal{I}$. Likewise, a strict Nash network satisfies the strict inequality in (5.12).

Claim 5.6.1 *Nash networks are connected.*

Proof. Assume that a Nash network, say $\mathcal{G}$, is not connected. Then there exist $i, j \in \mathcal{I}$ such that $j \in \mathcal{N}_i$ and $(i \leftarrow j) \notin \mathcal{G}$. Adding (i, j) to $\mathcal{G}$ increases the reward of agent i, since $\delta > \kappa_0 + \kappa_1$, and therefore it increases the reward of agent i. Thus, the network $\mathcal{G}$ is not a Nash network, which contradicts our initial assumption. Thus,

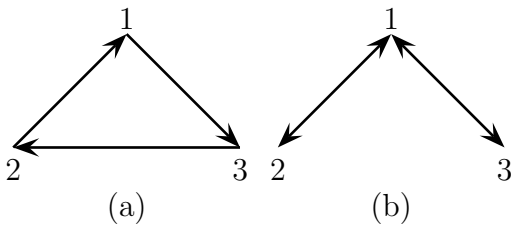


Figure 5.2: Nash equilibria in case of the *connections model* of [JW96].

any Nash network is a connected network. □

The Nash networks for $n = 3$ agents and no decay are shown in Fig. 5.2. For example, in Fig. 5.2(a), assuming that $\kappa_0 > 0$ and $\kappa_1 = 0$, all agents realize the same utility, which is equal to $2 - \kappa_0$. This is a strict Nash network since each agent can only be worse off by unilaterally changing its links. Likewise, in Fig. 5.2(b), agents 2 and 3 realize utility $2 - \kappa_0$, while agent 1 realizes $2 - 2\kappa_0$. Moreover, agent 2 is indifferent between creating a link with agent 1 or agent 3, since both links provide $2 - \kappa_0$. Similarly, agent 3 is indifferent between creating a link with agent 1 or agent 2. Instead, agent 1 can only decrease its utility by changing its strategy. Hence, for the case of $\kappa_1 = 0$, network (b) is *not* a strict Nash network.

However, in case $\kappa_1 > 0$, both Nash networks in Fig. 5.2 are strict, since each deviation from the equilibrium play is charged by an extra cost of order κ_1. For example, in Fig. 5.2(b), agent 2 is no longer indifferent between creating a link with agent 1 or agent 3.

According to the definition of a Nash network and local stability analysis of Proposition 5.5.1, we conclude that:

Proposition 5.6.1 *Under the hypotheses of Proposition 5.5.1, a stationary point $x^* = \alpha^* + \lambda w^*(\lambda)$, such that α^* is a strict Nash network, is a locally asymptotically stable point of the ODE (5.7) for sufficiently small $\lambda > 0$.*

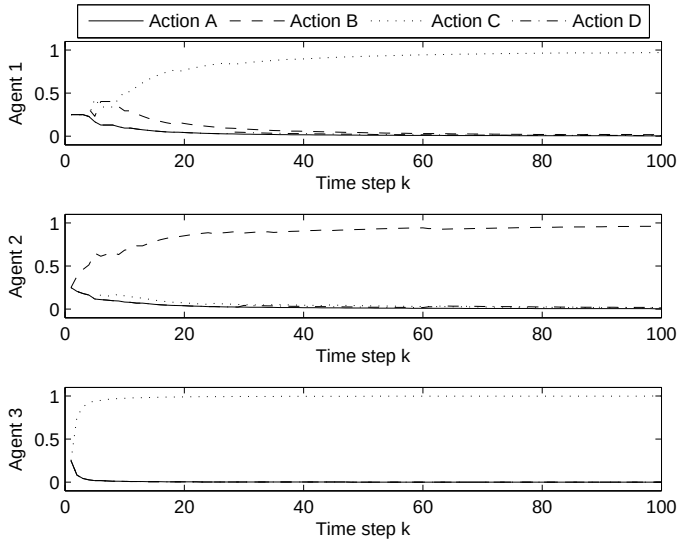


Figure 5.3: A typical response of the stochastic iteration (5.1), for $\delta = 1$, $\kappa = 1/2$, $\kappa_1 = 0$, $\lambda = 0.01$. Convergence to the efficient formation of Fig. 5.2(a) is observed.

Therefore, finding the set of strict Nash networks α^* reveals the set of stationary points x^* that are locally stable.

Note that according to Propositions 3.8.3–3.8.4, convergence to non-strict Nash network need not be excluded.[10] Figs. 5.3–5.4, simulate two characteristic responses of the stochastic recursion (5.1) where we consider the following action spaces $\mathcal{A}_1 = \{\varnothing, \{2\}, \{3\}, \{2,3\}\}$, $\mathcal{A}_2 = \{\varnothing, \{1\}, \{3\}, \{1,3\}\}$, $\mathcal{A}_3 = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$, denoted by $\mathcal{A}_i = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$, $i = 1, 2, 3$. In Fig. 5.3, the recursion converges to the efficient formation of Fig. 5.2(a), while in Fig. 5.4 the recursion converges to the non-efficient network of Fig. 5.2(b).

5.6.1 Frictionless benefit flow ($\delta = 1$)

In order to characterize the Nash networks in the general case of $n > 3$ agents, we need to define a general class of networks called *critically linked networks*.

[10] By Proposition 5.5.1, the linearization of the ODE (5.7) about a Nash network may exhibit some zero eigenvalues.

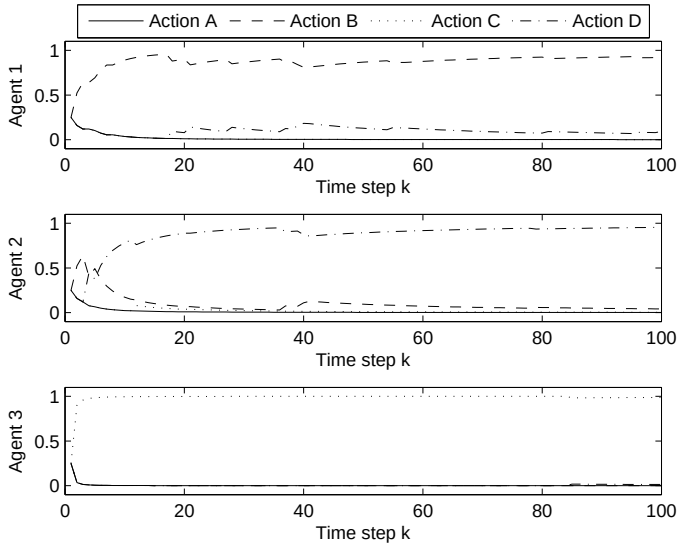


Figure 5.4: A typical response of the stochastic recursion (5.1), for $\delta = 1$, $\kappa = 1/2$, $\kappa_1 = 0$, $\lambda = 0.01$. Convergence to the non-efficient formation of Fig. 5.2(b) is observed.

Definition 5.6.2 (Critically linked network) *A network, $\mathcal{G}$, is critically linked if i) it is connected and ii) for all $(i,j) \in \mathcal{G}$, the unique path $(i \leftarrow j)$ is (i,j).*

In words, a critically linked network is such that if agent i drops a direct link to (neighboring) agent j, then i loses connectivity to j by any means.

Proposition 5.6.2 (Nash networks) *For $\delta = 1$, $n > 2$, and $\kappa_0, \kappa_1 > 0$, a network is a strict Nash network if and only if it is a critically linked network.*

Proof. (Critically linked $\Rightarrow$ Strict Nash) Let $\alpha^* \in \mathcal{A}$ correspond to a critically linked network, $\mathcal{G}^*$. Suppose for some agent $i \in \mathcal{I}$ and some action $\alpha_i' \in \mathcal{A}_i$, $\alpha_i' \neq \alpha_i^*$,

$$v_i((\alpha_i', \alpha_{-i}^*), \alpha_i^*) \geq v_i((\alpha_i^*, \alpha_{-i}^*), \alpha_i^*), \tag{5.13}$$

i.e., agent i's utility of α_i' is at least that of α_i^*. Denote the resulting network by $\mathcal{G}'$.

The distinction between α^* and α' lies in the set of neighbors in $\mathcal{N}_i$ that were dropped, added, or kept. Denote these sets of neighbors by $N_{\text{drop}}, N_{\text{keep}}, N_{\text{add}} \subset \mathcal{N}_i$,

respectively. Accordingly, we can identify

$$\alpha^* = N_{\text{keep}} \cup N_{\text{drop}} \quad \& \quad \alpha' = N_{\text{keep}} \cup N_{\text{add}}.$$

Clearly if $N_{\text{drop}} = \varnothing$, then (5.13) cannot hold.

Assume for now that $\mathcal{G}'$ is connected. We will revisit this assumption later. Since $\mathcal{G}^*$ and $\mathcal{G}'$ are connected, the derived benefits in both cases equals $n-1$. Since the establishment coefficient κ_1 is strictly positive, the only possibility for (5.13) to hold is if $|N_{\text{add}}| < |N_{\text{drop}}|$. That is, the maintenance cost is strictly less using α', for if the maintenance costs were equal, the establishment cost would result in (5.13) being violated.

We now show that $|N_{\text{add}}| < |N_{\text{drop}}|$ contradicts $\mathcal{G}^*$ being a critically linked network.

- For each element of N_{add}, construct a path without loops in $\mathcal{G}^*$ to i. These paths must pass through $N_{\text{keep}} \cup N_{\text{drop}}$.
- Since $|N_{\text{drop}}| > |N_{\text{add}}|$, there exists a $k^* \in N_{\text{drop}}$ that is not part of any of these paths.
- Construct a path in $\mathcal{G}^*$ from k^* to any element in N_{add}. Since $\mathcal{G}'$ is connected and from the critically connected assumption, it is possible to construct such a path that does not pass through agent i.
- The conclusion is a path from k^* to an element of N_{add} to an element of $\alpha^* \backslash k^*$. This path contradicts the critically linked assumption, since the existence of this path implies that k^* could be dropped in $\mathcal{G}^*$ without loss of connectivity to agent i.

Returning to the assumption that $\mathcal{G}'$ is connected, if this were not the case, then adding a link to an appropriate element of N_{drop} results in an increased utility. We can repeat this process, each time relabeling α' and $\mathcal{G}'$ until $\mathcal{G}'$ is connected.

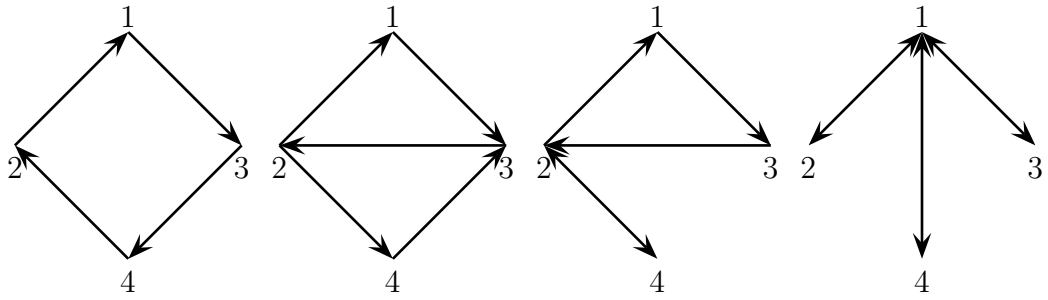


Figure 5.5: Flower networks in case of $n = 4$.

(Strict Nash $\Rightarrow$ Critically linked) Suppose a Nash network is not critically linked. Then there exists an agent i that can drop a direct link to an agent $j \in \mathcal{N}_i$ but still maintain connectivity to j, and hence receive the benefits of j without incurring the maintenance cost of j. Therefore, the original network cannot be a Nash network, which is a contradiction. $\square$

A special class of critically linked networks are so-called *flower networks*, defined in [BG00]. For example, Figs. 5.2 and 5.5 show all possible flower networks for the case of $n = 3$ and $n = 4$ agents, respectively. Contrary to prior work on best-reply dynamics, here we see that the introduction of a dynamic establishment cost function was able to justify the emergence of flower networks. Under best-reply dynamics, convergence to the entire family of flower networks is not possible [BG00].[11]

In the special case of i) no establishment cost ($\kappa_1 = 0$) and ii) unconstrained neighbors ($\mathcal{N}_i = \mathcal{I} \backslash i$), we can have a more explicit characterization of strict Nash networks. We first define the following.

Definition 5.6.3 (Wheel network) *A wheel network is a connected network uniquely*

[11] In [BG00], the introduction of a discount factor into the reward function resulted in the emergence of multiple Nash equilibria, the family of which includes the flower networks. Although, in [BG00], flower networks are defined as the union of distinct wheel sub-networks with only one common node (i.e., it is a subset of the set of flower networks defined in this paper), there is a specific sub-class of flower networks (the ones with two petals where one of them is a spoke) that cannot be strict Nash networks.

defined by a path $(i \leftarrow i)$ *for some* $i \in \mathcal{I}$ *where every agent in* $\mathcal{I}$ *is visited only once.*

Proposition 5.6.3 *For* $\delta = 1$*,* $n > 2$*, and* $1 > \kappa_0 > \kappa_1 = 0$*, the wheel network is the unique strict Nash network.*

Proof. It is straightforward to check that the wheel network is a strict Nash network. Let us consider any other (necessarily connected) Nash network, $\mathcal{G}$. If it is not a wheel network, there exists an agent that is a common link between two other agents, i.e., $(j, i), (k, i) \in \mathcal{G}$ for some $i, j, k \in \mathcal{I}$. This cannot be a strict Nash network, since agent j is indifferent between connecting to i or connecting to k. □

5.6.2 Decaying benefit flow ($\delta < 1$)

In this section, we analyze the case where the information flow is also subject to decay. We will see that a network being a Nash equilibrium imposes a structural constraint on the distances between neighboring agents.

Proposition 5.6.4 (Nash networks with decay) *Let* $0 < \delta < 1$*,* $n > 2$*,* $\kappa_0 > 0$*, and* $\kappa_1 \geq 0$*. Let* $\mathcal{G}$ *be a Nash network corresponding to the joint action* $\alpha \in \mathcal{A}$*. For any agent* i*, if* $|\alpha_i| < |\mathcal{N}_i|$*, then*

$$\delta - \delta^{d_{ij}(\alpha)} \leq \kappa_0 + \kappa_1 \text{ for all } j \in \mathcal{N}_i.$$

The condition $|\alpha_i| < |\mathcal{N}_i|$ means that agent i is not using all of its available links. The inequality (5.14) is revealing only for neighbors of i for which there is not a direct link. This could be of interest, for example, in the unconstrained neighbors case with a large number of agents. **Proof.** Let α_i^* satisfy the assumptions of Proposition 5.6.4, and compare an alternative action $\alpha_i' \in \mathcal{A}_i$ that consists of adding a direct link to

neighbor j, i.e., $\alpha_i' = \alpha_i \cup \{j\}$. The resulting utility to agent i can be bounded by

$$v_i((\alpha_i', \alpha_{-i}), \alpha_i) \geq v_i((\alpha_i, \alpha_{-i}), \alpha_i) + (\delta - \delta^{d_{ij}(\alpha)}) - (\kappa_0 + \kappa_1).$$

That is, the consequence of adding a link to j shortens the distances to other links; adds the direct benefit of a link to j; loses the indirect benefit of a link to j; incurs additional maintenance cost; and incurs additional establishment cost. Therefore, if

$$(\delta - \delta^{d_{ij}(\alpha)}) - (\kappa_0 + \kappa_1) > 0, \tag{5.14}$$

there is an incentive to add a link to j, and so α cannot be a Nash network. Conversely, asserting that α is a Nash network implies the desired result. □

This theorem can be used to bound distances to neighbors as follows. Inequality (5.14) is equivalent to

$$d_{ij}(\alpha) \leq \frac{\log(\delta - (\kappa_0 + \kappa_1))}{\log(\delta)}.$$

A sufficient condition to bound the distance to neighbors by $d_{\max}$ is then

$$\kappa_0 + \kappa_1 \leq \delta - \delta^{d_{\max}}.$$

Let us assume for example that

$$\delta - \delta^2 < \kappa_0 + \kappa_1 < \delta - \delta^3.$$

It is straightforward to check that the networks shown in Fig. 5.6 are Nash networks. We observe that a distance of 3 among any two agents is not supported in any of these networks.

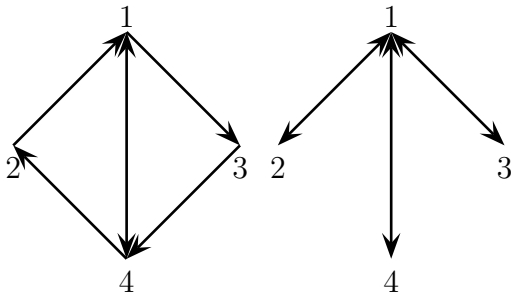


Figure 5.6: Two Nash networks in case of $n = 4$ agents and $\delta - \delta^2 < \kappa_0 + \kappa_1 < \delta - \delta^3$.

5.7 Dynamic reinforcement

In Chapter 4 it was shown that a dynamic reinforcement scheme of the form of (5.2) can destabilize all non-efficient Nash equilibria in coordination games. We wish to answer the question of whether such a reinforcement scheme can be used for distributed reinforcement to desirable networks.

Intuitively, the dynamic reinforcement scheme of (5.2) reinforces *changes* in strategy. Following the language of Propositions 3.8.3–3.8.4, let x^* be a joint equilibrium strategy associated with some joint action $\alpha^* \in \mathcal{A}$ for sufficiently small λ. Dynamic reinforcement effectively skews the perceived payoff benefits of unilateral action deviations. For example, suppose α_i' is an alternative action for agent i. Under dynamic reinforcement, the perceived benefit of a deviation is

$$(1 + \gamma_i)\overline{v}_i(\alpha_i', x^*) - (\overline{v}_i(\alpha_i^*, x^*) + 1),$$

as opposed to the actual benefit in the absence of dynamic reinforcement, which is

$$\overline{v}_i(\alpha_i', x^*) - \overline{v}_i(\alpha_i^*, x^*).$$

If α^* corresponds to a Nash equilibrium strategy, the actual deviation benefit will be negative for all alternatives, α_i'. Under dynamic reinforcement, the perceived benefit

can be positive and induce a departure for that agent from the action α_i^*.

This departure can, in turn, induce other agents to abandon their Nash equilibrium actions. For example, in Fig. 5.2(b), if agent 2 is able to support dropping a link with agent 1 in favor of a link with agent 3, then agent 1's best reply will be to drop the link with agent 3 maintaining only the link with agent 2, which gives rise to the efficient wheel formation of Fig. 5.2(a).

On a cautionary note, excessive dynamic reinforcement can induce deviations from all Nash equilibria. The key to evoking efficient outcomes lies in finding the correct level of dynamic reinforcement, as measure by the coefficients γ_i, to induce deviations from non-efficient equilibria while maintaining stability of efficient equilibria. The following claims explicitly carry out this procedure for a special case of network formation.

Claim 5.7.1 *Assume that for each $i \in \mathcal{I}$, $\mathcal{N}_i = \mathcal{I} \backslash i$. Let $\delta = 1$, $n > 2$, and $\kappa_0, \kappa_1 \geq 0$. Let x^{non} be a stationary point corresponding to an non-efficient Nash network configuration, α^{non}, according to (5.8) for sufficiently small $\lambda > 0$. There exists an agent i and constant $\gamma^{\text{non}} > 0$ such that if agent i applies the dynamic reinforcement scheme of (5.2) with coefficient $\gamma_i > \gamma^{\text{non}}$, then the non-efficient equilibrium formation, x^{non}, is linearly unstable point for (5.10).*

Proof. For any non-efficient configuration (in this case, anything other than the wheel network) by following the proof of Proposition 5.6.3, we can identify an agent i that would be indifferent between its current action, α_i^{non}, and an alternative, α_i', if κ_1 were equal to 0. We will use Proposition 5.5.2 to compute a destabilizing level of γ_i for this agent (regardless of the value of κ_1). For this agent, under the assumptions of Claim 5.7.1, $\overline{v}_i(\alpha_i^{\text{non}}, x^{\text{non}}) = (n-1) - \kappa_0 + O(\lambda)$. If agent i deviates to α_i', its expected utility becomes $\overline{v}_i(\alpha_i', x^{\text{non}}) = (n-1) - \kappa_0 - \kappa_1 + O(\lambda)$. Applying Proposition 5.5.2

and setting

$$\gamma^{\text{non}} = \frac{1+\kappa_1}{(n-1)-(\kappa_0+\kappa_1)}$$

gives the desired result. □

Claim 5.7.1 shows that there exists an agent who is able to destabilize a non-efficient network. The process could get attracted to another steady-state configuration that is not efficient. However, if each agent $i \in \mathcal{I}$ applies derivative action with $\gamma_i > \gamma^{\text{non}}$, then all non-efficient networks will be linearly unstable.

The following claim computes an upper bound on the γ_i so that stability of the efficient (wheel) network is maintained.

Claim 5.7.2 *Assume that for each $i \in \mathcal{I}$, $\mathcal{N}_i = \mathcal{I}\backslash i$. Let $\delta = 1$, $n > 2$, and $\kappa_0, \kappa_1 \geq 0$. Let x^{eff} be a stationary point corresponding to the efficient Nash network wheel configuration, α^{eff}, according to (5.8) for sufficiently small $\lambda > 0$. There exists a $\gamma^{\text{eff}} > \gamma^{\text{non}}$ such that if any agent i applies the dynamic reinforcement scheme of (5.2) with coefficient γ_i, $0 < \gamma_i < \gamma^{\text{eff}}$, then x^{eff} is a locally asymptotically stable equilibrium for (5.10).*

Proof. Again, we will use Proposition 5.5.2 to compute an upper bound for γ^{eff}. Assume that agents are currently playing the efficient equilibrium configuration with associated strategy x^{eff}. Consider any agent $i \in \mathcal{I}$, who is currently playing the corresponding equilibrium action α_i^{eff}. Agent i realizes utility $\overline{v}_i(\alpha_i^{\text{eff}}, x^{\text{eff}}) = (n-1) - \kappa_0 + O(\lambda)$, since the network is connected (by Proposition 5.6.1) and each agent maintains only one link (by Proposition 5.4.1).

Assume now that agent i deviates by selecting a different action $\alpha_i' \neq \alpha_i^*$, such that $\overline{v}_i(\alpha_i', x^{\text{eff}})$ is as large as possible. This deviation corresponds to the tightest upper bound in Proposition 5.5.2. Since $\kappa_0 < 1$, the desired action α_i' corresponds to the case of establishing two links, one of which is the link of α_i^{eff} and the other link

arbitrary. In this case, $\overline{v}_i(\alpha_i', x^{\text{eff}}) = (n-1) - 2\kappa_0 - \kappa_1 + O(\lambda)$, and by Proposition 5.5.2, for any $\gamma_i > 0$ such that

$$\gamma_i < (1 + \kappa_0 + \kappa_1)/((n-1) - 2\kappa_0 - \kappa_1) + O(\lambda) = \gamma^{\text{eff}},$$

the efficient equilibrium x^{eff} is locally asymptotically stable for (5.10). Since

$$\frac{1+\kappa_1}{(n-1) - 2\kappa_0 - \kappa_1} > \frac{1+\kappa_1}{(n-1) - \kappa_0 - \kappa_1},$$

we conclude that $\gamma^{\text{eff}} > \gamma^{\text{non}}$ for small λ. □

Based on the previous claims, we are ready to describe convergence and non-convergence properties of the dynamic reinforcement scheme in the case of frictionless benefit flow.

Proposition 5.7.1 *In the framework of Claims 5.7.1–5.7.2, if* $\gamma_i \in [\gamma^{\text{non}}, \gamma^{\text{eff}})$ *for all* $i \in \mathcal{I}$, *then* $\text{Prob}\{\lim_{k\to\infty} x(k) = x^{\text{eff}}\} > 0$, *and* $\text{Prob}\{\lim_{k\to\infty} x(k) = x^{\text{non}}\} = 0$.

Proof. This is a direct consequence of Propositions 3.8.3–3.8.4. □

For example, let us consider the case of $n = 3$ agents and $\kappa_0 = 1/2$, $\kappa_1 = 0$, $\lambda = 0.01$ and $\delta = 1$. According to Claims 5.7.1–5.7.2, $\gamma^{\text{non}} = 2/3$ and $\gamma^{\text{eff}} = 3/2$. In Fig. 5.7 we have simulated the stochastic recursion (5.1) with initial conditions that are close to the non-efficient network of Fig. 5.2(b) when all agents apply the dynamic reinforcement scheme of (5.2) with $\gamma = 1$. Note that since all agents apply dynamic reinforcement with $\gamma \in [\gamma^{\text{non}}, \gamma^{\text{eff}})$, according to Proposition 5.7.1 the non-efficient network Fig. 5.2(b) will be linearly unstable.[12] We observe that deviation

[12] In fact, it is only sufficient either agent 2 or 3 to apply dynamic reinforcement in order for the non-efficient network to be destabilized, according to Claim 5.7.1.

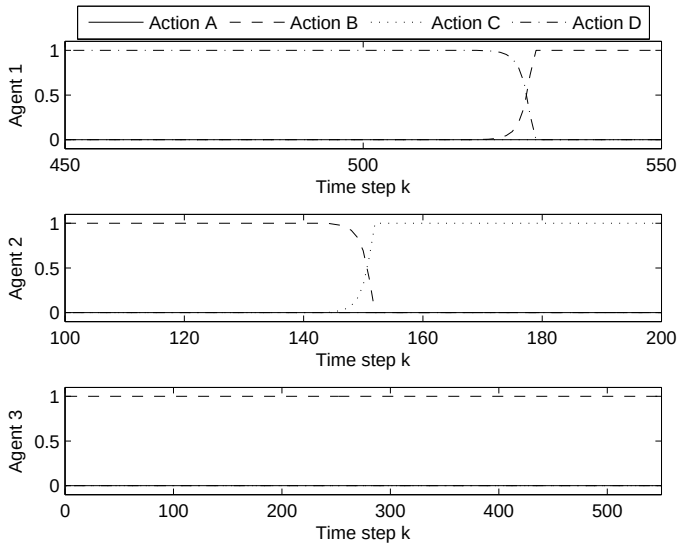


Figure 5.7: A typical response of the stochastic iteration (5.1), for $\delta = 1$, $\kappa_0 = 1/2$, $\kappa_1 = 0$, $\lambda = 0.01$ when all agents apply dynamic reinforcement with $\gamma = 1$ and for an initial condition that is close to the non-efficient formation of Fig. 5.2(b).

from the non-efficient network is achieved and the process converges to the efficient configuration.

5.8 Application: Topology control of ad-hoc wireless sensor networks

5.8.1 Motivation

Recent advances in wireless communications and electronics have enabled the developments of low-cost, low-power, multifunctional sensor nodes that are small in size and communicate untethered in short distances. These tiny sensor nodes, which consist of sensing, data processing, and communicating components, leverage the idea of sensor networks [ASS02].

The position of sensor nodes need not be engineered or predetermined. This allows random deployment in inaccessible terrains or disaster relief operations. On the

other hand, this also means that *sensor network protocols must possess self-organizing capabilities.*

Some of the application areas are health, military, and home. In military, for example, the rapid deployment, self-organization, and fault tolerance characteristics of sensor networks make them a very promising sensing technique for military command, control, communications, computing, intelligence. In health, sensor nodes can also be deployed to monitor patients and assist disabled patients. Some other commercial applications include managing inventory, monitoring product quality and monitoring disaster areas.

Realization of these and other sensor network applications require wireless ad hoc networking techniques. *Although many protocols and algorithms have been proposed for traditional wireless ad-hoc networks, they are not well suited to the unique features and application requirements of sensor nodes.* To illustrate this point the differences between sensor networks and ad hoc networks are:

- The number of sensor nodes in a sensor network can be several orders of magnitude higher than the nodes in an ad hoc network.
- Sensor nodes are densely deployed.
- Sensor nodes are prone to failures.
- The topology of sensor nodes changes very frequently.
- Sensor nodes mainly use a broadcast communication paradigm, whereas most ad hoc networks are based on point-to-point communications.
- Sensor nodes are limited in power, computational capacities, and memory.
- Sensor nodes may not have global *identification* (ID) because of the large amount of overhead and large number of sensors.

We are going to introduce some basic ideas regarding the current research issues in this emerging field. We also attempt an investigation into pertaining design constraints and outline the use of certain tools to meet the design objectives.

5.8.2 Sensor networks: Communication architecture

The sensor nodes are usually scattered in a *sensor field*. Each of these scattered sensor nodes has the capabilities to collect data and route data back to the *sink*. Data are routed back to the sink by a multihop infrastructureless architecture through the sink. The sink may communicate with the task manager node via Internet or Satellite. The design of the sensor network is influenced by many factors, including

- fault tolerance (nodes might fail, and such failures should not affect the overall task of the network),
- scalability (nodes may be thousands, and network should be able to work well),
- network topology (topology might change over time, or additional nodes might be added),
- power consumption (the wireless sensors have a limited power source. The malfunctioning of a few nodes can cause significant topological changes and might require rerouting of packets and reorganization of the network. At the same time, large communication distances are costly. Therefore, power conservation and power management is quite important).

Of course, there are also several challenging hardware design problems, which we will not be considered here. More information about these issues can be found at [ASS02] and the references therein.

5.8.3 The protocol hierarchy

Sensors communicate and execute tasks through a protocol hierarchy (starting from the lowest level) which includes the:

1. physical layer,
2. data link layer,
3. network layer,
4. transport layer,
5. application layer.

The physical layer includes the modulation, transmission and receiving techniques. The data link layer establishes the communication rules among nodes (usually executed by a medium access control (MAC) protocol) and minimizes collision among neighboring nodes' broadcasts. The network layer takes care of routing the data supplied by the transport layer. The transport layer helps helps to maintain the flow of data if the sensor networks application requires it. Finally, depending on the sensing tasks, different types of application software can be built and used on the application layer.

Also, there might be other protocols (usually called *planes*) that control power, movement and task distribution among nodes that may affect each one of the above layers. These planes help the sensor nodes coordinate the sensing task and lower overall power consumption.

In this section, we will deal with design questions posed about the *data link layer*.

5.8.4 The data link layer

The data link layer is responsible for enabling the communications among nodes in the network. In particular, a link-layer infrastructure needs to be established and also channel access needs to be regulated among the nodes. This process is captured by the Medium Access Control (MAC) protocol.

Several MAC schemes have already been developed for other wireless networks (e.g., cellular systems, bluetooth, etc). In cellular systems power conservation is of secondary importance, while every mobile node is only a hop away from the nearest base station. Therefore, MAC protocol in cellular systems is only assigned a resource allocation task. However, in sensor networks there is no base station and such a scheme will not be useful. Bluetooth and mobile ad-hoc networks are probably closer to the design question in sensor networks.

The Bluetooth topology is a star network where a master node can have up to seven slave nodes wirelessly connected to it. The MAC protocol in a mobile ad-hoc network has the task of forming the infrastructure and maintain it in the face of mobility. Power consumption is again of secondary importance. On the other hand, a sensor network comprises of a much larger number of nodes. Also topology changes might be more frequent and can be attributed to both node mobility and failure. Therefore, new MAC protocols need to be developed for sensor networks.

More specifically, the methods for channel access in the existing ad-hoc networks is done by two different methods: *contention* or *explicit organization* in time or frequency or code domains. The MAC-layer design for 802.11 is an example of the former method. The contention-based channel access scheme is not suitable for sensor networks, due to their requirement for radio transceivers to monitor the channel at all times, which is considered quite expensive for the low radio ranges of sensor networks [SGA00]. In sensor networks instead, it will be useful to turn off the radio when there is no need for communication. The *organized* channel access [BE81], attempts to

determine network connectivity first (i.e., discover the radio neighbors of each node) and then assign collision-free channels to links. To ease the assignment problem, a hierarchical structure is formed in the network to localize groups of nodes and make the task of channel assignment more manageable. The problem, however, is how to determine the cluster memberships and the cluster heads so that the whole network is covered while the sensors move [SGA00].

Several attempts have been done to this end, including the Self-Organizing Medium Access Control (SMACS) in [SGA00]. The SMACS scheme proposed in [SGA00] is a combination of the neighbor discovery and channel assignment phases. In SMACS a channel is immediately assigned to a link immediately after the link's existence is discovered. Thus, by the time all nodes hear all their neighbors, they will have formed a connected network, where there exists at least one multihop path between any two distinct nodes. However, there is the possibility for time collisions with slots assigned to adjacent links whose existence is not known at the time of channel assignment. To reduce the likelihood of collisions, reference [SGA00] requires each link to operate on a different frequency, where each frequency is chosen from a large pool of frequencies. Reference [FFM05] argues that since the number of channels in SMACS is a function of the density of the links, the scheme will not be easily applicable.

5.8.5 Topology control

Besides the necessity of designing protocols that achieve a self-organization of the sensor network that is robust to failures or mobility, there several other criteria that a communication graph needs to satisfy. The schemes that were presented above do not distinguish among neighbors. Instead, by the time a neighbor is detected, a channel is assigned to it.

Other criteria that a communication graph needs to satisfy include

– minimal connectivity,

– bounded degree,

– minimum interference,

– small number of hops in each path (low *stretch*),

– energy efficiency.

In addition, we need to take into account that decentralized schemes for topology control do not typically have access to directional or positional information. Memory limitations in sensor nodes also impose the restriction that a node can only keep track of $O(1)$ neighbors. Furthermore, no global clock or other synchronizing mechanism is assumed, but all sensor nodes have the same clock frequency [FFM05].

It is reasonable to consider that implicitly a minimally connected graph that has minimum degree it will have minimum interference. Indeed interference was only implicitly considered in the early work of topology control [BLR03, SWL04, FFM05, San05, LSW05, DPP06]. For example, in [BLR03] a network formation protocol is introduced, where the node degree k is a constant tuned to ensure connectivity with high probability is given. A self-organizing method that provides bounded power stretch factor and bounded node degree is proposed in [SWL04], although distance estimation hardware is necessary for the application of this protocol. In [FFM05], a protocol is proposed that guarantees constant degree spanning graph with optimal hop-stretch. This approach does not make use of any positional or directional hardware, but only of the relative position among nodes.

As far as node interference is concerned, we would like to note that the intuition behind these approaches is that a low (constant) node degree at all nodes would solve the interference issue automatically. According to [LRW08], node interference need to be dealt and solved separately before deriving distributed algorithms that will

guarantee all the above requirements. Several studies on these lines include the work of [BRW04]. It is not clear yet what is the complexity of the optimization problem of minimizing the interference, and it is considered an open problem in sensor networks [LRW08].

5.8.6 An information-based learning approach

We would like to examine the utility of the proposed learning algorithms in the topology control problem in wireless ad-hoc networks. To this end, we will consider the *one-way benefit model* for network formation as described in Section 5.4.

Such a network formation scheme is **energy efficient**. Connected networks with directed links require smaller amount of communications among nodes than networks with non-directed links. The work on topology control cited above assumes *only* bidirectional links among nodes. Furthermore, this network formation model establishes **connected networks** with positive probability (see Claim 5.6.1 and Proposition 5.6.1).

Moreover, this scheme can be designed so that it establishes networks with **bounded number of hops** between any two *neighboring* nodes with positive probability. In particular, by applying a *decaying benefit flow* scheme as presented in Section 5.6.2 we showed that a decay value can be designed so that the distance between any two neighboring nodes is bounded above with positive probability (see Proposition 5.6.4).

Finally, this network formation model can establish networks with **bounded node degree**. Although the learning algorithm penalizes the establishment of a new link, the process may converge to a configuration where the number of links of a node is $n-1$ where n is the total number of nodes (e.g., the star network is a possible limit point of the process when a decaying benefit flow is considered, see Fig. 5.6). Therefore, we cannot rely on the fact that establishment of new links is costly in order to bound the node degree.

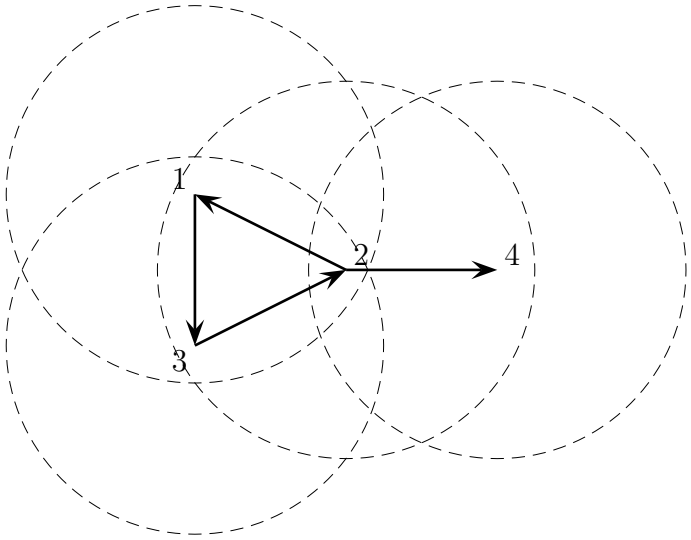


Figure 5.8: A Nash network in case of $n = 4$ nodes for some given neighborhood structure, frictionless benefit flow and maximum allowed number of links equal to 1.

Such criterion can be achieved with probability one by suppressing the number of actions (links) that can be established by each agent. However, we need to note that this constraint may result in losing network connectivity when the network is not dense enough. For example, when the benefit flow is frictionless (i.e., the benefits are not decayed) and each agent may establish at most one link, Fig. 5.8 shows a possible *Nash network*[13] of the learning process for some given neighborhood structure. Note that node 2 is not connected with node 4, which implies that the network is *not* connected.

However, when we do not constrain the neighborhood of each node, then Nash networks will be connected networks.

Proposition 5.8.1 *Consider a set of* $\mathcal{I} = \{1, 2, ..., n\}$ *nodes deployed on the plane, and let* $\mathcal{N}_i = \mathcal{I} \backslash i$ *be the neighboring nodes of node* $i \in \mathcal{I}$*. Assume the one-way benefit flow model of Section 5.4.1, where for each node* $i \in \mathcal{I}$ *the action space is* $\mathcal{A}_i \triangleq \{\alpha_i \in 2^{\mathcal{N}_i} : |\alpha_i| \leq M\}$ *for some* $1 \leq M < n - 1$*. Consider also the reward*

[13]See Definition 5.6.1.

function of (5.3) with $0 < \delta < 1$ *and the cost function of (5.4) with* $\kappa_0 + \kappa_1 \leq \delta$. *Then, a Nash network is connected when* δ *is sufficiently close to 1.*

Proof. Assume that a Nash network, say $\mathcal{G}$, is not connected. This implies that there exist $i, j \in \mathcal{I}$ such that $(i \leftarrow j) \notin \mathcal{G}$. Note that node i may establish at most M links where $1 \leq M < n-1$. There are two possibilities: (a) node i has currently established $< M$ links, (b) node i has currently established M links. It suffices to show that in both cases, network $\mathcal{G}$ is not a Nash network, which is a contradiction to our initial assumption.

(a) Assume that node i has currently established $< M$ links. In that case, node i can always benefit by establishing a new link with node j since $\delta \geq \kappa_0 + \kappa_1$, which implies that $\mathcal{G}$ is not a Nash network.

(b) Assume now that node i has currently established M links. Let us first define the set of nodes where node i is connected to either directly or indirectly as $I \triangleq \{h \in \mathcal{I} : (i \leftarrow h) \in \mathcal{G}\}$. Similarly, define for node j the set $J \triangleq \{k \in \mathcal{I} : (j \leftarrow k) \in \mathcal{G}\}$. There are three possible cases: (b1) $I \cap J \neq \varnothing$, (b2) $i \in J$ and $I \cap J = \varnothing$, and (b3) $i \notin J$ and $I \cap J = \varnothing$.

(b1) Let $I \cap J \neq \varnothing$. In this case, there exist $h \in I \backslash (I \cap J)$ and $k \in I \cap J$ such that $(h, k) \in \mathcal{G}$. For δ sufficiently close to one, node h has the incentive to drop its link with k and establish a link with j, since in this case it can still access the benefits from k plus the new benefits from j. This implies that the network is not a Nash network.

(b2) Let $i \in J$ and $I \cap J = \varnothing$. In this case, we consider the following possibilities: (1) there exists $h \in I$ such that $(h, i) \in \mathcal{G}$, or (2) there is no $h \in I$ such that $(h, i) \in \mathcal{G}$. In the first case, for δ sufficiently close to 1, node h has the incentive to drop its link with i and connect with j, since it gets access to the benefits of both j and i. In the second case, and since $I \cap J = \varnothing$, there exist nodes $h, k \in I$ such that $(h, k) \in (i \leftarrow k)$ and $(k, h) \notin (i \leftarrow k)$. For δ sufficiently close to one, node k has the incentive to drop

its link with h and establish a link with j, since in this case it can access the benefits from i and i's links (including h), and the new benefits from j. This implies that the network is not a Nash network.

(b3) Let $i \notin J$ and $I \cap J = \varnothing$. In other words, there is no link from the set $i \cup I$ to $j \cup J$ and vice versa. Consider a node $h \in i \cup I$ and a node $k \in j \cup J$. Let α be the set of nodes node h has access to directly or indirectly. Note that $\alpha \subseteq i \cup I$. Similarly, define β as the set of nodes node k is connected to. Note also that $\beta \subseteq j \cup J$. We observe that either $|\alpha| \leq |\beta|$, $|\alpha| > |\beta|$. If $|\alpha| \leq |\beta|$, then node h can benefit from dropping any one of its links and connect to k when δ is sufficiently close to 1. In the case when $|\alpha| > |\beta|$, then node k can benefit from dropping any one of its links and connect with h, when δ is sufficiently close to 1. In either case, we see that the network is not a Nash network when δ is sufficiently close to 1. □

In other words, Proposition 5.8.1 states that for any M such that $1 \leq M < n-1$, a Nash network is connected as long as the decay factor is sufficiently close to 1 or equal to 1. Fig. 5.9(a) shows a connected network where the maximum internode distance among any two nodes is $n-1$ (dependent on n), which is a Nash network if δ is sufficiently close to 1 and $M = 1$.

If instead we prefer to bound the internode distance by $d_{\max}$, that might result into a non-connected Nash network when M is small compared to n. For example, Fig. 5.9(b) shows a Nash network when δ is not close to 1 and M is small compared to n. We summarize these remarks as follows:

We summarize the above observations for the unbounded neighborhood case as follows:

Proposition 5.8.2 *Consider a set of $\mathcal{I} = \{1, 2, ..., n\}$ nodes deployed on the plane, and let $\mathcal{N}_i = \mathcal{I} \backslash i$ be the set of neighboring nodes of node $i \in \mathcal{I}$. Assume that each node i applies the reinforcement learning scheme of (5.1), where the reward function*

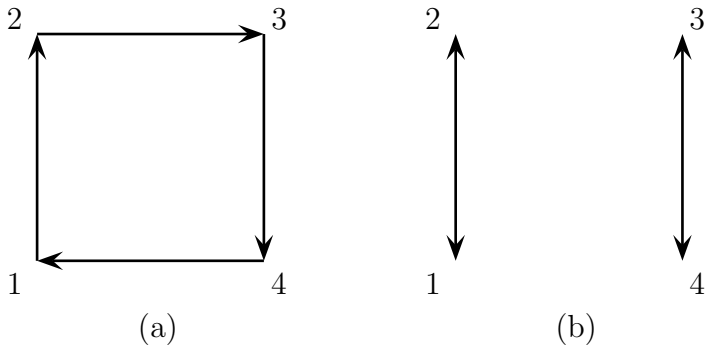


Figure 5.9: A Nash network in case of $n = 4$ nodes for unbounded neighborhood for each node, with (a) frictionless benefit flow and maximum allowed number of links per node $M = 1$, (b) $\delta^2 < \kappa_1$ and maximum allowed number of links per node $M = 1$.

is given by (5.3) for some $0 < \delta \leq 1$ *and the cost function is given by (5.4) with* $0 < \kappa_0 + \kappa_1 \leq \delta$. *Assume finally that for some* $1 \leq M \leq n-1$ *and for each* $i \in \mathcal{I}$ *the action set* $\mathcal{A}_i$ *is* $\mathcal{A}_i \triangleq \{\alpha_i \in 2^{\mathcal{N}_i} : |\alpha_i| \leq M\}$. *Then a Nash network*

1. *is connected and has maximum node degree* $1 \leq M < n-1$, *when* δ *is sufficiently close to 1,*

2. *is connected and has maximum internode distance* $d_{\max} \in \{2, 3, ...\}$, *when* $0 < \kappa_0 + \kappa_1 \leq \delta - \delta^{d_{\max}}$ *and* $M = n - 1$.

Proof. The first conclusion follows from Proposition 5.8.1. The second conclusion follows from Claim 5.6.1 and Proposition 5.6.4. □

Another observation is:

Remark 5.8.1 *When* $1 \leq M < n-1$ *and* δ *is sufficiently close to 1, then the wheel network is the efficient network*[14] *and furthermore it is a Nash network. Also, the dynamic reinforcement scheme introduced in Section 5.4 can be applied to reinforce convergence to the efficient network as shown by Proposition 5.7.1.*

[14]See Definition 5.4.3.

5.9 Remarks

We presented a method for distributed network formation and reinforcement of efficient networks by dynamic reinforcement. Some key distinguishing features of this work include: i) payoff based dynamics, in which each agent adapts according to realized link rewards and costs; ii) incorporation of state dependent link establishment costs in addition to link maintenance costs; and iii) reinforcement of efficient networks through dynamic reinforcement. We presented various characterizations and properties of Nash network configurations, in terms of the structure of their connectivity or the distances between nodes. We also provided accompanying convergence results that show how these network configurations can be the outcome of a learning process. Finally, we illustrated the utility of this approach in topology control of wireless ad-hoc sensor networks.

CHAPTER 6

Conclusions and Future Work

6.1 Conclusions

This thesis was a small contribution to the problem of equilibrium selection in coordination problems for multiagent systems. Our objective was to model interactions among agents that are adaptive and robust to possible environmental changes. We accomplish that by considering a learning approach where agents learn their behavior through time based on reinforcement learning (payoff-based dynamics). Agents do not have access to the selections and strategies of other agents which makes the proposed techniques attractive for designing multiagent coordination problems as demonstrated clearly in the example of distributed network formation.

In Chapter 3, we assumed that agents apply a reinforcement learning scheme which is a small modification of the classical reward-inaction scheme of learning automata. We showed that when agents are involved in a coordination problem multiple equilibria might emerge as the asymptotic outcome of the learning algorithm. For the unperturbed reinforcement scheme, we were able to show that the learning process converges w.p.1 to the set of vertices of the probability space (for both constant and diminishing step size sequences).

For the perturbed reinforcement scheme with constant step size and for a special class of coordination games, we showed that any small neighborhood of the vertices of the probability space is a recurrent set of the process when the perturbation is sufficiently small. Further characterization of the possible asymptotic outcomes can

be based on ODE methods for stochastic approximations. We apply the ODE method for the case of diminishing step size sequences, although the results are also valid for constant step size (in the context of weak-convergence). According to this method, the stochastic process converges with positive probability to a locally stable set of the relevant ODE.[1]

In Chapter 4, we specialized our results for the case of coordination games. Besides characterizing the possible outcomes of the process, our intention was to also control its outcome. To this end, new forms of agent's decision rules were introduced that are based on feedback control. According to these rules, recent observations count more in agent's decisions than older observations. Under this new framework, we showed that predictions of the final outcome of a coordination problem can be changed considerably in a distributed way. In particular, although the reinforcement scheme without dynamic reinforcement may converge to multiple outcomes, including possibly non-efficient ones, the reinforcement scheme with dynamic reinforcement can be designed so that it does not converge to non-efficient outcomes. Note also that this is possible even when dynamic reinforcement is applied only by a *single* agent. From the one hand, this demonstrates the utility of feedback control techniques in distributed learning, while, on the other hand, it reveals the possible fragility of agent-based simulation models.

The results where demonstrated in coordination games (which is a special class of a coordination problem) and distributed network formation. In coordination games, prior work has shown that the risk-dominant equilibrium "seems" to be the only reasonable prediction, even when it is not payoff-dominant. Our intention here was to show that transient phenomena in learning dynamics can be exploited in a distributed manner to alter the convergence properties in a desirable way. In particular, it was shown that the dynamic reinforcement scheme, when applied by a single agent, is able

[1]Defined in Chapter C.

to destabilize the risk-dominant equilibrium independently of the number of agents playing the game.

In Chapter 5, we analyzed the problem of distributed network formation under the proposed framework of reinforcement learning, which is a problem of independent interest. Prior learning models used for modeling network formation, such as best-reply learning models, usually assume that each agent has access to the previous actions of all other agents. The proposed model of reinforcement learning assumes minimal amount of information available to each agent. Furthermore, we illustrated the flexibility of this scheme to incorporate various design criteria, including dynamic cost functions that reflect link establishment and maintenance, and distance-dependent benefit functions. We showed that the dynamic process may converge to multiple stable configurations (i.e., strict Nash networks), which need not emerge under alternative processes such as best-reply dynamics. Finally, we showed that in the case of frictionless benefit flow (i.e., when there is no discount) a single agent can reinforce the emergence of an efficient network through dynamic reinforcement.

The utility of the network formation model was also illustrated in the problem of topology control for wireless ad-hoc networks. In particular, the problem is to design distributed techniques that guarantee several modeling criteria, such as bounded node degree, bounded internode distance, connectivity, energy efficiency and robustness to possible failures. Under the proposed reinforcement scheme, we can design reward functions that support stable configurations (strict Nash networks) with these criteria. In particular, we showed that the node degree can be bounded by restricting the actions set of each agent, the internode distance can be bounded by applying a discount factor on the flow of benefits, while we found conditions under which Nash networks are connected networks. Finally, the dynamic reinforcement scheme can also be used to destabilize non-efficient network structures when the benefit flow is assumed frictionless.

6.2 Future directions

The analysis presented for equilibrium selection in the case of dynamic reinforcement and distributed network formation in Chapters 4–5 was only local. A first attempt towards the global characterization of convergence was presented in Chapter 3, where the proposed algorithm (perturbed or unperturbed) was analyzed using martingale convergence theorems. Recall that for the perturbed reinforcement scheme with constant step size, we showed that any small neighborhood of the vertices of the probability space is a recurrent set of the process when the perturbation is sufficiently small. This, however, does not ensure where the process will converge.

In order to characterize where the process will converge we may apply results from stochastic approximations similar to the ones applied for diminishing step size in Chapters 4–5. This analysis is based on weak convergence techniques and shows that the probability that the process lies in an chain recurrent set of the corresponding ODE goes to one as the step size approaches zero.

However, such a statement does not exclude the possibility that there are chain recurrent sets that are other than pure Nash equilibria. In order to exclude that possibility, we may apply techniques based on large deviations [DZ93], based on which we can compute the robustness of each stable set to noise. The process will spend most of its time to the stable set that is the most robust to noise, which constitutes a different notion of stability, namely *stochastic stability* [FW84].

Simulation results has shown that certain Nash equilibria are more robust than others. In particular, risk-dominant Nash equilibria are more robust than non-risk-dominant equilibria when the perturbed reinforcement scheme with constant step size is applied. In other words, the process spends most of its time at the risk-dominant equilibrium (i.e., this equilibrium is *stochastically stable*). This property is attributed to the fact that the risk-dominant equilibrium has the largest basin of attraction,

where the basin of attraction corresponds to the smallest deviation cost from that equilibrium, or equivalently, the largest eigenvalue of the linearized dynamics of the corresponding ODE (as the analysis of Chapter 4 showed). Therefore, to guarantee convergence to a "desirable" equilibrium, it is sufficient to make its basin of attraction larger than the basin of attraction of all other equilibria. Thus, it is not really necessary to destabilize those equilibria.

On the other hand, large deviations techniques do not seem promising when applied in large games (such as the network formation problem) under the proposed reinforcement scheme. So far, stochastic stability results for infinite-space learning dynamics have only been derived for games with small number of agents and actions, e.g., the analysis under fictitious play dynamics in [Wil02], but the results were only numerical. An analytical characterization of the global convergence properties of the proposed reinforcement scheme is necessary. Such an analytical characterization will also increase the utility of the proposed techniques into a large class of problems that can fit into the framework of coordination problems.

APPENDIX A

Martingales

A.1 Martingale convergence theorem

Let $\{X(k)\}_{k\geq 0}$ be a sequence of real random variables on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_k$ be a sequence of sub-σ-fields of $\mathcal{F}$ with

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset ...\mathcal{F}_{k-1} \subset \mathcal{F}_k... \subset \mathcal{F} \tag{A.1}$$

and let $X(k)$ be measurable with respect to $\mathcal{F}_k$. $\mathcal{F}_k$, for example, can be the σ-field generated by random variables $Y(0), Y(1), ..., Y(k)$, where the sequence $\{Y(k)\}_{k\geq 0}$ is also defined on $(\Omega, \mathcal{F}, P)$. We can consider $\mathcal{F}_k$ as containing the information available at stage k. Also, $X(k)$ is measurable with respect to $\mathcal{F}_k$ if it is determined by $\{Y(0), Y(1), ..., Y(k)\}$. The definition of a martingale in terms of σ-fields may be stated as follows:

Definition A.1.1 (Martingale) *Let $\{X(k)\}_{k\geq 0}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{F}_k$ be a sequence of sub-σ-field of $\mathcal{F}$ satisfying equation (A.1). Then $\{X(k)\}$ is called a submartingale with respect to $\{\mathcal{F}_k\}$ if for all k*

1. *$X(k)$ is measurable with respect to $\mathcal{F}_k$,*
2. *$E[X(k)^+] < \infty$, where $X(k)^+ = \max\{0, X(k)\}$, and*
3. *$E[X(k+1)|\mathcal{F}_m] \geq X(m)$, for $m \leq k$.*

Note that if $\{-X(k)\}$ is a submartingale, then $\{X(k)\}$ is a supermartingale. If both $\{-X(k)\}$ and $\{X(k)\}$, then $\{X(k)\}$ is a martingale with respect to $\{\mathcal{F}_k\}$.

Theorem A.1.1 (Martingale Convergence Theorem) *(a) Let $\{X(k), \mathcal{F}_k\}_{k\in\mathbb{N}}$ be a nonnegative supermartingale satisfying*

$$\sup_{k\geq 0} E[X(k)] < \infty.$$

Then there exists a random variable X_∞ to which $\{X(k)\}$ converges w.p.1, i.e.,

$$P[\lim_{k\to\infty} X(k) = X_\infty] = 1.$$

(b) If $\{X(k), \mathcal{F}_k\}_{k\in\mathbb{N}}$ is a positive supermartingale and for some $\alpha > 1$ the function $E[|X(k)|^\alpha]$ is bounded, then in addition to part (a),

$$\lim_{k\to\infty} E[X(k)] = E[X_\infty] = E[\lim_{k\to\infty} X(k)].$$

For a submartingale $\{X(k)\}$,

$$\sup_{k\geq 0} E[X(k)^+] < \infty \Rightarrow \sup_{k\geq 1} E[|X(k)|] < \infty.$$

Hence by the theorem, every non-positive submartingale, nonnegative supermartingale, or martingale that is uniformly bounded from above converges with probability 1.

A corollary of Theorem A.1.1 is the following:

Corollary A.1.1 *Under the conditions of Theorem A.1.1,*

$$E[X(k+1) - X(k)|X(0), X(1), ..., X(k)] \to 0 \quad \textit{w.p.1 as } k \to \infty.$$

APPENDIX B

Convergence of Markov Processes

B.1 Discrete-time Markov processes

Let $X(k)$, $k = 0, 1, 2, ...$, be a stochastic process with values in E_l (Euclidean l-space). $X(k)$ is a Markov process if the probability of an event, say $X(k) \in \Gamma$, given $X(s) = x$ is not affected if the behavior of the process up to time s is also known. This probability, known as the *transition probability* of the Markov process, will be denoted by $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$. Clearly, $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$ must be a probability measure as a function of Γ and $\mathcal{B}_l$-measurable as a function of x, where $\mathcal{B}_l$ is the Borel σ-algebra of dimension l.

A transition function needs to satisfy the *Chapman-Kolmogorov relations*

$$\mathrm{P}(X(k) \in \Gamma | X(s) = x) = \int_A \mathrm{P}(X(u) \in dy | X(s) = x) \cdot \mathrm{P}(X(k) \in \Gamma | X(u) = y), \quad \text{(B.1)}$$

where $A \in \mathcal{B}_l$.

Definition B.1.1 (Markov process) *Let* $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$ *be a transition function in* A, $A \in \mathcal{B}_l$. *A process* $X(k)$, $k = 0, 1, 2, ...$ *in* A *is called a Markov process with transition function* $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$ *if, for* $s < u < k$,

$$P[X(k) \in \Gamma | X(s), X(s+1), ..., X(u)] = \mathrm{P}(X(k) \in \Gamma | X(u)) \quad \textit{(a.s.)}$$

A transition function $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$ is said to be *homogeneous* if $\mathrm{P}(X(k+$

$s) \in \Gamma | X(s) = x)$ is independent of s.

When dealing below with various problems related to a Markov process $X(k)$, $k = k_0, k_0+1, ...$, for some $k_0 > 0$, with values in E_l and transition function $\mathrm{P}(X(k) \in \Gamma | X(s) = x)$, we shall find it convenient to use operator Δ defined on functions $V(t, x)$, $x \in E_l$, by

$$\Delta V(k, x) = \int \mathrm{P}(X(k+1) \in dy | X(k) = x) \cdot [V(k+1, y) - V(k, x)]. \qquad \text{(B.2)}$$

B.1.1 Markov processes and supermartingales

Martingales and supermartingales may be used as tools in studying the limiting behavior of sample functions of Markov processes.

Consider a Markov process $X(k)$, $k \geq k_0$, in E_l. Consider also a sequence of imbedded σ-algebras, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$, so that $X(k)$ is $\mathcal{F}_k$-measurable. Let τ_G denote the first exit time of the sample function of the process from a domain G, and

$$\tau_G \wedge k \triangleq \min(\tau_G, k).$$

Theorem B.1.1 *Let $V(k, x)$ be a nonnegative function, $V(k, x) \in D_L$ for $k \geq k_0$ and $x \in D$, such that*

$$\Delta V(k, x) \leq 0 \quad (\Delta V(k, x) = 0)$$

for all these points (k, x), and $\mathrm{E}V(k_0, X(k_0)) < \infty$. Then

$$\{Y(k) = V(\tau_G \wedge k, X(\tau_G \wedge k)), \mathcal{F}_k\}$$

is a supermartingale (martingale) for $k \geq k_0$.

Proof. See pg. 32 of [NH76]. □

B.1.2 Exit of sample functions from a domain

It is important to have conditions under which a Markov process $X(k) = X(k, \omega)$, $k \geq 0$, with some initial distribution, will almost surely (i.e., with probability 1) leave an open domain G in E_l in a finite time. We will assume that $X(k)$ is a Markov process with generating operator Δ and arbitrary initial distribution.

Theorem B.1.2 *Suppose that there exists a nonnegative function, $V(k, x)$ in the domain $k \geq 0$, $x \in G$, such that $\Delta V(k, x) \leq -\epsilon(k)$ in this domain, where $\epsilon(k)$ is a sequence such that*

$$\epsilon(k) > 0, \quad \sum_{k=0}^{\infty} \epsilon(k) = \infty. \tag{B.3}$$

Then a process $X(k)$ leaves G in a finite time with probability 1.

Proof. Here we follow the proof of Theorem 5.1 in [NH76]. Let x be any point of G. It will clearly suffice to prove the theorem for a process $\{X(k)\}_k$. Denote the first exit time of the sample functions of $X(k)$ from G by τ. Set

$$W(k, x) = V(k, x) + \beta(k), \tag{B.4}$$

where $\beta(k) = \sum_{u=0}^{k-1} \epsilon(u)$. It is clear, by (B.3), that

$$\lim_{k \to \infty} \beta(k) = \infty.$$

In addition,

$$\Delta W(k, x) = \Delta V(k, x) + \Delta\beta(k) = \Delta V(k, x) + \epsilon(k) \leq 0$$

for $k \geq 0$ and $x \in G$. Hence, by Theorem B.1.1, the pair

$$(W(\tau \wedge k, X(\tau \wedge k)), \mathcal{F}_k), k \geq 0,$$

is a supermartingale, and so, by Theorem A.1.1, a finite limit

$$\lim_{t \to \infty} W(\tau \wedge k, X(\tau \wedge k)) = \eta < \infty \tag{B.5}$$

exists with probability one. Since $V(k, x)$ is nonnegative, it follows from (B.4) and (B.5) that $\lim_{k\to\infty} \beta(\tau \wedge k)$ also exists and is finite with probability 1. But, since $\lim_{k\to\infty} \beta(k) = \infty$, we conclude that the r.v. $\tau \wedge k$ is bounded a.s. as a function of k. In other words, the process $X(k)$ leaves G in a finite time with probability 1. □

The following corollary of the above theorem is important in cases we would like to consider entrance of a stochastic process into the domain of attraction of an equilibrium.

Corollary B.1.1 *Let B be a subset of E_l, $\mathcal{B}_\varepsilon(B)$ its ε-neighborhood,*[1] *and $\mathcal{V}_\varepsilon(B) = E_l \backslash \mathcal{B}_\varepsilon(B)$. Suppose there exists a nonnegative function $V(k, x) \in D_L$ in the domain $k \geq 0$, $x \in E_l$ for which*

$$\Delta V(k, x) \leq -\epsilon(k)\varphi(k, x), \quad k \geq 0, x \in E_l, \tag{B.6}$$

where the sequence $\epsilon(k)$ satisfies (B.3) and $\varphi(t, x)$ satisfies

$$\inf_{k \geq T, x \in \mathcal{V}_\varepsilon(B)} \varphi(k, x) > 0 \tag{B.7}$$

[1]The distance $\mathrm{dist}(x, B)$ from a point x to a set B is defined as $\mathrm{dist}(x, B) \triangleq \inf_{y\in B} \mathrm{dist}(x, y)$, and $\mathcal{B}_\varepsilon(B) = \{x : \mathrm{dist}(x, B) < \varepsilon\}$. We shall also write $x \to B$ if $\mathrm{dist}(x, B) \to 0$.

for all $\varepsilon > 0$ *and some* $T = T(\varepsilon)$. *Then*

$$P[\liminf_{k\to\infty} \text{dist}(X(k), B) = 0] = 1.$$

Proof. First, we can show that for some $\varepsilon_1 > 0$ the stochastic process enters $\mathcal{B}_{\varepsilon_1}(B)$ in finite time τ_1 with probability 1.[2] Then by taking $0 < \varepsilon_2 < \varepsilon_1$, we can show that the stochastic process will enter $\mathcal{B}_{\varepsilon_2}(B)$ in finite time $\tau_2 > \tau_1$ with probability 1. In this way, we can define a subsequence of time instants $\tau_1, \tau_2, ...$ for which $\text{dist}(X(\tau_n), B) < \varepsilon_n$ for all $n = 1, 2, ...$ Therefore, the conclusion of the corollary follows. □

Note that the corollary implies that the stochastic process enters an arbitrarily small neighborhood of a set B infinitely often with probability one. Such a conclusion is important when we are discussing convergence of stochastic processes.

[2]We can show that by following the proof of Theorem B.1.2.

APPENDIX C

ODE Method for Stochastic Approximations (SA)

C.1 Convergence analysis for SA

Consider the stochastic approximation

$$x(k+1) = x(k) + \epsilon(k) \cdot g(x(k), \alpha(k)) \tag{C.1}$$

that evolves in the domain $\mathbb{R}^r$. We will assume that the step size sequence will satisfy the fundamental condition

Assumption C.1.1 $\sum_{k=0}^{\infty} \epsilon(k) = \infty, \quad \epsilon(k) \geq 0, \quad \epsilon(k) \to 0, \ \textit{for } k \geq 0; \quad \epsilon(k) = 0, \ \textit{for } k < 0.$

Also, $g(x(k), \alpha(k))$ denote the $\mathbb{R}^r$-valued "observation" at time k that depends on the current state $x(k)$ and a noise term $\alpha(k)$. We assume that the noise term $\alpha(k)$ takes values in some topological space $\mathcal{A}$. Let $\mathcal{F}(k)$ denote the σ-algebra determined by the initial condition $x(0)$ and observations $g(x(i), \alpha(i))$, $i < k$.

In general, we will consider a noise process $\{\alpha(k)\}_k$ that is Markov state-dependent. For each x, let $\mathrm{P}(\cdot, \cdot|x)$ be a Markov transition function parameterized by x such that $\mathrm{P}(\cdot, A|\cdot)$ is Borel measurable for each Borel set A in the range space $\mathcal{A}$ of $\alpha(k)$. Suppose that the law of evolution of the noise satisfies

$$P[\alpha(k+1) \in \cdot | \alpha(i), x(i), i \leq k] = \mathrm{P}(\alpha(k), \cdot | x(k)), \tag{C.2}$$

where $\mathrm{P}(\alpha, \cdot|x)$ denotes the one-step transition probability with starting point x and parameterized by x. If (C.3) holds, the noise process $\{\alpha(k)\}_k$ is called *Markov state-dependent*.

In the stochastic recursions we will consider in this thesis, we will only encounter the special case where

$$P[\alpha(k+1) \in \cdot|\alpha(i), x(i), i \leq k] = \mathrm{P}(\cdot|x(k)), \tag{C.3}$$

which implies that the "*next*" noise depends only on the "*current*" state. In this case, the noise process will be called "*state-dependent*".

The analysis of the stochastic approximation (C.1) when the noise is state-dependent, as described by (C.3), coincides with the analysis of the stochastic approximation with a *fixed-x noise process*. In particular, if $\mathrm{P}(\cdot|x)$ is the transition function of a Markov chain; the fixed-x noise process, denoted by $\alpha_k(x)$, will correspond to the random variables of that chain. We expect that the probability law of this chain for given x is close to the probability law of the true noise $\{\alpha(k)\}_k$ if $x(k)$ varies slowly around x, and hence that the mean ODE can be obtained in terms of this fixed-x chain. This turns out to be the case. Of special interest is the fixed-x process $\{\alpha_i(x(k)), i \geq k\}_i$, defined for each k with initial condition $\alpha_k(x(k)) = \alpha(k)$. Thus, this process starts at value α_k at time k and then evolves as if the parameter were fixed at $x(k)$ forever after.

Assumption C.1.2 $\sup_k E[|g(x(k), \alpha(k))|] < \infty$

Assumption C.1.3 *$g(x, \alpha)$ is continuous in x for each α.*

Assumption C.1.4 *There is a continuous function $\overline{g}(\cdot)$ such that for $x \in \mathbb{R}^r$, the expression*

$$v_k(x, \alpha(k)) = \sum_{i=n}^{\infty} \epsilon(i) E[g(x, \alpha_i(x)) - \overline{g}(x)|\alpha(k)]$$

is well defined when the initial condition for $\{\alpha_i, i \geq k\}$ *is* $\alpha_k(x) = \alpha(k)$*, and*

$$v_k(x(k), \alpha(k)) \to 0 \quad w.p.1.$$

The convergence analysis of the stochastic process $\{x(k)\}_k$ will be expressed in terms of the continuous time interpolation of $x(k)$. A natural time scale for the interpolation is defined in terms of the step size sequence. In particular, define $t_0 = 0$ and $t_k = \sum_{i=0}^{k-1} \epsilon_i$. Define the *continuous-time interpolation* $x^0(\cdot)$ on $(-\infty, \infty)$ by $x^0(t) = x(0)$ for $t \leq 0$, and for $t \geq 0$,

$$x^0(t) = x(k), \quad \text{for } t_k \leq t \leq t_{k+1}.$$

Define also the sequence of *shifted* processes $x^k(\cdot)$ by

$$x^k(t) = x^0(t_k + t), \quad t \in (-\infty, \infty).$$

Proposition C.1.1 (Convergence of SA) *Assume C.1.1, C.1.2, C.1.3 and C.1.4. Then there is a null set* N *such that for* $\omega \notin N$*, the set of functions* $\{x^k(\omega, \cdot), k < \infty\}$ *has a subsequence that converges to some continuous limit, uniformly on each bounded interval. Let* $x(\omega, \cdot)$ *denote the limit of some convergent subsequence.*

(a) If $\{x(k)\}_k$ *is bounded with probability one, then for almost all* ω*, the limits* $x(\omega, \cdot)$ *of convergent subsequences of* $\{x^k(\omega, \cdot)\}$ *are trajectories of*

$$\dot{x} = \overline{g}(x) \tag{C.4}$$

in some bounded invariant set and $\{x(k)\}$ *converges to this invariant set.*

(b) If $A \subset \mathbb{R}^r$ *is locally asymptotically stable in the sense of Lyapunov for (C.4) and* $x(k)$ *is in some compact set in the domain of attraction of* A *infinitely often with probability* $\geq \rho$*, then* $x(k) \to A$ *with at least probability* ρ*.*

(c) Suppose there is a unique solution for each initial condition. There is a null set N such that if $\omega \notin N$ and if for points x and $\overline{x}$

$$x(k) \in \mathcal{B}_\delta(x), \quad x(k) \in \mathcal{B}_\delta(\overline{x}), \quad \textit{infinitely often} \tag{C.5}$$

for all $\delta > 0$. Then x and $\overline{x}$ are chain connected.[1] *Thus the assertions concerning convergence to an invariant or limit set of the mean ODE (C.4) can be replaced by convergence to a set of chain recurrent points within that invariant or limit set.*

Proof. This proposition is a special case of Theorem 6.1 in Chapter 6 of [KY97]. □

Note that for a noise process that satisfies (C.3), we have

$$E[g(x, \alpha_i(x)|\xi(k)] \equiv E[g(x, \alpha_i(x))] \equiv E[g(x, \alpha(i))|x(i) = x] \quad \text{for all } i \geq k.$$

Therefore, if we define

$$\overline{g}(x) \triangleq E[g(x, \alpha(k))|x(k) = x]$$

then, the process $\{v_k(x, \alpha(k))\}_k$ is well-defined since it is identically zero. Thus, for a stochastic approximation (C.1) where the noise sequence satisfies (C.3), we just need to check assumptions C.1.2 and C.1.3 in order for Proposition C.1.1 to hold.

C.2 Non-convergence analysis for SA

Invariant sets of the ODE (C.4) may also include points that are linearly unstable. In practice, these points are often seen to be unstable points for the stochastic recursion. However, this does not follow directly from the convergence result of Proposition C.1.1. These points tend to be unstable for the recursion because of the

[1] We allow $x = \overline{x}$.

combination of the instability properties of the ODE near those points with the perturbing noise. Therefore, under appropriate "directional nondegeneracy" conditions on the noise, instability theorems based on Lyapunov function or large deviations methods can be used to prove the repelling property of the linearly unstable points of the ODE (C.4).

Let $D \subseteq \mathbb{R}^r$ be an open subset of an affine subspace in $\mathbb{R}^r$. Let $g : D \to TD$ be a class C^1, where TD is the translation of the affine subspace that contains the origin. We rewrite the stochastic approximation (C.1) as

$$x(k+1) = x(k) + \epsilon(k) \cdot \overline{g}(x(k)) + \epsilon(k) \cdot \xi(k) \tag{C.6}$$

where

$$\xi(k) \triangleq E[g(x(k), \alpha(k)) - \overline{g}(x) | x(k) = x]. \tag{C.7}$$

Note that $E[\xi(k)|x(k)] = 0$, and assume that it is such that $x(k)$ always remain in D.

Let x^* be a stationary point of the ODE (C.4), and suppose some eigenvalue of $\overline{g}_x(x^*)$ has a positive real part. Then the point is said to be *linearly unstable*.

Proposition C.2.1 (Non-convergence of SA) *Let the stochastic process* $\{x(k) : k \geq 0\}$ *be defined so that it satisfies (C.6) for some sequence of random variables* $\{\xi(k)\}$ *as defined in (C.7). Let* x^* *be any point of* D *with* $\overline{g}(x^*) = 0$*, let* $\mathcal{B}$ *be a neighborhood of* x^* *and assume that there are constants* $\nu \in (1/2, 1]$ *and* $c_1, c_2, c_3, c_4 > 0$ *for which the following conditions are satisfied whenever* $x(k) \in \mathcal{B}$ *and* k *is sufficiently large:*

1. x^* *is a linearly unstable critical point,*
2. $c_1/k^\nu \leq \epsilon(k) \leq c_2/k^\nu$*,*
3. $E[(\langle \xi(k), x \rangle)^+ | x(k) = x] \geq c_3/k^\nu$ *for every unit vector* $x \in TD$*,*

4. $\xi(k) \leq c_4/k^{\nu}$,

where $(\langle \xi(k), x \rangle)^+ = \max\{\langle \xi, x \rangle, 0\}$, *and* $\langle \xi, x \rangle$ *denotes the inner product of* ξ *and* x. *Assume* $\overline{g}$ *is smooth enough to apply the stable manifold theorem: at least* C^2. *Then*

$$P[x(k) \to x^*] = 0.$$

Proof. This proposition coincides with Theorem 1 in [Pem90]. □

APPENDIX D

Proofs

D.1 Proofs of Chapter 3

D.1.1 Proof of Claim 3.6.1

According to the definition of the conditional expectation of the change in payoff, we have

$$\begin{aligned}
\Delta R(k) &= E[R(\alpha(k+1)) - R(\alpha(k))|x_1(k), x_2(k)] \\
&= E[x_1(k+1)^{\mathrm{T}} D x_2(k+1) - x_1(k)^{\mathrm{T}} D x_2(k)|x_1(k), x_2(k)] \\
&= E[[x_1(k+1) - x_1(k)]^{\mathrm{T}} D[x_2(k+1) - x_2(k)] + \\
&\quad x_1(k)^{\mathrm{T}} D[x_2(k+1) - x_2(k)] + [x_1(k+1) - x_1(k)]^{\mathrm{T}} D x_2(k)|x_1(k), x_2(k)] \\
&= E[\delta x_1(k)^{\mathrm{T}} D \delta x_2(k) + x_1(k)^{\mathrm{T}} D \delta x_2(k) + \delta {x_1}^{\mathrm{T}} D x_2(k)|x_1(k), x_2(k)] \\
&= E[\delta x_1(k)^{\mathrm{T}} D \delta x_2(k)|x_1(k), x_2(k)] + x_1(k)^{\mathrm{T}} D \Delta x_2(k) + \Delta x_1(k)^{\mathrm{T}} D x_2(k)
\end{aligned}$$

which completes the proof.

D.1.2 Proof of Proposition 3.6.2

When both automata apply the $\tilde{L}_{R-I}$ scheme, we have

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2|x_1(k), x_2(k)] \\
&= [\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{11} x_{21} (d_{11})^2 + [\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{11} x_{22} (d_{12})^2 + \\
&\quad [\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{12} x_{21} (d_{21})^2 + [\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{12} x_{22} (d_{22})^2
\end{aligned}$$

By expanding the first term of the above expression, we get

$$[\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{11} x_{21} (d_{11})^2 = \epsilon^2 (d_{11} - d_{12} - d_{21} + d_{22}) x_{11} x_{12} x_{21} x_{22} (d_{11})^2$$

Similarly, we get

$$\begin{aligned}
[\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{11} x_{22} d_{12} &= -\epsilon^2 (d_{11} - d_{12} - d_{21} + d_{22}) x_{11} x_{12} x_{21} x_{22} (d_{12})^2 \\
[\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{12} x_{21} d_{21} &= -\epsilon^2 (d_{11} - d_{12} - d_{21} + d_{22}) x_{11} x_{12} x_{21} x_{22} (d_{21})^2 \\
[\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{12} x_{22} d_{22} &= \epsilon^2 (d_{11} - d_{12} - d_{21} + d_{22}) x_{11} x_{12} x_{21} x_{22} (d_{22})^2
\end{aligned}$$

It is straightforward to show now that

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] = \\
&\quad \epsilon^2 x_{11} x_{12} x_{21} x_{22} (d_{11} - d_{12} - d_{21} + d_{22}) ((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2)
\end{aligned}$$

which completes the proof.

D.1.3 Proof of Proposition 3.6.3

Assume that automaton 1 has three actions, i.e., $|\mathcal{A}_1| = 3$, while automaton 2 has two actions, i.e., $|\mathcal{A}_2| = 2$ actions. In this case, we have

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] \\
&= E[[x_1(k+1) - x_2(k)]^{\mathrm{T}} D[x_2(k+1) - x_2(k)] | x_1(k), x_2(k)] \\
&= E[[\epsilon R_1(\alpha_1 - x_1(k))]^{\mathrm{T}} D[\epsilon R_2(\alpha_2 - x_2(k))] | x_1(k), x_2(k)] \\
&= [\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{11} x_{21} (d_{11})^2 + [\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{11} x_{22} (d_{12})^2 + \\
&\quad [\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{12} x_{21} (d_{21})^2 + [\epsilon(e_2 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{12} x_{22} (d_{22})^2 + \\
&\quad [\epsilon(e_3 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{13} x_{21} (d_{31})^2 + [\epsilon(e_3 - x_1)]^{\mathrm{T}} D[\epsilon(e_2 - x_2)] x_{13} x_{22} (d_{32})^2
\end{aligned}$$

By expanding the first term of the above expression, we get

$$\begin{aligned}
&[\epsilon(e_1 - x_1)]^{\mathrm{T}} D[\epsilon(e_1 - x_2)] x_{11} x_{21} (d_{11})^2 \\
&= \epsilon^2[(1 - x_{11})(1 - x_{21})d_{11} - x_{12}(1 - x_{21})d_{21} - x_{13}(1 - x_{21})d_{31} - \\
&\quad (1 - x_{11})x_{22}d_{12} + x_{12}x_{22}d_{22} + x_{13}x_{22}d_{32}]x_{11}x_{21}(d_{11})^2 \\
&= \epsilon^2[(x_{12} + x_{13})x_{22}d_{11} - x_{12}x_{22}d_{21} - x_{13}x_{22}d_{31} - \\
&\quad (x_{12} + x_{13})x_{22}d_{12} + x_{12}x_{22}d_{22} + x_{13}x_{22}d_{32}]x_{11}x_{21}(d_{11})^2 \\
&= \epsilon^2(d_{11} - d_{12} - d_{21} + d_{22})x_{11}x_{12}x_{21}x_{22}(d_{11})^2 + \\
&\quad \epsilon^2(d_{11} - d_{12} - d_{31} + d_{32})x_{11}x_{13}x_{21}x_{22}(d_{11})^2
\end{aligned}$$

Similarly we derive the rest of the terms and we finally get

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] \\
&= \epsilon^2[x_{11}x_{12}x_{21}x_{22}(d_{11} - d_{12} - d_{21} + d_{22})((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2) + \\
&\quad x_{11}x_{13}x_{21}x_{22}(d_{11} - d_{12} - d_{31} + d_{32})((d_{11})^2 - (d_{12})^2 - (d_{31})^2 + (d_{32})^2) + \\
&\quad x_{12}x_{13}x_{21}x_{22}(d_{21} - d_{22} - d_{31} + d_{32})((d_{21})^2 - (d_{22})^2 - (d_{31})^2 + (d_{32})^2)]
\end{aligned}$$

By induction, for the general case of $|\mathcal{A}_1| > 2$ actions for the automaton 1 and $|\mathcal{A}_2| = 2$ actions for the automaton 2, it can be shown that if:

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] = \\
&\quad \epsilon^2 \sum_{i,j=1,i\neq j}^{|\mathcal{A}_1|-1,|\mathcal{A}_1|-1} x_{1i}x_{1j}x_{21}x_{22}(d_{i1} - d_{j1} - d_{i2} + d_{j2})((d_{i1})^2 - (d_{j1})^2 - (d_{i2})^2 + (d_{j2})^2),
\end{aligned}$$

then,

$$\begin{aligned}
&E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] = \\
&\quad \epsilon^2 \sum_{i,j=1,i\neq j}^{|\mathcal{A}_1|,|\mathcal{A}_1|} x_{1i}x_{1j}x_{21}x_{22}(d_{i1} - d_{j1} - d_{i2} + d_{j2})((d_{i1})^2 - (d_{j1})^2 - (d_{i2})^2 + (d_{j2})^2),
\end{aligned}$$

Thus, the expression holds for arbitrary values of $|\mathcal{A}_1|$.

The same arguments carry over to the case where $|\mathcal{A}_2|$ strategies are available to the second agent. For example, if we consider the case of $|\mathcal{A}_1| = |\mathcal{A}_2| = 3$, then we have:

$$
\begin{aligned}
& E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)] \\
&= \epsilon^2 [x_{11}x_{12}x_{21}x_{22}(d_{11} - d_{12} - d_{21} + d_{22})((d_{11})^2 - (d_{12})^2 - (d_{21})^2 + (d_{22})^2) + \\
&\quad x_{11}x_{13}x_{21}x_{22}(d_{11} - d_{12} - d_{31} + d_{32})((d_{11})^2 - (d_{12})^2 - (d_{31})^2 + (d_{32})^2) + \\
&\quad x_{12}x_{13}x_{21}x_{22}(d_{21} - d_{22} - d_{31} + d_{32})((d_{21})^2 - (d_{22})^2 - (d_{31})^2 + (d_{32})^2) + \\
&\quad x_{11}x_{12}x_{21}x_{23}(d_{11} - d_{13} - d_{21} + d_{23})((d_{11})^2 - (d_{13})^2 - (d_{21})^2 + (d_{23})^2) + \\
&\quad x_{11}x_{13}x_{21}x_{23}(d_{11} - d_{13} - d_{31} + d_{33})((d_{11})^2 - (d_{13})^2 - (d_{31})^2 + (d_{33})^2) + \\
&\quad x_{12}x_{13}x_{21}x_{23}(d_{21} - d_{23} - d_{31} + d_{33})((d_{21})^2 - (d_{23})^2 - (d_{31})^2 + (d_{33})^2) + \\
&\quad x_{11}x_{12}x_{22}x_{23}(d_{12} - d_{13} - d_{22} + d_{23})((d_{12})^2 - (d_{13})^2 - (d_{22})^2 + (d_{23})^2) + \\
&\quad x_{11}x_{13}x_{22}x_{23}(d_{12} - d_{13} - d_{32} + d_{33})((d_{12})^2 - (d_{13})^2 - (d_{32})^2 + (d_{33})^2) + \\
&\quad x_{12}x_{13}x_{22}x_{23}(d_{22} - d_{23} - d_{32} + d_{33})((d_{22})^2 - (d_{23})^2 - (d_{32})^2 + (d_{33})^2)]
\end{aligned}
$$

In the more general case of $|\mathcal{A}_1| > 2$ and $|\mathcal{A}_2| > 2$ actions, the expression for $E[\delta {x_1}^{\mathrm{T}} D \delta x_2 | x_1(k), x_2(k)]$ contains $[|\mathcal{A}_1|\,(|\mathcal{A}_1| - 1)][|\mathcal{A}_2|\,(|\mathcal{A}_1| - 1)]/4$ terms.

D.2 Proofs of Chapter 4

D.2.1 Proof of Proposition 4.5.1

(sketch) We first linearize the first part of the vector field, $\overline{g}(x, y, \rho)$. The partial derivatives of $\overline{g}$ with respect to x are captured by the matrix $A^{\lambda,\gamma}$. It is straightforward to show that the non-diagonal blocks of $A^{\lambda,\gamma}$ are factored by λ. In particular, the change of the vector field of agent $i \in \mathcal{I}$, $\overline{g}_i(\cdot, \cdot, \cdot)$, when agent $l \in \mathcal{I} \backslash \{i\}$ perturbs his

strategy is captured by

$$A_{il}^{\lambda,\gamma} = \lim_{h \to 0} \frac{\overline{g}_i(\tilde{x}_l + h\delta x_l, \tilde{x}_{-l}, \tilde{y}, \tilde{\rho})}{h},$$

where we write the perturbed strategy profile as $(\tilde{x}_l + h\delta x_l, \tilde{x}_{-l}, \tilde{y}, \tilde{\rho})$ for some arbitrarily small $h > 0$, since only the state of agent l is perturbed. For convenience, we will use the notation $\Delta x_l \triangleq (\tilde{x}_l + h\delta x_l, \tilde{x}_{-l}, \tilde{y}, \tilde{\rho})$. Note also that the vector δx_l determines the direction of the perturbation, and assume for now that $\tilde{x}_l + h\delta x_l \in \Delta(|\mathcal{A}_i|)$.

The sth entry of the expected reward vector of agent i at Δx_l is

$$\overline{r}_{is}(\Delta x_l) = [(1-\lambda)\tilde{x}_{is} + \lambda/\left|\mathcal{A}_i\right|] \cdot \overline{v}_i(s, \Delta x_l),$$

where $\overline{v}_i(\cdot, \Delta x_l)$ is the conditional reward (4.4) of agent i evaluated at Δx_l. We can write

$$\overline{v}_i(s, \Delta x_l) = \overline{v}_i(s, (\tilde{x}, \tilde{y}, \tilde{\rho})) + h(1+\gamma_l)(1-\lambda) \sum_{q \in \mathcal{A}_i} \overline{v}_i(s, (\tilde{x}, \tilde{y}, \tilde{\rho}) | \alpha_l = q) \delta x_{lq},$$

where $\overline{v}_i(s, (\tilde{x}, \tilde{y}, \tilde{\rho}))$ is the conditional reward (4.4) evaluated at the equilibrium $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\rho})$, which from now on we simply denote by $\overline{v}_i(s, \tilde{z})$. Also, $\overline{v}_i(s, (\tilde{x}, \tilde{y}, \tilde{\rho}) | \alpha_l = q)$ denotes the conditional reward (4.4) evaluated at the equilibrium $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{\rho})$ given also that agent l selects action q.

The expected component sum of the reward vector is

$$\overline{R}_i(\Delta x_l) \triangleq \sum_{s \in \mathcal{A}_i} \overline{r}_{is}(\Delta x_l),$$

which implies that the sth entry of the vector field of agent i at Δx_l will be

$$\overline{g}_{is}(\Delta x_l) = \overline{r}_{is}(\Delta x_l) - \overline{R}_i(\Delta x_l) \cdot \tilde{x}_{is}.$$

If we replace the expected reward terms in the above expression, divide by h, and take the limit as $h \to 0$, we take a quantity that is factored by λ, so that

$$A_{il}^{\lambda,\gamma} = \lambda W_{il}^{\lambda,\gamma}, \quad l \in \mathcal{I} \backslash i,$$

for some matrix $W_{il}^{\lambda,\gamma}$ whose entries are of order of the reward function and it satisfies $\lim_{\lambda \to 0} \lambda W_{il}^{\lambda,\gamma} = 0$.

We now compute the diagonal blocks of the Jacobian matrix, $A_{ii}^{\lambda,\gamma}$, $i \in \mathcal{I}$, that is

$$A_{ii}^{\lambda,\gamma} = \lim_{h \to 0} \frac{\overline{g}_i(\tilde{x}_i + h\delta x_i, \tilde{x}_{-i}, \tilde{y}, \tilde{\rho})}{h},$$

where the perturbed strategy profile is $(\tilde{x}_i + h\delta x_i, \tilde{x}_{-i}, \tilde{y}, \tilde{\rho})$ for some arbitrarily small $h > 0$, since only the state of agent i is perturbed. Again, for convenience, we will use the notation $\Delta x_i \triangleq (\tilde{x}_i + h\delta x_i, \tilde{x}_{-i}, \tilde{y}, \tilde{\rho})$.

In this case, the expected reward of agent i is

$$\overline{r}_{is}(\Delta x_i) = [(1-\lambda)(\tilde{x}_{is} + (1+\gamma_i)h\delta x_{is}) + \lambda / |\mathcal{A}_i|]\overline{v}_i(s, \tilde{z}),$$

for all $s \in \mathcal{A}_i$. Also, the component sum of the expected reward vector is $\overline{R}_i(\Delta x_i) = \sum_{s \in \mathcal{A}_i} \overline{r}_{is}(\Delta x_i)$. The corresponding entries of the vector field will be

$$\begin{aligned}\overline{g}_{ij^*}(\Delta x_i) = &-[1 + (1+\gamma_i)h\delta x_{ij^*(i)}]\overline{v}_i(j^*, \tilde{z})h\delta x_{ij^*} \\ &- \sum_{q \in \mathcal{A}_i \backslash \{j^*\}} (1+\gamma_i)h\delta x_{iq}\overline{v}_i(q, \tilde{z})(1 + h\delta x_{ij^*}) + \lambda\times,\end{aligned}$$

and, for any $s \in \mathcal{A}_i \backslash \{j^*\}$,

$$\begin{aligned}\overline{g}_{is}(\Delta x_i) = &[-\overline{v}_i(j^*, \tilde{z}) + (1+\gamma)\overline{v}_i(s, \tilde{z})]h\delta x_{is} \\ &- \sum_{q \in \mathcal{A}_i} (1+\gamma_i)h\delta x_{iq}\overline{v}_i(q, \tilde{z})h\delta x_{is} + \lambda\times,\end{aligned}$$

where $\times$ denotes quantities that are factored by terms of the form $h\delta x_{is}$, $s \in \mathcal{A}_i$.

If we divide by h and take the limit as $h \to 0$, we conclude that the Jacobian matrix $A_{ii}^{\lambda,\gamma}$ can be written as

$$A_{ii}^{\lambda,\gamma} = A_{ii}^{\gamma} + \lambda W_{ii}^{\lambda,\gamma},$$

where $W_{ii}^{\lambda,\gamma}$ is a matrix whose entries are of order of the reward function, therefore bounded, and it satisfies $\lim_{\lambda\to 0} \lambda W_{ii}^{\lambda,\gamma} = 0$. Also,

$$A_{ii}^{\gamma} = U_i \begin{pmatrix} -\overline{v}_i(j^*, \tilde{z}) & \text{row}\{-(1+\gamma_i)\overline{v}_i(s,\tilde{z})\}_{s\neq j^*} \\ 0 & \text{diag}\{-\overline{v}_i(j^*,\tilde{z}) + (1+\gamma_i)\overline{v}_i(s,\tilde{z})\}_{s\neq j^*} \end{pmatrix} U_i^{\mathrm{T}}$$

for some unitary matrix U_i in $\mathbb{R}^{|\mathcal{A}_i|\times|\mathcal{A}_i|}$ which rearranges the sequence of states so that action j^* corresponds to the first row.

Regarding the evaluation of matrix $B^{\lambda,\gamma}$, it is straightforward to show that its non-diagonal blocks are factored by λ. Its diagonal blocks are defined by

$$B_{ii}^{\lambda,\gamma} \triangleq \lim_{h\to 0} \frac{\overline{g}_i(\tilde{x}, \tilde{y}_i + h\delta y_i, \tilde{y}_{-i}, \tilde{\rho})}{h},$$

where the perturbed strategy profile is $(\tilde{x}, \tilde{y}_i + h\delta y_i, \tilde{y}_{-i}, \tilde{\rho})$, which will be denoted by Δy_i. In this case, the sth entry of the expected reward function is given by

$$\overline{r}_{is}(\Delta y_i) = [(1-\lambda)(\tilde{x}_{is} - \gamma_i h\delta y_{is}) + \lambda/\left|\mathcal{A}_i\right|]\overline{v}_i(s,\tilde{z}).$$

We conclude that the Jacobian matrix $B_{ii}^{\lambda,\gamma}$ can be written in the following form:

$$B_{ii}^{\lambda,\gamma} = B_{ii}^{\gamma} + \lambda V_{ii}^{\lambda,\gamma},$$

where $V_{ii}^{\lambda,\gamma}$ is a matrix whose entries are of the order of the reward function, and it

satisfies $\lim_{\lambda \to 0} \lambda V_{ii}^{\lambda,\gamma} = 0$. Also,

$$B_{ii}^{\gamma} = U_i \begin{pmatrix} 0 & \text{row}\{\gamma_i \overline{v}_i(s, \tilde{z})\}_{s \neq j^*} \\ 0 & \text{diag}\{-\gamma_i \overline{v}_i(s, \tilde{z})\}_{s \neq j^*} \end{pmatrix} U_i^{\mathrm{T}}.$$

As far as the partial derivatives of $\overline{g}$ with respect to ρ are concerned, it is straightforward to show that are zero. In particular, any perturbation of ρ about an equilibrium point $(\tilde{x}, \tilde{y}, \tilde{\rho})$ perturbs only the feedback gain $\gamma_i(\cdot)$. However, the feedback gain multiplies $x - y$, which at an equilibrium is identically zero, and therefore any perturbation of ρ does not perturb $\overline{g}$.

The second part of the vector field, $x - y$, is linear and the computation of the partial derivatives is straightforward. It is also straightforward to show that partial derivative of the last part of the vector field $\overline{R}(z) - \rho$ with respect to ρ is simply $-I$.

So far, we have not taken into account that the perturbations δx_i and δy_i belong to the invariant subspace of the unit vector. Consider an orthonormal matrix N as defined by (4.11) and (4.12). Then, the linearization (4.10) has the system matrix of (4.13).

Index

References

[Ale00] J. M. Alexander. "Evolutionary Explanations of Distributive Justice." *Philosophy of Science*, **67**(3):490–516, 2000.

[AM88] R. Aumann and R. Myerson. *Endogenous formation of links between players and coalitions: An application of the Shapley value*, pp. 175–191. In Roth, A. (ed.), *The Shapley Value*, Cambridge University Press, 1988.

[Art93] W. B. Arthur. "On designing economic agents that behave like human agents." *Journal of Evolutionary Economics*, **3**:1–22, 1993.

[ASS02] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. "A Survey on Sensor Networks." *IEEE Communications Magazine*, pp. 102–114, August 2002.

[BE81] D. J. Baker and A. Ephremides. "The Architectural Organization of a Mobile Radio Network via a Distributed Algorithm." *IEEE Transactions on Communications*, **COM-29**(11):1694–1701, 1981.

[BG00] V. Bala and S. Goyal. "A noncooperative model of network formation." *Econometrica*, **68**(5):1181–1229, 2000.

[BHO05] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis. "Convergence in Multiagent Coordination, Consensus, and Flocking." In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 2996–3000, Seville, Spain, December 2005.

[BL96] J. Bergin and B. L. Lipman. "Evolution with state-dependent mutations." *Econometrica*, **64**(4):943–956, 1996.

[BL03] P. Bonacich and T. M. Liggett. "Asymptotics of a matrix valued Markov chain arising in sociology." *Stochastic Processes and their Applications*, **104**:155–171, 2003.

[BLR03] D.M. Blough, M. Leoncini, G. Resta, and P. Santi. "The k-neigh protocol for symmetric topology control in ad-hoc networks." In *MobiHoc*, 2003.

[Blu93] L. E. Blume. "The Statistical Mechanics of Strategic Interactions." *Games and Economic Behavior*, **5**:387–424, 1993.

[Blu03] L. E. Blume. "How noise matters." *Games and Economic Behavior*, **44**:251–271, 2003.

[BRW04] M. Burkhart, P. von Rickenbach, R. Wattenhoffer, and A. Zollinger. "Does Topology Control Reduce Interference?" In *Proc. of the 5th ACM International Symposium on Mobile Ad-Hoc Networking and Computing*, 2004.

[BS92] K. Binmore and L. Samuelson. "Evolutionary stability in repeated games played by finite automata." *Journal of Economic Theory*, **57**:278–305, 1992.

[BS97] K. Binmore and L. Samuelson. "Muddling Through: Noisy Equilibrium Selection." *Journal of Economic Theory*, **74**:235–265, 1997.

[BST04] M. Benaim, S.J. Schreiber, and P. Tarres. "Generalized urn models of evolutionary processes." *The Annals of Applied Probability*, **14**(3):1455–1478, 2004.

[BV04] V. Bhaskar and F. Vega-Redondo. "Migration and the evolution of conventions." *Journal of Economic Behavior & Organization*, **3**:397–418, 2004.

[CM00] S. Currarini and M. Morelli. "Network formation with sequential demands." *Review of Economic Design*, **5**:229–249, 2000.

[CM05] I. K. Cho and A. Matsui. "Learning aspiration in repeated games." *Journal of Economic Theory*, **124**:171–201, 2005.

[CS07] G. C. Chasparis and J. S. Shamma. "Distributed Dynamic Reinforcement of Efficient Outcomes in Multiagent Coordination." In *Proc. European Control Conference*, Kos, Greece, July 2007.

[DGJ04] E. Droste, R. P. Gilles, and C. Johnson. "Evolution of Conventions in Endogenous Social Networks." *Journal of Economic Literature*, 2004.

[DJ00] B. Dutta and M. Jackson. "The stability and efficiency of directed communication networks." *Review of Economic Design*, **5**:251–272, 2000.

[DM97] B. Dutta and S. Mutuswami. "Stable Networks." *Journal of Economic Theory*, **76**:322–344, 1997.

[DPP06] M. Damian, S. Pandit, and S. Pemmaraju. "Local Approximations Schemes for Topology Control." In *Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing*, 2006.

[DZ93] A. Dembo and O. Zeitouni. *Large Deviations Techniques*. Jones and Barlett Publishers, London, England, 1993.

[Ell93] G. Ellison. "Learning, local interaction, and coordination." *Econometrica*, **61**:1047–1071, 1993.

[Ely02] Jeffrey C. Ely. "Local Conventions." *Advance in Theoretical Economics*, **2**(1):1–29, 2002.

[FFM05] M. Farach-Colton, R.J. Fernandes, and M.A. Mosteiro. "Bootstrapping a Hop-Optimal Network in the Weak Sensor Model." In *Proc. 13th Annual European Symposium on Algorithms (ESA)*, volume 3669, 2005.

[FL98] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

[Fri01] N. E. Friedkin. "Norm formation in social influence networks." *Social Networks*, **23**:167–189, 2001.

[FW84] M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*. Springer-Verlag, New York, NY, 1984.

[FY90] D. Foster and H. P. Young. "Stochastic Evolutionary Game Dynamics." *Theoretical Population Biology*, **38**:219–232, 1990.

[GG97] B.M. Galesloot and S. Goyal. "Costs of flexibility and equilibrium selection." *Journal of Mathematical Economics*, **28**:249–264, 1997.

[GV05] S. Goyal and F. Vega-Redondo. "Network formation and social coordination." *Games and Economic Behavior*, **50**:178–207, 2005.

[Hoj04] D. Hojman. "Interaction Structure and the Evolution of Conventions." 2004.

[HS88] J. C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA, 1988.

[Jac03] M. Jackson. "A survey of models of network formation: Stability and Efficiency." January 2003.

[JLM03] A. Jadbabaie, J. Lin, and S. A. Morse. "Coordination of groups of mobile agents using nearest neighbor rules." *IEEE Transactions on Automatic Control*, **48**(6):988–1001, June 2003.

[Joh86] E. C. Johnsen. "Structure and process: aggreement models for friendship formation." *Social Networks*, **8**:257–306, 1986.

[JW96] M. Jackson and A. Wolinsky. "A strategic model of social and economic networks." *Journal of Economic Theory*, **71**:44–74, 1996.

[JW02a] M. Jackson and A. Watts. "The evolution of social and economic networks." *Journal of Economic Theory*, **106**(2):265–295, 2002.

[JW02b] M. Jackson and A. Watts. "On the formation of interaction networks in social coordination games." *Games and Economic Behavior*, **41**:265–291, 2002.

[KMR93] M. Kandori, G. Mailath, and R. Rob. "Learning, mutation, and long-run equilibria in games." *Econometrica*, **61**:29–56, 1993.

[KY97] Harold J. Kushner and G. George Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag New York, Inc., 1997.

[Lew02] D. Lewis. *Convention: A Philosophical Study*. Blackwell Publishing, 2002.

[LRW08] T. Locher, P. von Rickenbach, and R. Wattenhofer. "Sensor Networks Continue to Puzzle: Selected Open Problems." In *International Conference of Distributed Computing and Networking*, 2008.

[LSW05] X.Y. Li, W.Z. Song, and W. Wan. "A Unified Energy Efficient Topology for Unicast and Broadcast." In *Proceedings of the 11th Int. Conf. on Mobile Computing and Networking*, 2005.

[Mor05] L. Moreau. "Stability of Multiagent Systems with Time-Dependent Communication Links." *IEEE Transactions on Automatic Control*, **50**(2):169–182, February 2005.

[MSS01] G. Mailath, L. Samuelson, and A. Shaked. "Endogenous Interactions." In A. Nicita and U. Pagano, editors, *Evolution of Economic Diversity*. Routledge, New York, NY, 2001.

[Mye77] R. Myerson. "Graphs and Cooperation in Games." *Math. Operations Research*, **2**:225–229, 1977.

[Mye91] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, MA, 1991.

[NH76] M. B. Nevelson and R. Z. Hasminskii. *Stochastic Approximation and Recursive Estimation*. American Mathematical Society, Providence, RI, 1976.

[Nor68] M. F. Norman. "On Linear Models with Two Absorbing States." *Journal of Mathematical Psychology*, **5**:225–241, 1968.

[NP94] K. Najim and A. S. Poznyak. *Learning Automata: Theory and Applications*. Pergamon Press Inc, 1994.

[NT89] K. Narendra and M. Thathachar. *Learning Automata: An introduction*. Prentice-Hall, 1989.

[Oec99] J. Oechssler. "Competition among conventions." *Mathematical and Computational Organization Theory*, **5**:31–44, 1999.

[OFM07] R. Olfati-Saber, J. A. Fax, and R. M. Murray. "Consensus and cooperation in networked multi-agent systems." In *Proceedings of the IEEE (to appear)*, 2007.

[Olf06] R. Olfati-Saber. "Flocking for multi-agent dynamic systems: algorithms and theory." *IEEE Transactions on Automatic Control*, **51**(3):401–420, 2006.

[Pem90] R. Pemantle. "Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations." *The Annals of Probability*, **18**(2):698–712, 1990.

[Rob90] A.J. Robson. "Efficiency in evolutionary games: Darwin, Nash and the secret handshake." *Journal of Theoretical Biology*, **144**:379–396, 1990.

[Rob93] A. Robson. *The Adam and Eve Effect and Fast Evolution of Efficient Equilibria in Coordination Games*. mimeo, 1993.

[RV96] A. Robson and F. Vega-Redondo. "Efficient Equilibrium Selection in Evolutionary Games with Random Matching." *Journal of Economic Theory*, **70**:65–92, 1996.

[SA05] J. S. Shamma and G. Arslan. "Dynamic Fictitious Play, Dynamic Gradient Play, and Distributed Convergence to Nash Equilibria." *IEEE Transactions on Automatic Control*, **50**(3):312–327, March 2005.

[Sam97] L. Samuelson. *Evolutionary Games and Equilibrium Selection*. The MIT Press, Cambridge, MA, 1997.

[San05] P. Santi. *Topology Control in Wireless Ad Hoc and Sensor Networks*. Wiley, 2005.

[SB98] R.S. Sutton and A.G. Barto. *Reinforcement learning: an introduction*. MIT Press, Cambridge, MA, 1998.

[Sch06] T. Schelling. *The Strategy of Conflict*. Harvard University Press, 2006.

[SGA00] K. Sohrabi, J. Gao, V. Ailawadhi, and G. J. Pottie. "Protocols for self-organization of a wireless sensor network." *Personal Communications, IEEE*, **7**(5):16–27, 2000.

[Sky02] B. Skyrms. "Signals, Evolution and the Explanatory Power of Transient Information." *Philosophy of Science*, **69**:407–428, 2002.

[Sky04] B. Skyrms. *The Stag-Hunt and the Evolution of the Social Structure*. Cambridge University Press, New York, NY, 2004.

[Sky07] B. Skyrms. "Dynamic Networks and the Stag Hunt: Some Robustness Considerations." *Biological Theory*, **2**(1):7–9, 2007.

[Smi82] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.

[SN69] I. J. Shapiro and K. S. Narendra. "Use of Stochastic Automata for Parameter Self-Organization with Multi-Modal Performance Criteria." *IEEE Transactions on Systems Science and Cybernetics*, **5**:352–360, 1969.

[SN00] M. Slikker and A. van den Nouweland. "Network formation models with costs for establishing links." *Review of Economic Design*, **5**:333–362, 2000.

[SP00] B. Skyrms and R. Pemantle. "A dynamic model of social network formation." *Proc. of the National Academy of Sciences of the USA*, **97**:9340–9346, 2000.

[Str03] S. H. Strogatz. *Sync: The Emerging Science of Spontaneous Order*. Hyperion, New York, NY, 2003.

[SWL04] W. Song, Y. Wang, X. Li, and O. Frieder. "Localized Algorithms for Energy Efficient Topology in Wireless Ad Hoc Networks." In *MobiHop*, pp. 98–108, Roppongi, Japan, May 2004.

[VV63] V. I. Varshavskii and I. P. Vorontsova. "On the Behaviour of Stochastic Automata with a Variable Structure." *Automation and Remote Control*, **24**:327–333, 1963.

[Wat01] A. Watts. "A Dynamic Model of Network Formation." *Games and Economic Behavior*, **34**:331–341, 2001.

[Wil02] N. Williams. "Stability and Long Run Equilibrium in Stochastic Fictitious Play." 2002.

[You93] H. P. Young. "The evolution of conventions." *Econometrica*, **61**:57–84, 1993.

[You98] H. P. Young. *Individual Strategy and Social Structure*. Princeton University Press, Princeton, NJ, 1998.

[You04] H. P. Young. *Strategic Learning and Its Limits*. Oxford University Press, New York, NY, 2004.

[Zeg94] E. Zeggelink. "Dynamics of structure: an individual oriented approach." *Social Networks*, **16**:295–333, 1994.

[Zeg95] E. Zeggelink. "Evolving friendship networks: an individual-oriented approach implementing similarity." *Social Networks*, **17**:83–110, 1995.