POLITECNICO DI TORINO

Corso di Laurea in Physics of Complex Systems

Tesi di Laurea Magistrale

Will we ever learn?

An attempt to clarify convergence of learning in games

Relatore
prof. Luca Dall'Asta
Correlatore:
prof. J. Doyne Farmer

Luca Mungo
matricola: 242249

Anno accademico 2017 – 2018

Summary

In this work, we have addressed the problem of understanding which features of a game influence the convergence of a learning algorithm. We focused on the relation between the empirical convergence frequency for a given game and its best reply structure. We tried to broaden this framework and obtain more precise predictions by including quasi-best replies. In order to do so, we developed an analogy between the execution of a game by two players and a diffusive process on a fully connected weighted graph. We looked at the stationary distribution of such a process and tried to see if it could be used to calibrate the strength (e.g to infer the relative size) of each attractor. The aforementioned analogy was based on a logit one parameter (β) transformation that mapped the payoff matrix in a stochastic one. We ran extensive simulations to see how our predictions performed as β varied. Unfortunately, the new framework seemed to give little or no improvement to the old one. Arguing that the problem was choosing the same value of β for all the games included in the simulation, we introduced the notion of optimal beta β^* and tried to see whether this could be directly inferred from the payoff matrix. To this aim, we developed and tested four different measures that were coherent with the properties of the learning algorithm. We finally shared some thoughts about why these methods failed and which issues they are not able to cope with.

Acknowledgements

I would like to express my gratitude to the *Complexity Economics Programme* research group of the Institue of New Economic Thinking at the Oxford Martin School, which welcomed me and allowed me to discover an interesting field of research I had no previous knowledge on. I especially thank my supervisor there, prof. Doyne Farmer, who gave me this opportunity as well as encouragement and guidance. My deep gratitude also goes to D Phil student Marco Pangallo, who put his faith in me, arranged my visit, helped and sustained me throughout my whole time in Oxford. I would also like to say thank you to Dr. Torsten Heinrich for the helpful discussions and to all the other members of the group for making these months so enjoyable and stimulating.

I would then like to thanks my parents, for the relentless love and support. Thanks to my sister, who never has a doubt on me. And finally thanks to all the friends that have come with me through this long journey that was graduating, and especially to Giacomo and Ludovico, for the thoughtfulness and the good company of these last two years - looking back on it now, it has been a little miracle.

Contents

List of Figures 5

1 Theoretical Remarks 9
- 1.1 Game Theory . . . 9
- 1.2 Experience-Weighted Attraction Learning Algorithm . . . 11
- 1.3 Best Reply Structure . . . 14
- 1.4 Markov Processes . . . 16
- 1.5 Previous Work . . . 18

2 Research Work 23
- 2.1 A new estimate for the attractors' size . . . 23
  - 2.1.1 Softening the booleanization . . . 23
  - 2.1.2 A new measure for the attractors' size . . . 24
  - 2.1.3 Results . . . 30
- 2.2 Predicting the optimal β . . . 34

3 Conclusions 41

Bibliography 53

List of Figures

1.1 15
1.2 Example of a convergent EWA learning dynamics 19
1.3 Example of EWA learning dynamics in a cycle 19
1.4 21
2.1 26
2.2 An example of the process 29
2.3 Behaviour of R^2 31
2.4 Distance between transformed and boolean matrices 32
2.5 Behaviour of R^2 excluding matrices with only cycles 33
2.6 Results obtained by using the different metrics 36
2.7 38
2.8 39
2.9 40
3.1 46
3.2 Occurrence of cycles as a function of the number of strategies 50

Introduction

The concept of equilibrium permeates mainstream economic theory on many different levels; it is not of simple definition and can't be easily captured in a formula as it often happens in natural sciences, nevertheless it could be - vaguely - defined as a state in which beliefs match outcomes.

To be more precise, we must restrict our focus on specific areas. In microeconomic theory, for example, this intuition (called *Walrasian*[3] equilibrium in honour of *Leon Walras* who first formalized it and showed its existence) is declined as the assertion that demand and supply are perfectly balanced: both producers and customers manage to agree on the same prices for all the goods, it is produced nothing more than what is required, and in this sense beliefs match outcomes. The proof of the existence of such a state is a strong result and lies at the core of what is known as the *General Equilibrium Theory*, describing the formation and evolution of economic variables (like prices) from the interaction of a set of markets.

Another field where this concept is pervasive is *Game Theory*. Game theory is the mathematical study of interaction among independent, self-interested agents [4]. It was first introduced by J. Von Neumann et O. Morgenstern in 1944 [2], it has gone through a dramatic increase of interest in many fields and is now intensively studied in disciplines as diverse as political science, biology, psychology, economics, linguistics, sociology and computer science [5, 6, 7, 8].

In this framework, as we will see later, an equilibrium is a state where the expectations of one player (and consquently his actions, that are always a best reply to such beliefs) regarding the others' behaviors indeed match the choices of his rivals: once again essentially beliefs match outcomes. Following the pioneering work by John Nash [9], who proved that such a state exists for any game with a finite number of players, the traditional approach is based on equilibrium (also known as *Nash equilibrium*) and on the assumption that all the players can instantly coordinate to it.

In the example we provided, the existence of the equilibrium is just half of the story: for both theories to have a real value, it should be also shown that agents are capable of finding it. In microeconomic theory Walras himself devised a method, known as *tâtonemment*, through which all markets manage to balance demand and supply and economic agents maximize their utilities or profits. An equivalent reflection for Game Theory only

arose some years after Nash's seminal papers and took essentially two directions. The first is the field of *Algorithmic Game Theory*, which investigates the computational complexity of computing a Nash equilibrium[11]. The second is based on the assumption that players learn the equilibrium by repeating the game. On this intuition, a number of different *learning algorithm* were devised with the aim of mimicking the way actual humans learn. This seems to be a reasonable, *reality-grounded* approach, but opens a new, interesting question: when do this algorithms converge to a solution? Or equivalently, under which circumstances are players able to learn?

There is not a unique way to address the problem. Analytically, some results have been obtained for some subclasses of games (so-called dominance-solvable[12], coordination[13], potential[14], supermodular[15] or "weakly acyclic"[16] games). Another field of research applied *Non Linear Dynamical Systems Theory* to learning algorithms.
In this framework, a pioneering paper by T. Galla and J.D. Farmer [18] addressed the issue by studying *ensembles* of randomly-generated games and trying to estimate the average behavior of the expected convergence frequency in the parameter space[1]. By studying ensembles of 2-person, N-strategies games were able to divide such space into different regions: in one of them the dynamics of learning was essentially chaotic, meaning that players never manage to identify the equilibrium of the game. However, little understanding of the reasons for this erratic behaviour was provided.
In this direction, and always taking the investigation of ensembles of random games as a starting point, a further work [19] introduced a formalism that depends on a basic property of the game. This is its *best reply structure*, namely the set of cycles or fixed points that the players can get stuck into by myopically responding to each other with their *best replies*. The authors show that their formalism predicts non-convergence of several learning algorithms that have been used in behavioral economics and population biology. Their estimate works particularly well for *boolean* payoff matrices (which we will define later), but is not as efficient for normal ones. In this thesis, we tried to understand where this mismatch comes from and if the gap can be filled.

[1]The idea of studying randomly generated games was not new to the physics community. In [30] for example, the authors tried to approach game theory from a *Statistical Physics* point of view via the theory of *disordered systems*. It was nevertheless the first time where simulations were launched on ensembles of random games in the attempt of devising a phase space for the learning algorithm

Chapter 1

Theoretical Remarks

We report here a few notions that are essential to the following discussion.

1.1 Game Theory

Game theory studies the interactions among self-interested agents. By self-interested we don't necessarily mean that they want to cause harm to each other neither that they only care about themselves. We just assume that each agent "has his own description of which states of the world he likes — which can include good things happening to other agents — and that he acts in an attempt to bring about these states of the world"[4]. Agents' interests are modelled via a *utility function*, a mapping from states of the world to real numbers. These numbers are interpreted as measures of an agent's level of happiness in the given states. When the agent is uncertain about which state of the world he faces, his utility is defined as the expected value of his utility function with respect to the appropriate probability distribution over states.
When utility functions are assigned to agents, acting optimally in an uncertain environment is conceptually straightforward, but things can get considerably more complicated when the world contains *two or more* utility-maximizing agents whose action can affect each others' utilities. This setting is investigated in *noncooperative* game theory.

We begin by giving the definition of a normal form game, the most familiar representation of strategic interactions in game theory; "a game written in this way amounts to a representation of every player's utility for every state of the world, in the special case where states of the world depend only on the players' combined actions"[4]. Many other representations of interest can be reduced to the normal form, making it arguably the most fundamental in game theory.

Definition 1 (Normal Form Game) *A (finite, n-person) normal-form game is a tuple (N, A, u), where:*

- *N is a finite set of n players, indexed by i;*

- $A = A_1 \times \ldots \times A_n$, *where* A_i *is a finite set of actions available to player* i. *Each vector* $a = (a_1, \ldots, a_n) \in A$ *is called an action profile;*

- $u = (u_1, \ldots, u_n)$ *where* $u_i : A \to \mathbb{R}$ *is a real-valued utility (or payoff) function for player* i

A natural way to represent games is via an n-dimensional matrix. In 2-players games, each row corresponds to a possible action for player 1 (also called *Row* or R), each column corresponds to a possible *action* or *move* for player 2 (also called *Column* or C), and each cell corresponds to one possible outcome. Each player's utility for an outcome is written in the cell corresponding to that outcome, with player 1's utility listed first. Here a simple example, the game of paper, rock, scissor:

$$
\begin{array}{cc} & \text{C} \\ & \begin{array}{ccc} r & p & s \end{array} \\ R \begin{array}{c} r \\ p \\ s \end{array} & \begin{pmatrix} (0,0) & (-1,1) & (1,-1) \\ (1,-1) & (0,0) & (-1,1) \\ (-1,1) & (1,-1) & (0,0) \end{pmatrix} \end{array}
$$

Although this is the most compact way to represent the game, in our study this matrix will always be deconstructed in the two matrices Π^R and Π^Ccontaining respectively only the payoffs for player R and C. It is important to stress that when considering the payoff matrix of a single player, the rows always correspond to the possible actions for that player, columns being the choices available to his opponent.
Let us now introduce *strategies*. *Strategies* can be considered as the available choices for the player. One kind of strategy is to select a single action[1] and play it. We call such a strategy a *pure strategy*. We call a choice of pure strategy for each agent a *pure-strategy profile*.
Players could also follow another, less obvious type of strategy: randomizing over the set of available actions according to some probability distribution. Such a strategy is called a *mixed strategy*. We define a mixed strategy for a normal-form game as follows.

Definition 2 (Mixed Strategy) *Let* (N, A, u) *be a normal-form game, and for any set* X *let* $\prod(X)$ *be the set of all the probability distributions over* X. *Then the set of mixed strategies for player* i *is* $S_i = \prod(A_i)$. *The set of mixed-strategy profiles is simply the cartesian productof the individual mixed-strategy sets* $S = S_1 \times \ldots \times S_n$. *By* $s_i(a_i)$ *we denote the probability that an action* a_i *will be played under mixed strategy* s_i.

[1] In this work we will use *action* and *move* as synonyms

Given a mixed strategy profile, its *expected utility* is defined as follows:

Definition 3 (Expected utility of a mixed strategy) *Given a normal-form game* (N, A, u)*, the expected utility (or expected payoff)* u_i *for player* i *of the mixed-strategy profile* $s = (s_1, \ldots, s_n)$ *is defined as:*

$$u_i(s) = \sum_{a \in A} u_i(a) \prod_{j=1}^{n} s_j(a_j)$$

This notation can be semplified in the case of 2-players games by renaming $s_1(a_j)$ as x_j and $s_2(a_j)$ as y_j, or writing in general $s_i(a_j) = s_j^i$.

If an agent knew how the others were going to play, his strategic problem would become simple: it would essentially become a single-agent problem of choosing a utility-maximizing action. We define $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, a strategy profile s without agent i's strategy, and can write $s = (s_i, s_{-i})$. If the agents other than i (whom we will collectively denote as $-i$) were to commito to play s_{-i}, a utility maximizing agent i would face the problem of determining his best response.

Definition 4 (Best response) *Player* i*'s best response to the strategy profile* s_{-i} *is a mixed strategy* $s_i^* \in S_i$ *s.t.* $u_i(s_i^*, s_{-i}) \geq u_i(s_i, s_{-i})$ *for all strategies* $s_i \in S_i$

Definition 5 (Nash equilibrium) *A strategy profile* $s = (s_1, \ldots, s_n)$ *is a Nash equilibrium if, for all agents* i*,* s_i *is a best response to* s_{-i}*.*

A Nash equilibrium is a *stable* strategy profile: no agent would be better off by changing his strategy if he knew what strategies the other agents were following. If the strategy of each player in the equilibrium is a pure one than we will have a *Pure Strategy Nash Equilibrium*, otherwise we will have a *Mixed Strategy* one.

For a more detailed introduction see [4].

1.2 Experience-Weighted Attraction Learning Algorithm

As we said, Game Theory originally assumed that players could instantly coordinate on equilibrium. Later on, a number of behaviorally plausible learning algorithms were developed in order to mimic how humans actually learn. These learning algorithms can be grouped in two classes: in reinforcement learning, players only learn based on the payoff they received; in belief learning, players only consider what the expected action of their opponent is. *Experience-Weighted Attraction* (EWA) has been proposed by Camerer and Ho [20] to generalize reinforcement these two classes. Essentially, players update their strategies by taking into account both their predictions about what the other players will do (as in belief learning) and how their strategies performed in the past (as in reinforcement

learning). The connection between reinforcement and belief learning lies in the update of the moves that were not played, i.e. in considering the *foregone payoffs*. If only the probabilities of the moves that are played are updated, EWA reduces to a simple version of reinforcement learning. If all probabilities are updated with the same weight instead, EWA reduces to a belief learning algorithm (such as *fictitious play* or *best reply dynamics*, depending on the parameters). EWA has been extensively studied by experimental economists who have shown that it provides a reasonable approximation for how real people learn in games and has hence been taken as the reference learning algorithm in this and other ([18]) works.

Formally, the key quantities in EWA algorithm are the so-called *attractions* or *propensities* $Q_i^\mu(t)$ that quantify the appreciation for the move i by player μ at time t.The way to obtain the mixed strategies from them is through a logit model:

$$s_i^\mu(t) = \frac{e^{\beta Q_i^\mu(t)}}{\sum_j e^{\beta Q_j^\mu(t)}} \tag{1.1}$$

Attractions are updated through the equations (here we focus on a 2-players game):

$$Q_i^1(t+1) = \frac{(1-\alpha)\, N(t) Q_i^1(t) + (\delta + (1-\delta)\, x_i(t)) \sum_j \Pi^R(i,j) y_j(t)}{N(t+1)} \tag{1.2}$$

where:

$$N(t+1) = (1-\alpha)\,(1-\kappa)\, N(t) + 1 \tag{1.3}$$

$N(t)$ represents the *experience* as it increases monotonically with the number of rounds played; the more it grows, the smaller will be the influence of each round played on the update rate for the attractions.

As pointed out in [18] under this update rule, player 1 knows his own payoffs and also the frequency y_j with which player 2 makes each of his possible moves (and vice-versa). This approximates the situation in which the players vary their strategies slowly in comparison with the timescale on which they play the game, so that they can collect good statistics about the others before updating their strategy. In the machine learning literature, the practice of infrequent parameter updating is called "batch learning."

Parameter Choice and Convergence Criteria

Following papers [29], [18], we used the following set of parameters:

- $\alpha = 0.18$; sufficiently large to reach convergence in decent simulations' time, but small enough to avoid convergence to a trivial, meaningless fixed point.

- $\beta = \sqrt{N}$, N being the number of moves. The reason is that the expected payoffs $\sum_j \Pi^R(i,j)y_j$ and $\sum_j \Pi^C(i,j)x_j$ scale as $\frac{1}{\sqrt{N}}$; indeed the sum $\sum_j \Pi^\mu(i,j)$ scale as $\sqrt{N}$ due to the Central Limit Theorem (it is a sum of gaussian variables) while y_j and x_j go as $\frac{1}{\sqrt{N}}$ due to the normalization constraint. Therefore, increasing the size of the payoff matrix has the same effect as decreasing β. In the limit $\beta \to 0$ this eventually leads the players to choose uniformly at random between their possible moves, irrespectively of the payoff matrix.

- $\kappa = 1$, in order to keep the experience factor constant and assign the same weight at each round of the game.

- $\delta = 1$, again to avoid convergence to trivial fixed points[2].

To decide whether the process converged we adopted the following procedure (given in [19]). We run the EWA dynamics for 500 time steps and we consider the last 20% time steps to determine convergence. With the parameter values we chose for α, β, κ and δ, the transient is usually of the order of 100 time steps, so 500 time steps is enough to identify convergence. We then check that the average variance of the logarithms of the components of the mixed strategy vectors does not exceed a certain (very small) threshold, 10^{-2}. We look at the logarithms because the probabilities following the EWA dynamics vary on an exponential scale and can be of the order of, e.g, 10^{-100}.

Invariance Property

This choice of parameters makes the update equations invariant under the transformation:

$$\Pi^\mu(i,j) \to \Pi^\mu(i,j) + b_j \tag{1.4}$$

$\bar{b}$ being whatever N-dimensional vector.

We give proof of that for 2-player games (the choise is made to simplify notation; the proof for multiple players goes just the same). Let Q_i^R be the attraction of player Row towards the move i and A be the matrix whose rows are just copies of the vector $\bar{b}$. Then the transformation 1.4 can be rewritten as $\Pi^\mu(i,j) \to \Pi^\mu(i,j) + b_j = \Pi^\mu(i,j) + A(i,j) = \tilde{\Pi}^\mu(i,j)$. The update equations with a similar payoff matrix will be:

[2] See [29] for a more detailed explanation.

$$Q_i^R(t+1) = \frac{(1-\alpha)N(t)Q_i^R(t) + \sum_j \tilde{\Pi}_{i,j}^R y_j}{N(t+1)} =$$

$$\frac{(1-\alpha)N(t)Q_i^R(t) + \sum_j (\Pi_{i,j}^R + A_{i,j})y_j}{N(t+1)} = \tag{1.5}$$

$$Q_i^{old\,R} + \frac{1}{N(t+1)} \sum_j A_{ij} y_j$$

Since A_{ij} has equal rows, the sum $\sum_j A_{ij} y_j$ is independent on i and can be rewritten as a constant c. But then:

$$\tilde{x}_i(t+1) = \frac{e^{\beta Q_i^R(t+1)}}{\sum_j e^{\beta Q_j^R(t+1)}} = \frac{e^{\beta\left(Q_i^{old\,R}(t+1)+c\right)}}{\sum_j e^{\beta\left(Q_j^{old\,R}(t+1)+c\right)}} \tag{1.6}$$

Calling $e^{\beta c} = \alpha$ we will have:

$$\tilde{x}_i = \frac{\alpha e^{\beta Q_i^{old\,R}(t+1)}}{\alpha \sum_j e^{\beta Q_j^{old\,R}(t+1)}} = \frac{e^{\beta Q_i^{old\,R}(t+1)}}{\sum_j e^{\beta Q_j^{old\,R}(t+1)}} = x_i \tag{1.7}$$

Hence, the learning dynamics is not affected by a payoff transformation of the form 1.4.

1.3 Best Reply Structure

A *best reply* is the move that gives the best payoff in response to a given move played by the opponent. When in a 2 players game each player responds to his opponent by always playing his best reply we have a *best reply dynamics*.
The *best reply structure* is the arrangement of the best replies in the payoff matrix.

Figure 1.1

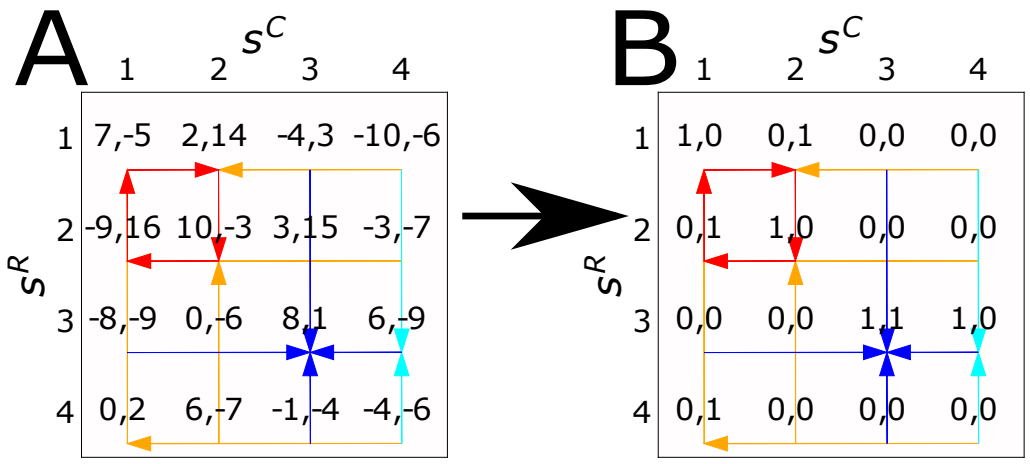


Illustration of the best reply structure. $S^R = 1, 2, 3, 4$ and $S^C = 1, 2, 3, 4$ are the possible moves of players Row and Column and each cell in the matrix represents their payoffs (Row is given first). The best response arrows point to the cell corresponding to the best reply. The vertical arrows correspond to player Row and the horizontal arrows to player Column. The arrows are colored red if they are part of a cycle, orange if they are not part of a cycle but lead to one, blue if they lead directly to a fixed point, and cyan if they lead to a fixed point in more than one step. The payoff matrix in B is the boolean version of matrix A (more details on this in the following). Image and caption from [19]

The best reply structure will then essentially be a collection of fixed points, *free best replies*[3] and cycles of various length; neglecting the free best replies, it can be encoded in a *best reply vector* $\bar{v}$ whose components will be

$$v_i = \#\textit{cycles of length } i$$

where fixed points will be simply considered cycles of length 1.

Cycles and fixed points are *basins of attraction* for the best reply dynamics (once the dynamics enters one of them it can get out) and their size can roughly be estimated by counting the number of actions they contain.

It is important to stress that, while learning algorithms are *synchronous* (strategies for each player are updated simultaneously), best reply dynamics is *asynchronous* (players choose their moves one after the other). In 2-players games, this asynchronicity has anyway no impact on the best reply vector[4].

[3] Best replies that lead to a fixed point or a cycle but are not part of it

[4] Meaning that in two players games one gets the same *best reply structure* independently from which of the two players is the first one to choose

One final remark is that best reply dynamics is deterministic. If we identify the decision to play move i by player 1 with the strategy profile $\bar{x} : x_j = \delta_{i,j}$, then the dynamics of the game will follow the equation:

$$\bar{x}(t+1) = B^R \cdot B^C \cdot \bar{x}(t) \tag{1.8}$$

where B_R and B_C are the *boolean payoff matrices* defined by:

$$B_R(i,j) = \begin{cases} 1 & \text{if } \Pi^R(i,j) = \max_k \left(\Pi^R(i,k)\right) \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

(similar equations hold for strategy $\bar{y}$ of player 2 and B_C)[5].

1.4 Markov Processes

A *stochastic process* is a system which evolves probabilistically in time or more precisely, "a systems in which a certain time-dependent random variable $X(t)$ exists"[22]. We can measure values $x_1, x_2, x_3, \ldots$, etc of $X(t)$ at times $t_1, t_2, t_3, \ldots$ and we assume that a set of joint probability densities exists

$$p(x_1, t_1; x_2, t_2; \ldots)$$

which describes a system completely.

A *Markov chain* is "a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event"[24]. A Markov process, named after the russian mathematician Andrey Markov, is a stochastic process that satisfies the *Markov assumption*; formulated in terms of conditional probabilities, it requires that if we take a sequence of ordered time steps in time $t_1, t_2, \ldots, t_n$ such that:

$$t_1 \leq t_2 \leq \ldots t_n$$

then the conditional probability is determined entirely by the knowledge of the most recent condition, i.e.:

$$p(x_n, t_n | x_{n-1}, t_{n-1}; x_{n-2}, t_{n-2}; x_{n-3}, t_{n-3}; \ldots) = p(x_n, t_n | x_{n-1}, t_{n-1})$$

[5]NB: the booleanization doesn't alter the best reply structure of the game

If the system can be described by a countable number of states, then its evolution can be depicted via a *stochastic* (or *Markov*) matrix S, whose entries S_{ij} represent the probabilities to go from state j to state i at each time step. If one takes the diffusion on a network as an example, then each state will correspond to a node and the entries of the matrix will be the probabilities to go from a node to another at each time step.

In order to be a (column) stochastic one, a $N \times N$ matrix S must satisfy the following conditions:

- $S_{ij} \geq 0$ for all i and j, as each entry is a transition probability
- $\sum_j S_{ij} = 1$ for all i; this is again very intuitive: if one is is on a site i at time t_{n-1}, with probability 1 he will be somewhere at time t_n.

Markov matrices have some interesting properties:

- Any Markox matrix always has an eigenvector with eigenvalue 1.
- If each entry of a Markov matrix A is strictly positive all other eigenvalues have absolute value strictly less than 1.[6]

The dynamics of a Markov Process obeys an evolution equation of the form:

$$\bar{p}\,(t+1) = S\bar{p}\,(t) \tag{1.10}$$

$\bar{p}\,(t)$ being the vector s.t. $\bar{p}_i\,(t)$ = probability of being in state i at time t. Under the conditions of stricly positive entries $S_{ij} > 0\,\forall\, i, j$, it can be proved that $\bar{p}(t)$ will converge in the limit $t \to \infty$ to the eigenvector associated with the eigenvalue 1. We don't give an exact proof here, but just sketch the reasoning. Let's consider the initial vector $\bar{p}\,(0)$. If we decompose it as a linear combination of eigenvalues $\bar{s}_k$ of S we will get:

$$\bar{p}\,(0) = \sum_k c_k^0 \bar{s}_k$$

Plugging it into the evolution equations:

$$\bar{p}\,(1) = S\bar{p}(0) = S\sum_k c_k^0 \bar{s}_k = \sum_k c_k^0 \lambda_k \bar{s}_k$$

[6] We don't give proof of these properties here; the demonstration is essentially based on the properties of the kernel and on *Perron-Frobenius* theorem. They can be found in [23]

And in general:

$$\bar{p}(t) = \sum_k c_k^0 (\lambda_k)^t \bar{s}_k$$

Each component hence grows accordingly to $(\lambda_k)^t$. We know that the largest eigenvalue is 1 and that there is a unique eigenvalue associated to it: in $t \to \infty$ limit the component related to the largest eigenvalue will then be much larger than the others, eventually making them negligible[7]. The eigenvector $\bar{e}_1$ is hence the *stationary probability* for the Markov Process.

1.5 Previous Work

The work done in this thesis is rooted in a paper by *M. Pangallo, T. Heinrich, J. D. Farmer*[19]. In an attempt to challenge the equilibrium assumption, the authors show that the convergence of learning processes over a given game is tied to its best reply structure.

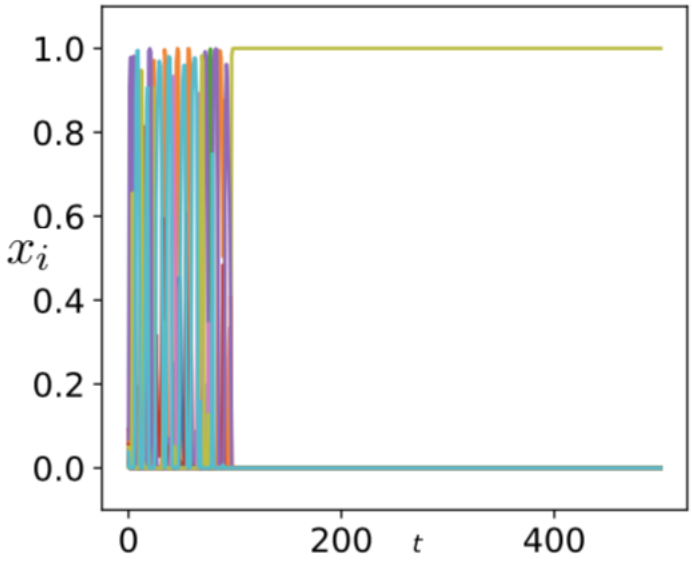
The idea is the following: best reply cycles and fixed points are attractors for the best reply dynamics. Cycles are associated with out-of-equilibrium behaviours, as the players are constantly changing the moves they are playing even if in a periodic fashion, while fixed points (*Pure Strategy Nash Equilibria* in game theory jargon) are associated with equilibrium.

The presence of such structures deeply influences the dynamics of a learning algorithm. Indeed, both in learning processes and best reply dynamics, players are trying to maximize their payoffs, so it is not unreasonable to believe that they can be related; our guess is that the existence of a cycle in the best reply dynamics increases the chances of having a basin of attraction different for the one of the fixed point also for the *EWA* learning dynamics.

To clarify that, we include two examples of EWA learning dynamics. In the first one (Fig 1.2) the dynamics gets stuck in a fixed point. In the second (Fig 1.3), instead, it enters into a cycle. In both cases, the presence of the fixed point/cycle could be predicted by looking at the best reply structure of the game.

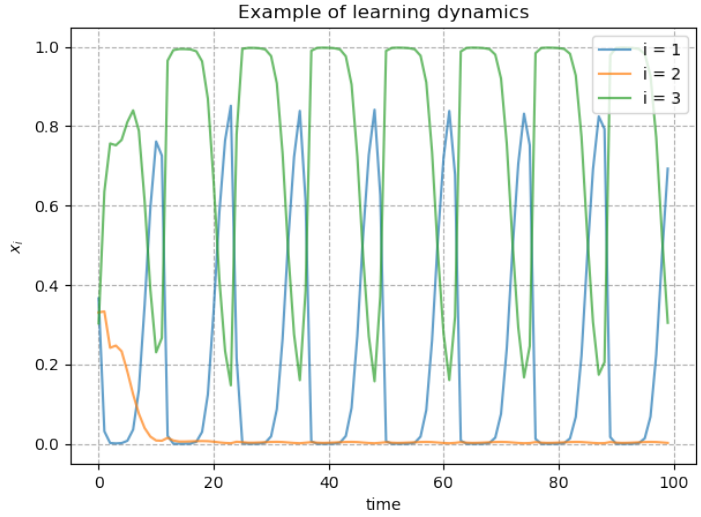
[7] This property was exploited in the simulations in order to find the stationary probability of the process

Figure 1.2: Example of a convergent EWA learning dynamics



Example of a convergent EWA learning dynamics in a 20x20 game with best reply vector $\bar{v} = (1,0,\ldots,0)$.

Figure 1.3: Example of EWA learning dynamics in a cycle



The figure shows an example of EWA learning dynamics caught in a cycle in a 3x3 game. In this case moves 1 and 2 (whose probabilities x_1 and x_2 are respectively the blue and green line in the plot) were part of a cycle of length 2. One can immediately understand in which sense the presence of a best reply cycle influences the dynamics of the learning process. The best reply vector was $\bar{v} = (0,1,0)$: a cycle of length 2 was predicted

An estimate of the ratio between the basin of attractions of the cycles and the basin of attractions of the fixed points could then give a good prediction of the probability of convergence of learning dynamics for a given game. The measure proposed for this goal is:

$$F(\bar{v}) = \frac{\sum_{j=2}^{N} (v_j \times j)}{\sum_{j=1}^{N} (v_j \times j)} \tag{1.11}$$

that estimates the probability of **non convergence** of learning over a game. In this formula:

- N is the number of possible strategies available to each player (equal to the size of the payoff matrix)

- the numerator is equal to the number of actions associated with cycles. This is obtained by multiplying the number of cycles of a given length (v_j) for their length (j) and then by summing over all the possible lengths (of course being N the total number of strategies, it will also be the maximum length for a cycle)[8].

- the denominator is equal to the total number of actions associated with attractors, and is indeed obtained by adding to the numerator the number of fixed points[9]

To sum up:

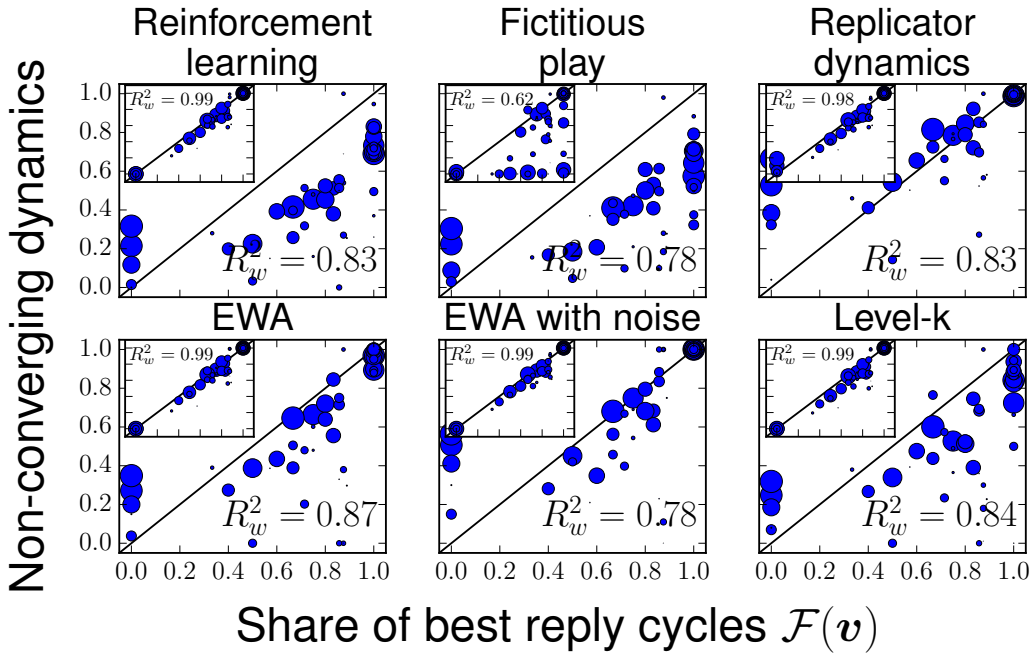
$$F(\bar{v}) = \frac{n_{\textit{moves associated to cycles}}}{n_{\textit{moves associated to cycles}} + n_{\textit{moves associated to fixed points}}} \tag{1.12}$$

The article shows that this prediction has a very good agreement with the empirical convergence probability for a large class of learning algorithms, suggesting a relationship between the attractors of the best reply dynamics and the ones of the learning processes. The prediction becomes essentially exact if learning processes are run on the boolean version of the payoff matrix.

[8]The length of a cycle is the number of different actions that **one** player plays once the dynamics enters in the cycle and not the **total** number of actions played by all players.

[9]that can be considered as cycles of length 1

Figure 1.4



Test for how well the best reply structure predicts non-convergence under several learning algorithms. Each circle corresponds to a specific best reply vector $\bar{v}$ and its size is the logarithm of the number of times a payoff matrix with $\bar{v}$ was sampled. The horizontal axis is the frequency of non-convergence under best reply dynamics F(v). The vertical axis gives the frequency of non-convergence in the simulations, as averaged over all payoff matrices and initial conditions having the same $\bar{v}$. In the insets simulations are based on Boolean approximations of payoff matrices. The identity line is plotted for reference. Image taken from [19]

Where does the mismatch between the two predictions come from? Can we find a way to make correct (or at least better) predictions for learning processes over normal, *non-boolean* games (meaning with this games whose payoff matrices have not gone through transformation 1.9)? In this thesis, we make an attempt to extend the result in this direction.

Chapter 2

Research Work

2.1 A new estimate for the attractors' size

2.1.1 Softening the booleanization

As we said, the probability of (non) convergence was initially estimated by taking the ratio between the size of basins of attractors related to non-equilibrium dynamics over the one of all the attractors of the best reply dynamics. Even when dealing with non boolean games this hypothesis seems plausible, still the way in which the attractor's size is computed might be modified to best capture the characteristics of the actual game.

The idea explored in this thesis relies on the concept of *quasi-best replies*. To give an intuition, imagine a situation in which a player has to choose between two different moves with a very narrow payoff gap. This proximity between the payoff could mislead the player, who could choose the *quasi best reply* (in other words, the move whose payoff is nearly as high as the one of the best reply) eventually leaving an attractor and getting trapped into another one.

We make an example to clarify this idea. Let us take a dummy payoff matrix:

$$\begin{pmatrix} (1,1) & (0,0) & (0,0) \\ (0.99,0) & (0,1) & (1,0) \\ (0,0) & (1,0) & (0,1) \end{pmatrix}$$

From the point of view of the best reply dynamics, there is one fixed point (upper left entry) and a cycle of length 2 (lower right side of the matrix). If the player C chooses to play move 1, then the best reply for R would be to play move 1 as well. Nevertheless, he could make a mistake and choose move 2, as the associated payoff is $= 0.99$, very close to the higher one. Once the mistake is made, if player C plays its best response, the dynamics

would leave the fixed point and enter the cycle.

To quantify this all, we made two observations:

- The boolean payoff matrix B_μ can be obtained by Π^μ through the transformation:

$$B^\mu(i,j) = \lim_{\beta\to+\infty} \frac{e^{\beta\Pi^\mu(i,j)}}{\sum_k e^{\beta\Pi^\mu(i,k)}} = \lim_{\beta\to+\infty} T^\mu_\beta(i,j) \tag{2.1}$$

- For any finite value of β, the equation:

$$\bar{x}(t+1) = T^R_\beta \cdot T^C_\beta \cdot \bar{x}(t) \tag{2.2}$$

  describes a Markov Process ($T^{R/C}_\beta$ are indeed column-stochastic matrices by definition[1]). We recover the best reply dynamics 1.8 in the limit $\beta \to +\infty$. In the following, we will occasionally refer to the matrix $T^R_\beta \cdot T^C_\beta$ as M for simplicity.

Can the stationary probability distribution of such stochastic process [2] help us build a better estimate of the attractors' size?

2.1.2 A new measure for the attractors' size

In the previous model, each cycle was assigned a weight proportional to its length (see 1.11). This essentially meant assigning an equal weight to all the moves. We investigated if any improvement could be achieved by weighting each move with the stationary probability of the Markov process associated with the matrix M.

There are some reasonable arguments to build a similar framework, we briefly look over them.

Trembling hand

As we said, the initial guess was that *quasi best replies* could have an impact on the choices of one player. This is what is known in game theory's jargon as the *trembling hand* [25], namely the possibility for the players to choose unintended, non-optimal strategies; in other words, the possibility for players to make mistakes.

We already introduced the *best reply dynamics*. If one carefully looks at the equation he will notice that it is essentially describing a diffusive process over a non-ergodic network,

[1] Please note that the scalar product of two stochastic matrices is still a stochastic matrix

[2] Of course this is just one of the many possible transformations that one could use. This one seemed anyway the most reasonable to us since it avoids problems with negative payoffs - differently from transformations of the form $\frac{x^\beta_{i,j}}{\sum_k x^\beta_{i,j}}$ - and moreover is, as well as the learning algorithm, invariant w.r.t the transformation 1.4. For supplementary information about Markov processes see [22]

whose adjacency matrix is given by $B^R B^C$. The components of this network are the cycles, the fixed points, and the free moves of the *best reply dynamics*. We give an example to clarify this concept.

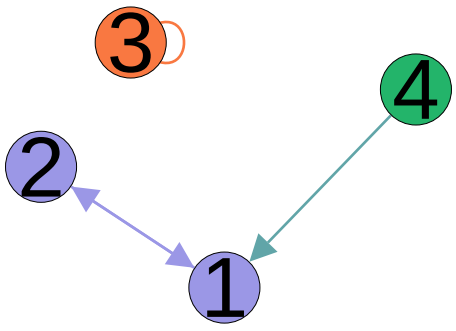
Let us take the (already booleanized) payoff matrix in 1.1. As can be seen from the figure, there is one cycle of length two, one free move and one fixed point. The payoff matrices for the two players are:

$$B^R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad B^C = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Hence:

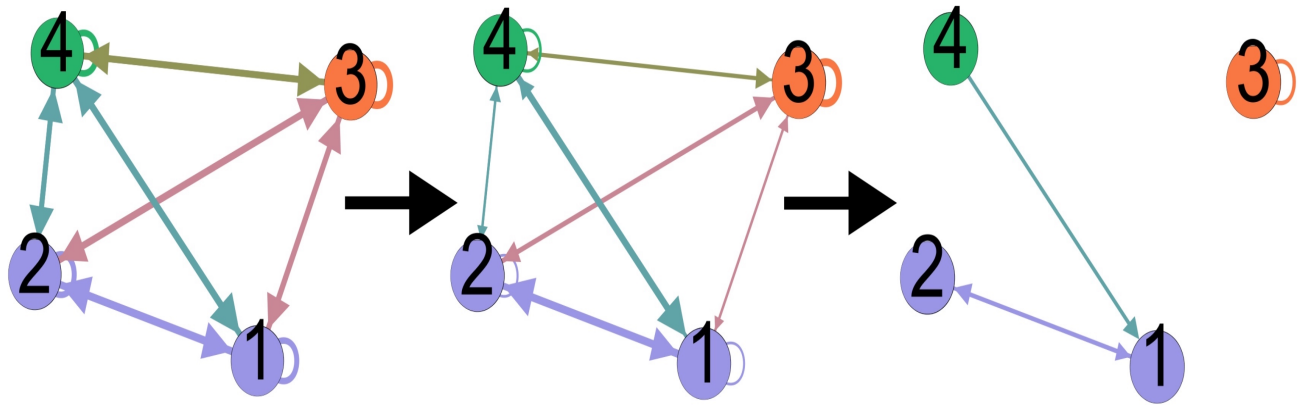
$$Adj = B^R B^C = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

We can already see the presence of a cycle and a fixed point in it, as expected. If we use it as an adjacency matrix for a graph we obtain:



In such a model, attractors are separate entities and, if a player starts in one of them, he will never reach another[3]. The transformation 2.1 instead allows us to create links between the attractors, whose strengths depend on the loss the player gets by deciding to change (i.e. the difference between the payoffs).

Figure 2.1



The size of the edges are proportional to their (*relative*) weight; increasing β we are left with only the links related to the best replies.

For a finite value of β, the graph is ergodic and allows us to find a stationary probability distribution for a dynamical process taking place on it.

Being $\bar{p}$ the stationary probability distribution of the Markov process, the new estimate for the **convergence** frequency will then be[4]:

$$f_{conv} = \frac{\sum_{i \in FP} p_i}{\sum_{i \in FP} p_i + \sum_{i \notin FP} p_i} = \sum_{i \in FP} p_i \tag{2.3}$$

being FP the set of all the moves associated to fixed points. The second equation comes from the fact that, being $\bar{p}$ a probability distribution, $\sum_{i=1}^{N} p_i = 1$.

[3]This is why the number of the moves contained in a cycle is the size of its basin of attraction for the best reply dynamics

[4]Note that the vector with entries equal to 1 in the moves belonging to an attractor and 0 in the free best replies is an eigenvector with eigenvalue = 1 of the boolean matrix. In this sense, in the limit $\beta \to \infty$ the two measures of f_{conv} are equivalent, the only problem being the degeneracy of eigenvalue 1 in the boolean case

Before proceeding, it seems necessary to make a few remarks:

- First, it should be stressed that the model seems particularly reasonable as it creates unbalanced transition probabilities within different attractors depending on how close the payoffs associated to quasi-best and best replies are. We again explain this through an example. Let us take the (single player) payoff matrix (we will assume it to be for player R)

$$\Pi^R = \begin{pmatrix} 1 & 0 & 0 \\ 0.99 & 0 & 100 \\ 0 & 100 & 0 \end{pmatrix}$$

Following our assumption, if player C choices move 1, then player R might be confused between move 1 and move 2 as their payoffs are very close; if instead player C selects either move 2 or 3, then the gap between the payoffs will be so high that player R will hardly ever get wrong. This is captured by our model. Indeed if one takes the logit of Π^R with e.g. $\beta = 5$, he will end up with:

$$T^R_\beta = \begin{pmatrix} 0.51 & 0 & 0 \\ 0.49 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

We see that there is a non null transition probability between the fixed point and the cycle: $p_{FP \to cycle} = T^R_{\beta\,2,1} + T^R_{\beta\,3,1} = 0.49$, while the probability of getting from the cycle to the fixed point $p_{cylce \to FP} = T^R_{\beta\,1,2} + T^R_{\beta\,1,3}$ is zero.

If we imagine that the situation for player C is reversed, i.e its payoff matrix is:

$$\Pi^C = \begin{pmatrix} 100 & 0.9 & 0.9 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \to \quad T^C_\beta = \begin{pmatrix} 1 & 0.49 & 0.49 \\ 0 & 0 & 0.51 \\ 0 & 0.51 & 0 \end{pmatrix}$$

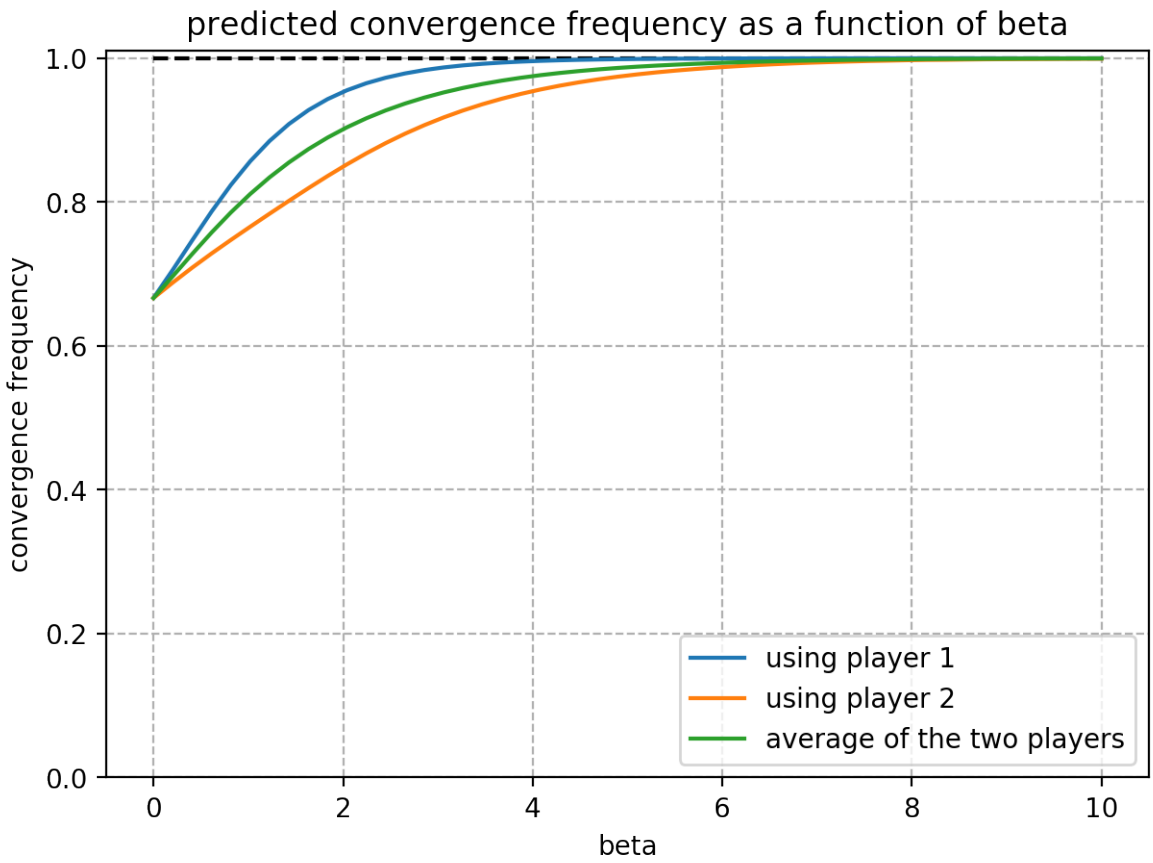
Then this will rebalance the transition probabilities and indeed the final stochastic matrix will be:

$$T_\beta^R T_\beta^C = \begin{pmatrix} 0.51 & 0.25 & 0.25 \\ 0.49 & 0.75 & 0.24 \\ 0 & 0 & 0.51 \end{pmatrix}$$

and essentially $p_{FP \to cycle} = p_{cycle \to FP}$.

- One could argue that, since $T_\beta^R \cdot T_\beta^C \neq T_\beta^C \cdot T_\beta^R$, we would end up having two different Markov processes depending on which of the two players moves first. We solved the problem by using the average of the two estimated convergence frequency. The behaviour of the two, varying β, is anyway very similar (see Fig. 2.2)

Figure 2.2: An example of the process



Here we show the behaviour of the expected convergence probability as a function of beta. In the game chosen, a 3x3 one, only one fixed point was present: the probability indeed goes to one as β increases

- There is then an issue for what concerns the ergodicity of the Markov process. Indeed, in the limit $\beta \to +\infty$ the matrix $M = \lim_{\beta\to+\infty} T_\beta^R \cdot T_\beta^C = B^R \cdot B^C$ would become non ergodic if there are multiple attractors in the best reply structure.The problem was tackled by assuming that for any *finite* value of β, even if large, the system would still be ergodic.

We then checked if the new estimate gave better predictions.

It might be remarked that, even if this work was not originally inspired by it, the idea we proposed is similar to the one behind the *PageRank* algorithm, invented by Google's founder Larry Page and still at the basis of the search engine's positioning algorithm for web pages. Indeed in both cases, a directed, unweighted network is transformed in a complete, weighted one (even if in a quite different fashion); a Markov Process is then used to

assign to each node a numeric value that is depictive of its importance within the network. [5]

Before showing the results, a few words need to be spent on the simulation protocol.

Simulation protocol

To obtain a result that could hold as generally as possible, random games were used throughout the study. With random games we mean games described by payoff matrices whose elements are randomly drawn from a given probability distribution. Following [18] at initialization we randomly generate N^2 pairs of payoffs (a couple for each possible combination (i, j) played; the elements in the couples are the payoffs that players get by respectively playing moves i and j) and we keep the payoff matrix fixed for the rest of the simulation (In the language of the theory of disordered system the random payoff matrix in our problem represents *quenched* disorder [28]). We consider an ensemble of payoff matrices constrained by the mean, variance and correlation of the pairs. The Maximum Entropy distribution that obeys these constraints is a bivariate Gaussian (again, see [18]), which we parametrize with zero mean, unit variance and correlation Γ. The meaning of including a constraint on the correlation was to have a control over the "*nature*" of the game, enhancing *competition* with $\Gamma < 0$ and *cooperation*[6] with $\Gamma > 0$. In the following simulations, only the parameter $\Gamma = 0$ was used (traditional definitions of competitive/cooperative games can be found in every game theory manual; see for example [21, 6, 4]).

For each game, the empirical convergence frequency was computed by performing the learning process 100 times. The number of games taken into account varies depending on the simulations.

In general, 2-player games with $N = 20$ moves were used. This allowed to simultaneously have a non-trivial best reply structure and a decent computational cost.

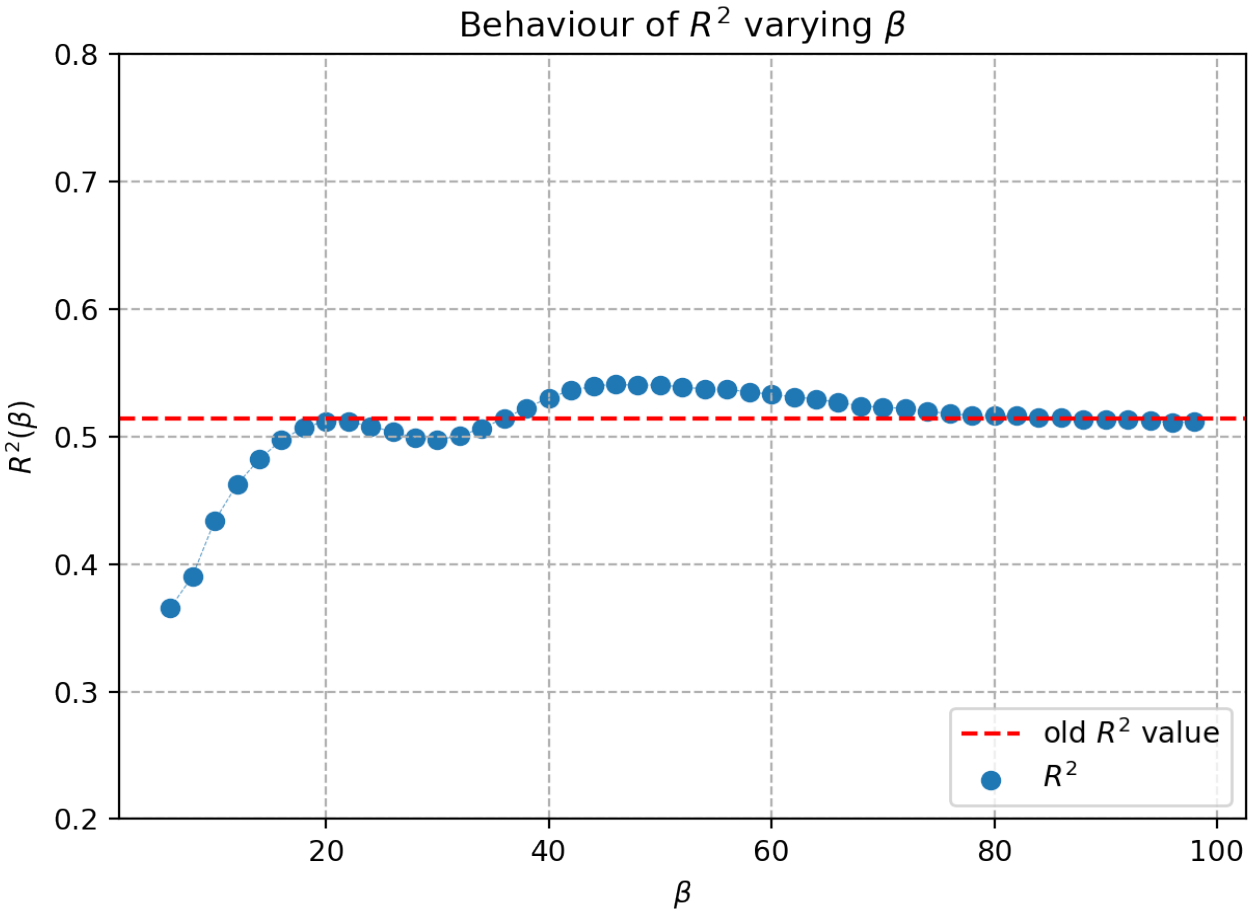
2.1.3 Results

Since the results of this process heavily depend on the value chosen for β, it seemed reasonable to start by studying how well the prediction performs varying β and if there exists an optimal value for the latter that could be chosen in order to have the best estimates.

[5] For further information on PageRank see [26]

[6] With cooperation we mean here the tendency of a combination of moves to be favorable for both players at the same time

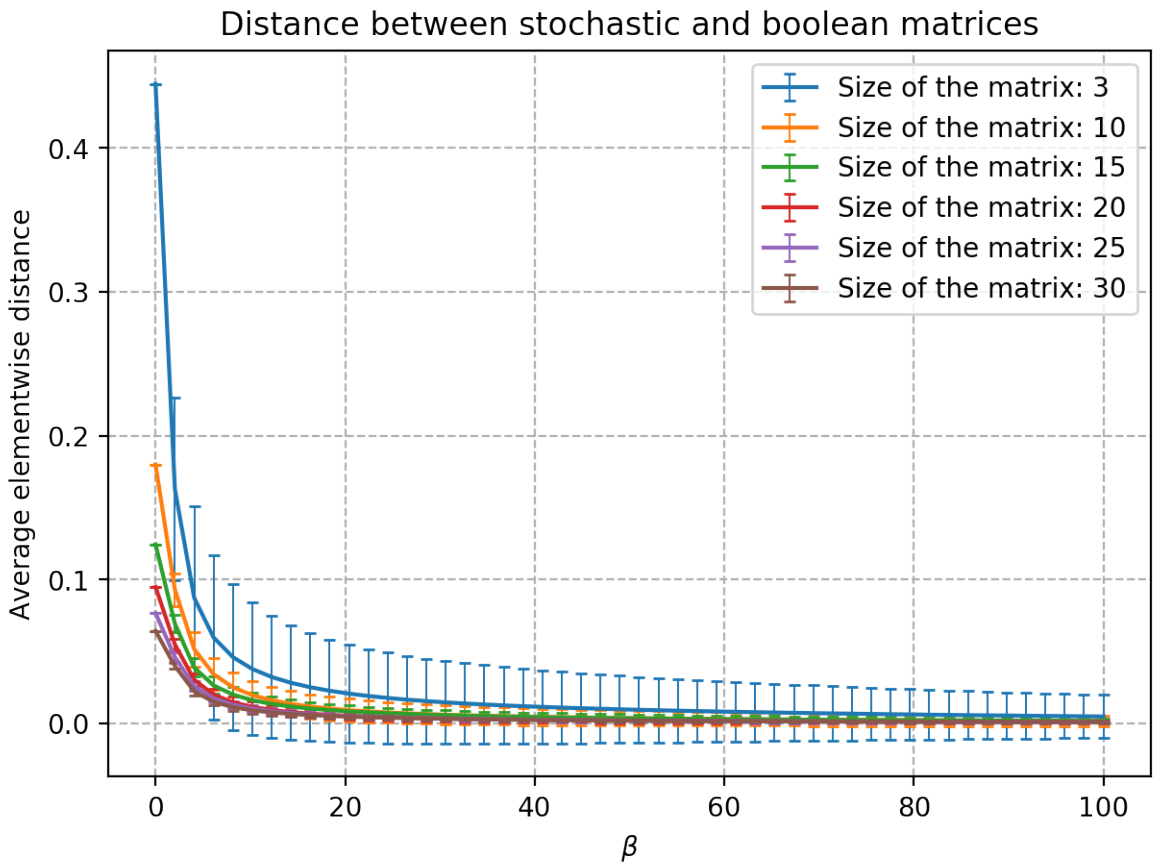
Figure 2.3: Behaviour of R^2



The picture shows the behaviour of the coefficient of determination R^2 as a function of the parameter β. The dashed red line shows the R^2 obtained by using the boolean estimate for the convergence frequency. We see that the new prediction only brings little improvement for $\beta \sim 50$. Fifty 20x20 games were used in the simulation. We want to remark that, even if the qualitative behavior of the function $R^2(\beta)$ seems to be essentially the same in all the simulations (with the only possible difference being that sometimes the peak doesn't manage to outperform the old predictions), the exact position of the peak and the value of the offset are subject to small but still significant variations; we consider the impossibility of exactly placing an optimal value for β a major flaw of the model, but are unfortunately unable to perform larger simulations.

$\beta_{max} = 100$ was chosen to be the maximum value of β since, for size $N = 20$, the matrices $T^{\mu}_{\beta_{max}}$ are already very close to the boolean ones (see following figure).

Figure 2.4: Distance between transformed and boolean matrices



The plot shows the behaviour of the distance between transformed and booleanized matrices for different values of N. As a measure of distance, we simply took $Dist. = \left\langle \left| T^{\mu}_{\beta_{max}}(i,j) - B^{\mu}(i,j) \right| \right\rangle_{(i,j)}$. For each value of N, 50 matrices were extracted randomly. The curve in the figure shows the *mean* distance between this ensemble

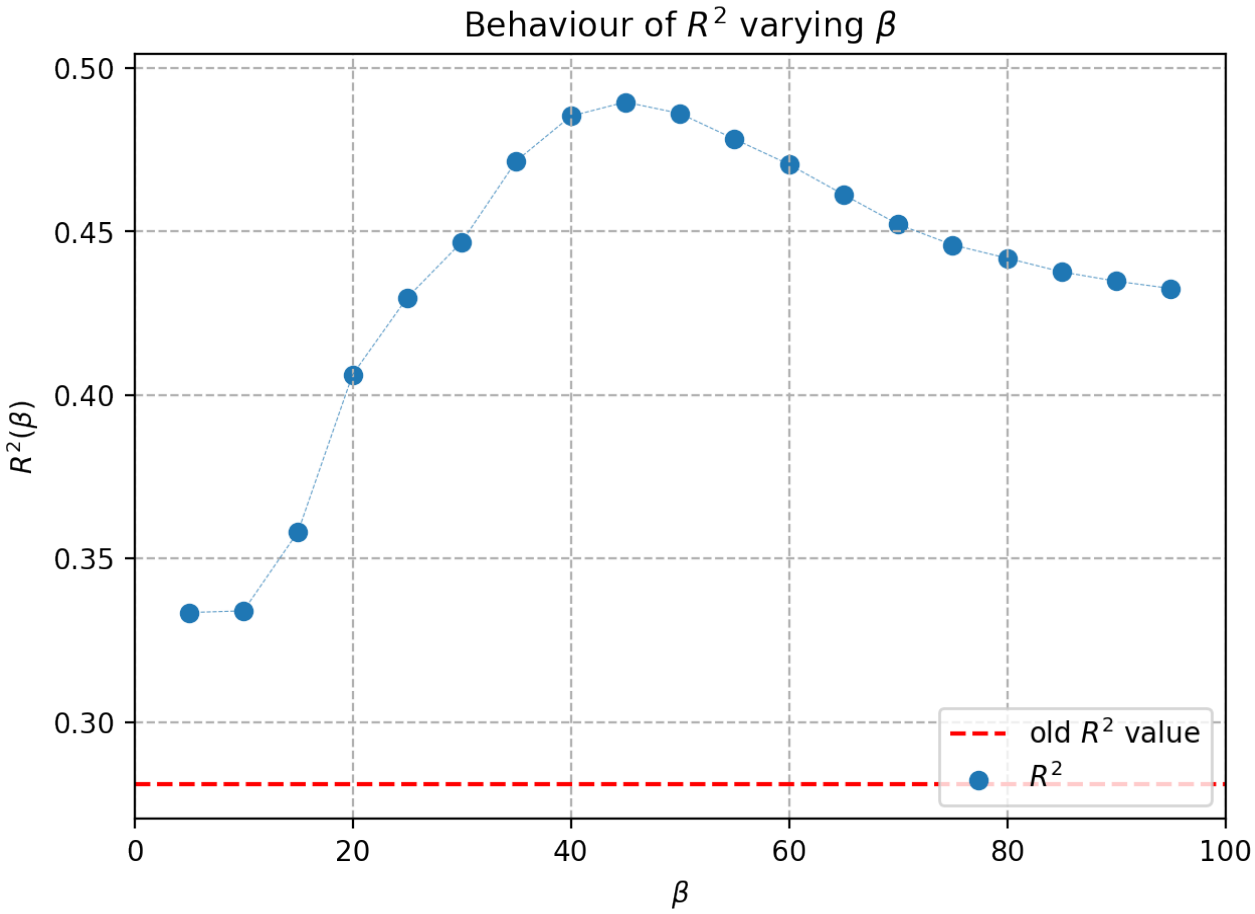
Unfortunately, as can be seen in the plot the new approach seemed to bring little or no improvement to the old one[7]. The behaviour as a function of β assumes, in general, a convex shape in the central part (even if in some simulation it appeared to be monotonically increasing) and, as expected, converges to the old value in the $\beta \to \infty$ limit.

We ran the same analysis on a sample of games which contained at least one fixed point. Indeed both the old approach and the new one (independently from β) assign to games

[7] The setup here is different from the one found in [19] for two different reasons. First of all, the authors of the paper managed to launch larger simulation and it is shown that the value of R^2 increases with the size of them; Secondly, in the paper an averaged way of computing R^2 was used, but its extension to our case is not straightforward. The plot hence must be intended to have a *comparative* value between the two predictions, not an *absolute* one.

which only have cycles a probability of convergence = 0, so it seemed reasonable to exclude them from the simulations and check which of the two performed better on the other games.

Figure 2.5: Behaviour of R^2 excluding matrices with only cycles



For numerical reasons we had to stop the value of β at the value $\beta_{max} = 100$; theory anyway assures us that the curve will asymptotically reach the setup value in the limit $\beta \to \infty$

We can immediately see two trends from the figure[8]; first of all, there is a more evident improvement of the predictions made via the new approach with respect to the original one; secondly there is a small drop in the setup value of R^2. This is not unexpected, as games with no fixed points are indeed very likely to be associated with a non-convergent dynamics (the only possibility of converging being related to a "*numerical coincidence*"[9]), hence on them predictions are often exact. It can be computed analytically that such games

[8]The same considerations made in the caption of the previous figure hold for this result

[9]More on this in 2.2

represent approximatively the 37% of all possible games with 2 players and $N = 20$ moves, hence removing them means essentially getting rid of a large number of exact results. Still, this drop should be taken in consideration when evaluating how good predictions are.

In none of the two cases we seemed to achieve a solid match between predictions and empirical results. Can we fix things or should we discard the model?

Trying to fix the model

A first attempt to interpret the result has been to question the legitimacy of expecting a unique, best-performing value of β to exist. Indeed, one could argue the following. β is responsible for the "*level of noise*" of the final Markov process (here and in the following, with process we mean the stochastic process described by the matrix M): in the limit $\beta \to +\infty$ the process becomes deterministic, while for $\beta \to 0$ the process becomes a random walk in a complete, homogeneous graph. Every intermediate, finite value of β gives rise to a process that is still equiparable to a random walk in a complete graph, but where the probabilities to jump from a node to another are different: the more they are different ($\leftrightarrow$ the more β grows), the less random will the process be.

Each game, on the other hand, has a peculiar conformation that makes learning the equilibrium more or less difficult, hence we could say that in each game is encoded a different "*level of noise*" (leaving intentionally this definition vague for the moment). Shouldn't then we expect each game to be best modeled (i.e. its convergence frequency to be best approximated) by a process obtained through a value of β that is peculiar to the game and related to the level of noise that it encodes?

2.2 Predicting the optimal β

The focus has now shifted on the problem of trying to predict the *optimal* $\beta = \beta^*$ for each game, meaning with that the value of β for which the transformation described in the previous section gives the best result.

We started by checking if any connection existed between β^* and various metrics that could fit the definition of "*level of noise*" encoded in a matrix. The results have unfortunately been discouraging.

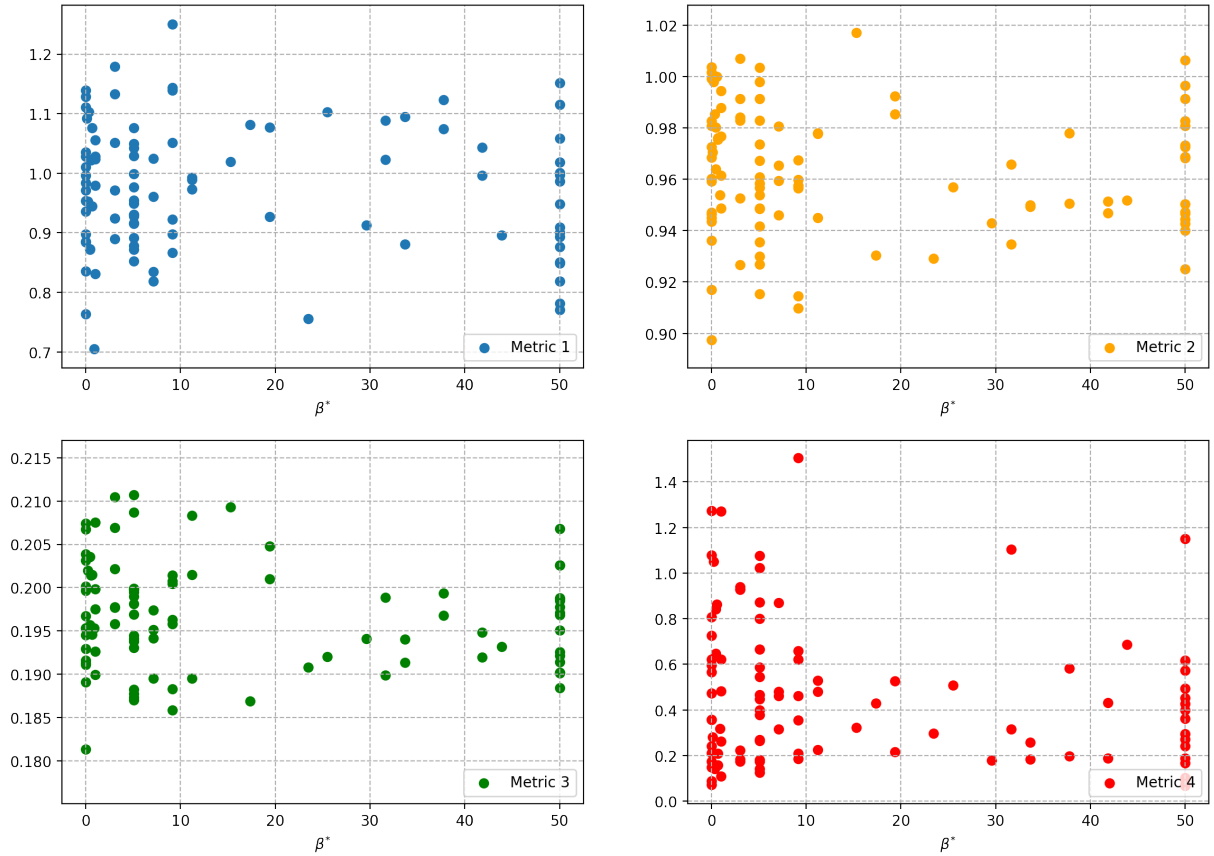
Before showing the results, a few words on how the metrics were chosen. First of all, as we already said, the learning process is invariant under the transformation 1.4; metrics should then have the same property. This leaves out a large ensemble of possible metrics that might seem reasonable: everything that is constructed including means, maximum values, ratios between elements etc. should be ruled out. It should be borne in mind that our original guess was convergence is influenced by *quasi-best replies*: the more close their payoffs are to the ones of the *best replies*, the more likely will be for the dynamics to escape an attractor. Our metrics should hence try to quantify this all.

Metrics tested were the following (when talking about columns we refer to the columns of the payoff matrices):

Metric 1	Average standard deviation $\langle\sigma\rangle$ of the columns
Metric 2	Standard deviation σ_{FP} of the column that contains the fixed point
Metric 3	Average distance $\langle d\rangle$ between payoffs in the same column
Metric 4	Distance $d_{1,2}$ between the payoff associated with the fixed point and the second largest payoff in the same column

Results of the tests are shown in the following plot:

Figure 2.6: Results obtained by using the different metrics



The plot shows the results obtained by using the different metrics. As evident from the figure, none of them seems to be correlated with β^*. Simulations were performed over 100 different games.

As we can see, there is unfortunately no clear correlation between any of this metrics and the value of β^*.

Some intuitions about why metrics don't work

As we saw, the metrics we used (and possibly the whole model previously built) seem incapable of improving the predictions of the old one. In the following, we will try to throw out a few thoughts about why this is the case.

The principal assumption of our model was that players can get "distracted" in their learning dynamics by quasi-best replies. This led naturally to a model where everything was conceived column-wise: the Markov process was implemented by applying a logit transformation on the columns of the payoff matrices and the metrics took into account quantities that were computed column-by-column. This still seems natural, but is not necessarily true. From the update equations of the EWA algorithm 1.2 we see that, if one of the two players sticks to a given move, its opponent will eventually learn to play his

best reply no matter how the distance between the largest and the second largest payoff is (even if the closer they are, the longer will learning take).

Another key issue might be that a move that is a best reply to a given move of the opponent might not be a convenient move to play in general. Let us take as an example the game:

$$\begin{pmatrix} (1,1) & (0,0) & (0,0) \\ (0,0) & (100,0) & (0,100) \\ (0,0) & (0,100) & (100,0) \end{pmatrix}$$

At the beginning of the learning process the strategy vectors $\bar{x}$ and $\bar{y}$ extracted randomly from a uniform distribution. We adopt the point of view of player R and define the *expected payoff* $\bar{\pi}^{exp}$ as

$$\pi_i^{exp} = \sum_j \Pi_{i,j}^R y_j$$

We can immediately see that it is highly unlikely (the probability obtained numerically is $\sim 10^{-4}$) that move 1 will be the one with the largest expected payoff, meaning that

$$P\left(1 = \arg\max_k \left(\pi_k^{exp}\right)\right) \sim 10^{-4}$$

Unfortunately, the growth rate of the probability associated with each move is proportional to its expected payoff (see 1.2 again), so move 1 will end up having a null probability of being played and indeed the empirical convergence frequency of such a game resulted to be zero.

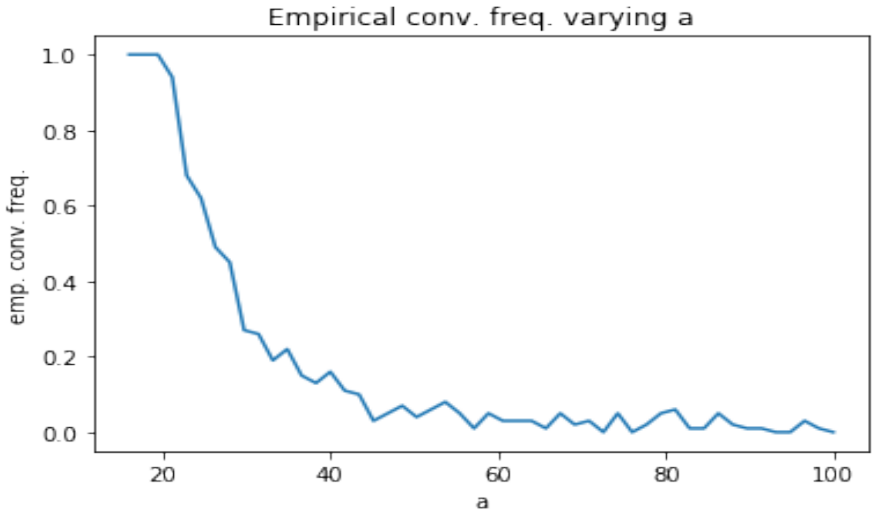
All this could be in principle be captured by the Markov process, but it's not clear whether efficiently or not. We report here an example to give a flavour of what happens. Let us take again the matrix in 1.1[10] and change two of its elements in the variables a and b.

[10] Actually a smaller version of it

$$\Pi = \begin{bmatrix} (7,-5) & (2,14) & (-4,3) \\ (-9,a) & (10,-3) & (3,15) \\ (b,-9) & (0,-6) & (8,1) \end{bmatrix}$$

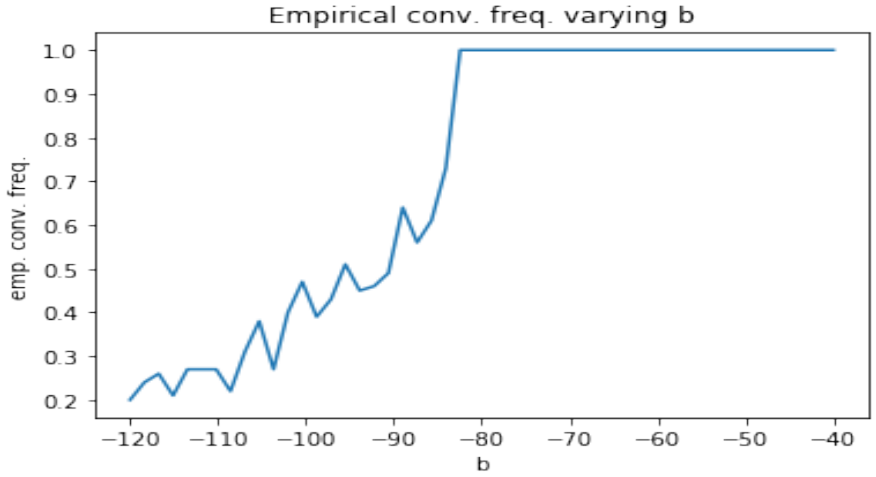
We can look at how the convergence frequency is affected by varying a and b[11].

Figure 2.7



[11]More specifically, we increase a and decrease b in order not to change the best reply structure of the game; when one of the two variables is varied, the other is maintained fixed to its original value

Figure 2.8



Increasing a (Fig. 2.7) we're increasing the expected payoff for move 1 of player column. Since the move belongs to a cycle we expect - and observe - a destabilizing effect. The same reasoning holds if we decrease b (Fig. 2.8), hence decreasing the expected payoff that player one has for move 3, the only fixed point of the game. Again, the destabilizing effect is observed.

Can our framework capture this phenomenon?

(a)

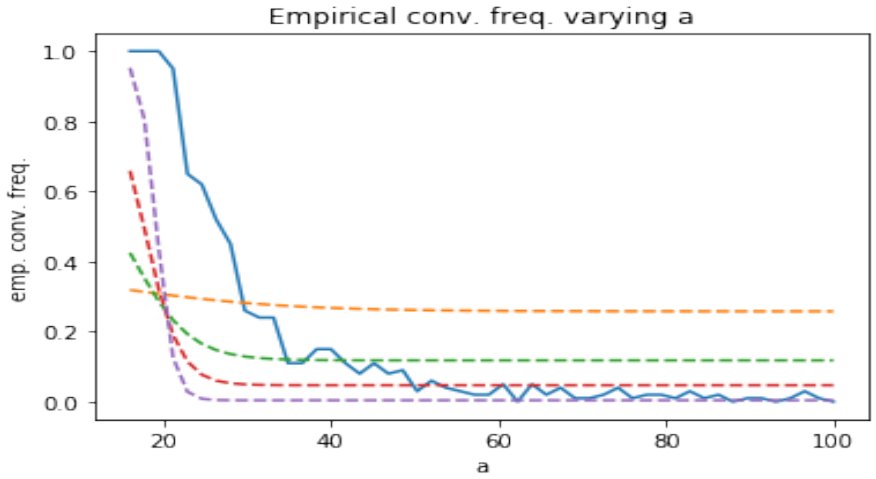
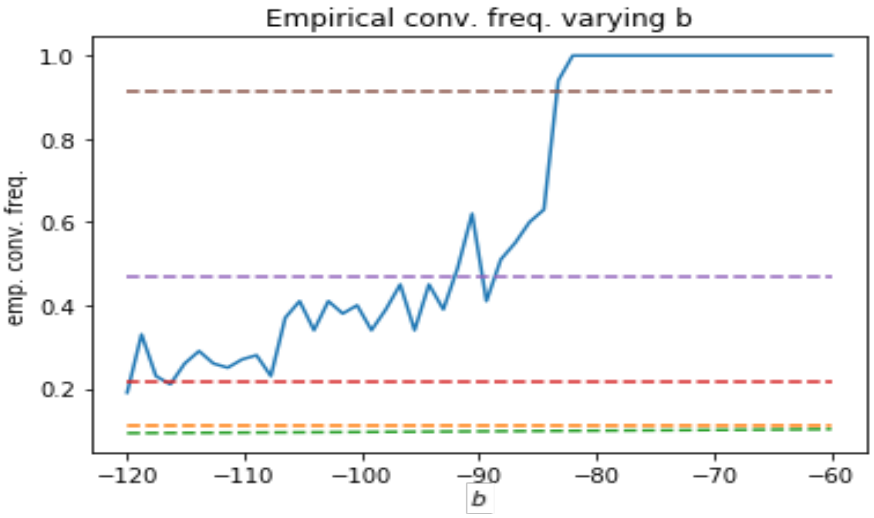


Figure 2.9

(b)



The dashed line represent the predicted probabilty for various values of β: 0.1,1,2.5,5; the steepest the decrease in figure (a), the larger is β

We see that the predictions work well (at least qualitatively) when a increases, but are unable to follow the behavior of the empirical convergence frequency when b is the one who varies. This would suggest that our predictions are not always able to capture how the initial expected payoffs influence the dynamics of the learning process.

Some issues arise from the nature of convergence as well. It is not uncommon to observe convergent simulations in games that lack any fixed point. What did the players learn? An optimistic guess would be to assume that players learned the *mixed strategy Nash equilibrium* of the game, that we know to exist([9]) for every game with a finite number of players and moves; unfortunately, this is not the case in the vast majority of simulations. Convergence is then to ascribe to an accidental balance between memory loss and growth rates of EWA updating equations, hence essentially casual or heavily dependent on the choice of parameters. All this makes the prediction of such kinds of convergence essentially impossible by an analysis of the game solely based on game theory and, in general, by simple models: they are heavily dependent on the details of the algorithm.

Chapter 3

Conclusions

In this work we have addressed the problem of understanding which features of a game influence the convergence of EWA learning algorithm. Taking recent literature as a starting point [19] we focused on the relation between the empirical convergence frequency for a given game and its *best reply structure*. We tried to broaden this framework and obtain more precise predictions by including *quasi best replies*. In order to do so, we developed an analogy between the execution of a game by two players and a diffusive process on a fully connected weighted graph. We looked at the stationary distribution of such a process and tried to see if it could be used to calibrate the strength (e.g to infer the relative size) of each attractor.

The aforementioned analogy was based on a *logit* one parameter (β) transformation that mapped the payoff matrix in a stochastic one. We ran extensive simulations to see how our predictions performed as β varied. Unfortunately, the new framework seemed to give little or no improvement to the old one.

Arguing that the problem was choosing the same value of β for all the games included in the simulation, we introduced the notion of *optimal beta* β^* and tried to see whether this could be directly inferred from the payoff matrix. To this aim, we developed and tested four different measures that were coherent with the properties of the learning algorithm. Unluckily, none of them seemed to have a clear predictive value. We finally shared some thoughts about why these methods failed and which issues they are not able to cope with.

Even if our guess might be considered unsuccessful, we think it was not unuseful. First of all, it gives a hint on how solid the methodologies of [19] are, showing that the very basic model used there is hardly improved by including more complex and detailed features of the game. Secondly, the whole work gives a flavor of how important details are when dealing with this problems: the parameters and the heuristics we choose have a dramatic impact on the results one obtains. This is why it is hard to capture them with simplistic models and, in our opinion, to outperform the very basic one; it could be argued that predictions based on the *best reply structure* are maybe the best that one could get and that what they fail to capture is essentially noise - and is there really a reason in trying to capture noise in random games?

Finally, we hope that this work could be a basis for further investigations of the topic.

Appendix A

Is equilibrium common?

In this thesis we tried to clarify what is the link between the structure in which best replies are arranged in a game and the frequency at which players manage to learn how to properly play it. A further step would be then to study the statistical properties of best reply structures: how likely is it for a generic game to have a given best reply structure? This was investigated in [19] using a *microcanonical* point of view, i.e. using combinatorial techniques to count how many configurations would show a given structure over the complete ensemble of possible games.

Another possible approach might exploit some well-known results in *random matrix theory*. To understand this link, two observations are necessary.
First of all, one should remember that the games we studied are formally described by a pair of payoff matrices whose entries are randomly extracted from a gaussian distribution: hence by two random matrices.
Secondly, after the booleanization, the spectrum of the matrix $A = B^R B^C$ we see in 1.8 (from now on *Adjacency Matrix*) encodes all the information about the best reply structure of the game [33]. Indeed, is quite straightforward to show that *free best replies* (best replies that lead to a best reply dynamics' attractor without being part of it) will correspond to a $\lambda = 0$ eigenvalue, fixed points correspond to a $\lambda = 1$ eigenvalue, while a cycle of length k will correspond to a set of eigenvalues such that $\sqrt[k]{\lambda} = 1$. Finding the probability distribution for the eigenvalues of this kind of adjacency matrices will then essentially give an estimate of how likely it would be to find a given best reply structure.

Up to now, most of the literature in random matrix theory focused on *Gaussian Ensembles*[1]. The matrices we are dealing with are, anyway, of a completely different nature. Their elements can only be zeros and ones; moreover there can be just one non-null element per column. Such matrices are called *permutation matrices*.

The most general result for random matrices eigenvalues can arguably be considered the circular law theorem, asserting that for any sequence of random $n \times n$ matrices whose

[1] Particularly appealing as they can be used to model several different kinds of Hamiltonians; for those ensembles (*Gaussian Unitary* or GUE, *Gaussian Orthogonal* or GOE, *Gaussian Symplectic* or GSE) the full joint probability distribution for the eigenvalues is known.

entries are independent and identically distributed random variables, all with mean zero and variance equal to $\frac{1}{n}$, the limiting spectral distribution is the uniform distribution over the unit disc. More precisely:

Definition 6 (Circular Law Theorem) *Let x be a complex random variable with mean zero and bounded variance σ^2. Let N_n be a random matrix of order n with entries being i.i.d. copies of x. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\frac{1}{\sigma\sqrt{n}} N_n$. Define the empirical spectral distribution μ_n of N_n by the formula:*

$$\mu_n(s,t) := \frac{1}{n} \sharp \{k \leq n | Re(\lambda_k) \leq s; Im(\lambda_k) \leq t\}$$

Then μ_n converges to the uniform distribution μ_∞ over the unity disk as n tends to infinity.

Permutation matrices, anyway, don't fall in this category, their entries being far from independent. Nevertheless we saw in 2.1 how such matrices can be obtained by ordinary ones, the interesting part being that for small values of β the matrix $T_\beta^R T_\beta^C$ can be approximately considered a Gaussian random matrix. The reasoning goes the following way: let R_{ij} be an entry of the payoff matrix for player row. Applying the logit transformation, it will become:

$$R_{ij} \to \frac{e^{\beta R_{ij}}}{\sum_j e^{\beta R_{ij}}}$$

Let R' be equal to $e^{\beta R_{ij}}$. By definition, R' will be a log-normal distributed random variable with well defined moments. Now let's assume that, in the large N limit, we can approximate the denominator of the previous expression with $\sum_j e^{\beta R_{ij}} \sim N \langle R' \rangle$. We have now found a matrix whose entries are i.i.d log-normal variables, with mean value $\sim \frac{1}{N}$ and variance $\sim \frac{1}{N^2}$. The same reasoning applies to the payoff matrix C of player column. To construct the adjacency matrix A, we have to take the scalar product of the two. The element A_{lk} of such a matrix will be:

$$A_{lk} = \sum_i R'_{ki} C'_{il}$$

Each term $R'_{ki} C'_{il}$, being the product of two independent lognormal variables, is itself a lognormal variable[2]; A_{lk} is then given by the sum of N i.i.d random variables with well defined moments and hence can be approximated with a gaussian random variable due tu the Central Limit Theorem; moreover, it's independent from the other entries in the matrix. Indeed:

[2]This comes by the properties of the lognormal distribution

$$E\left[A_{kl}A_{qr}\right] = E\left[\sum_i R'_{ki}C'_{il}\sum_j R'_{pj}C'_{jr}\right] = E\left[\sum_{ij} R'_{ki}R'_{pj}C'_{il}C'_{jr}\right]$$

Now:

- if *l,p,k,r* are all different, then the four variables are all independent and:

$$E\left[\sum_{ij} R'_{ki}R'_{pj}C'_{il}C'_{jr}\right] = \sum_{ij} E\left[R'_{ki}\right] E\left[R'_{pj}\right] E\left[C'_{il}\right] E\left[C'_{jr}\right] =$$

$$= N\left(E\left[R'_{ki}\right]\right)^4 \sim N\frac{1}{N^4} \to 0 \text{ in the large N limit} \quad (3.1)$$

- if $k = p$ (or equivalently $l = r$):

$$E\left[\sum_{ij} R'_{ki}R'_{kj}C'_{il}C'_{jr}\right] = \sum_{ij} E\left[R'_{ki}R'_{kj}\right] E\left[C'_{il}\right] E\left[C'_{jr}\right] \simeq \frac{1}{N^2}\sum_{ij} E\left[R'_{ki}R'_{kj}\right] \quad (3.2)$$

  Let's now decompose the sum in two terms:

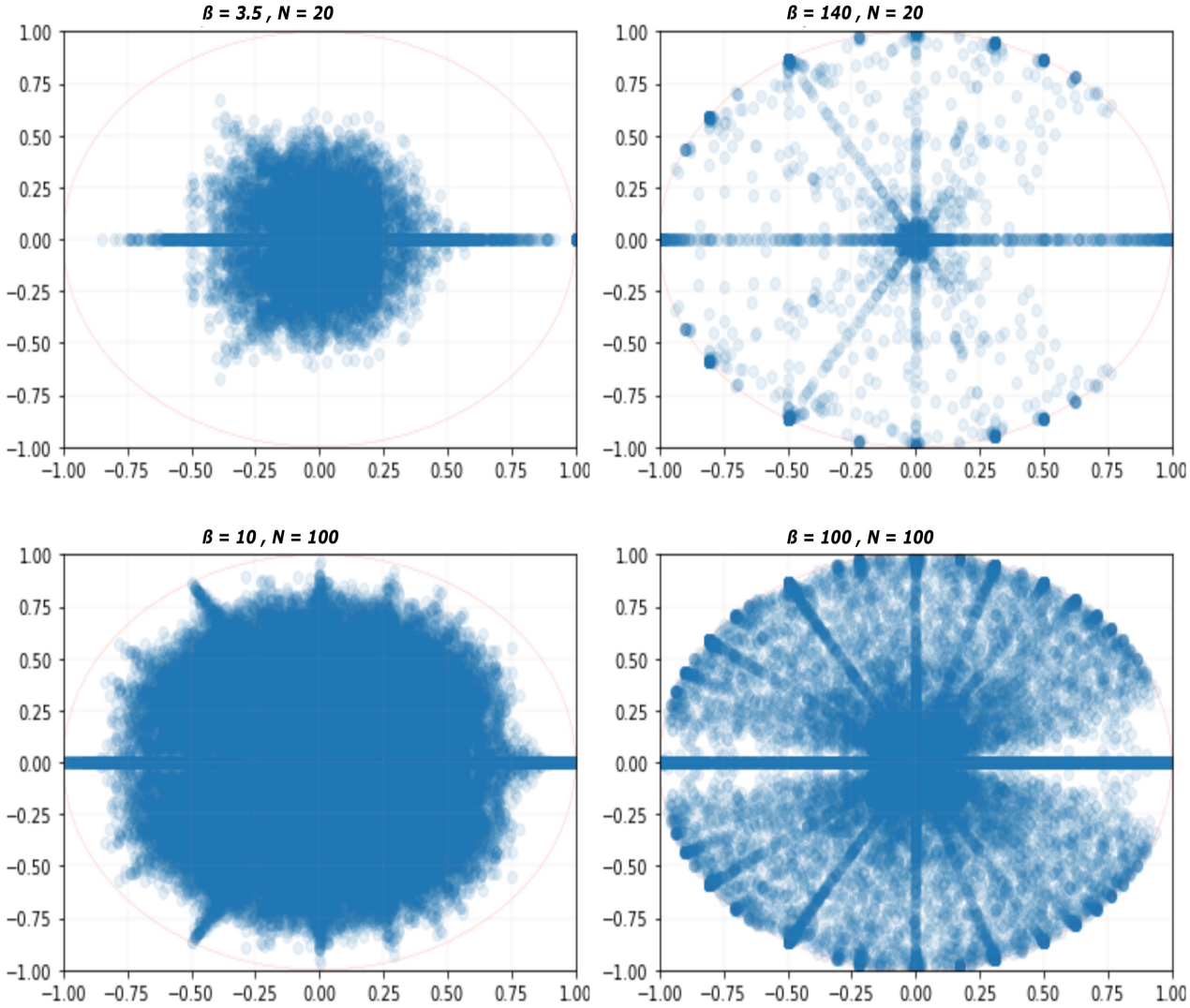
$$\sum_{ij} E\left[R'_{ki}R'_{kj}\right] = \sum_{i\neq j} E\left[R'_{ki}R'_{kj}\right] + \sum_i E\left[R'^2_{ki}\right]$$

  Now, for $i \neq j$ the variables R'_{ki} and R'_{kj} are independent. The first term hence goes like $N^2E^2\left[R'\right] \sim N^2\frac{1}{N^2} \sim 1$ and gives no contribution when multiplied by the factor $\frac{1}{N^2}$. The second term instead goes like $N\frac{1}{N^2} \sim \frac{1}{N}$ and again goes to zero.

To sum up our results, it seems like the entries of the matrix A can be well approximated, at least for small values of β, by independent Gaussian variables; we then expect its spectrum to follow the Circular Law for such βs. We also now that, in the $\beta \to \infty$ limit, A will become a permutation matrix, whose eigenvalues can only be on the center or at the border of the unit circle. We expect then to observe a gradual transformation of the spectrum from a uniform distribution to a very inhomogeneous one as β grows.
Simulations seem to confirm this intuition[3]:

[3] Note that it is normal to have a radius different from 1 for the uniform circle if the random variables have not been rescaled

Figure 3.1



The higher N, the more resemblant the distribution is to an homogeneous circle for low values of β; The smaller N, the faster the eigenvalues tend to recollocate themselves either on the center or on the border of the cycle.

Note that, the eigenvalues on the unit circle are only located on angles $\alpha = \frac{2\pi}{n}$, with $n \in [1, N]$. This is why they are more distanced from one another in the $N = 20$ case.

Investigating in detail how the eigenvalues' distribution deviates from the circular law

as β varies[4], and hence trying to find the probability distribution for the eigenvalues of permutation matrices, might be interesting for future analyses.

[4]Note that β influences the actual values of the moments of the distribution of R'

Appendix B

More than two players

The results we showed in this thesis - both originals and not - were focused on two players games. It would be quite straightforward to ask how they are affected by increasing the number of players. In this regard, the only work available up to now is a paper by Sanders, Galla and Farmer [28] who, taking [18] as a starting point, managed to show both analytically and numerically that chaotic behavior of learning processes seems to become more common when the number of players increases (i.e. the region of the parameter space associated with chaotic dynamics is larger for games with a higher number of players).

A similar extension of [19] has not been realized yet. It is anyway possible to make some preliminary remarks. First of all, the definition of best reply structure can be expanded also to games with a higher number of players. As long as one assumes a fixed order of choice[5], the very same formalism is recovered: we will still find free moves, fixed points and cycles for the best reply dynamics and work with them. Moreover, if the same network analogy is applied to a p-players game, the best reply dynamics can be depicted via a p-partite graph. This network, if one makes a sequence of *bipartite network projections* can in turn be projected in a monopartite one with N^{p-1} nodes. [6].

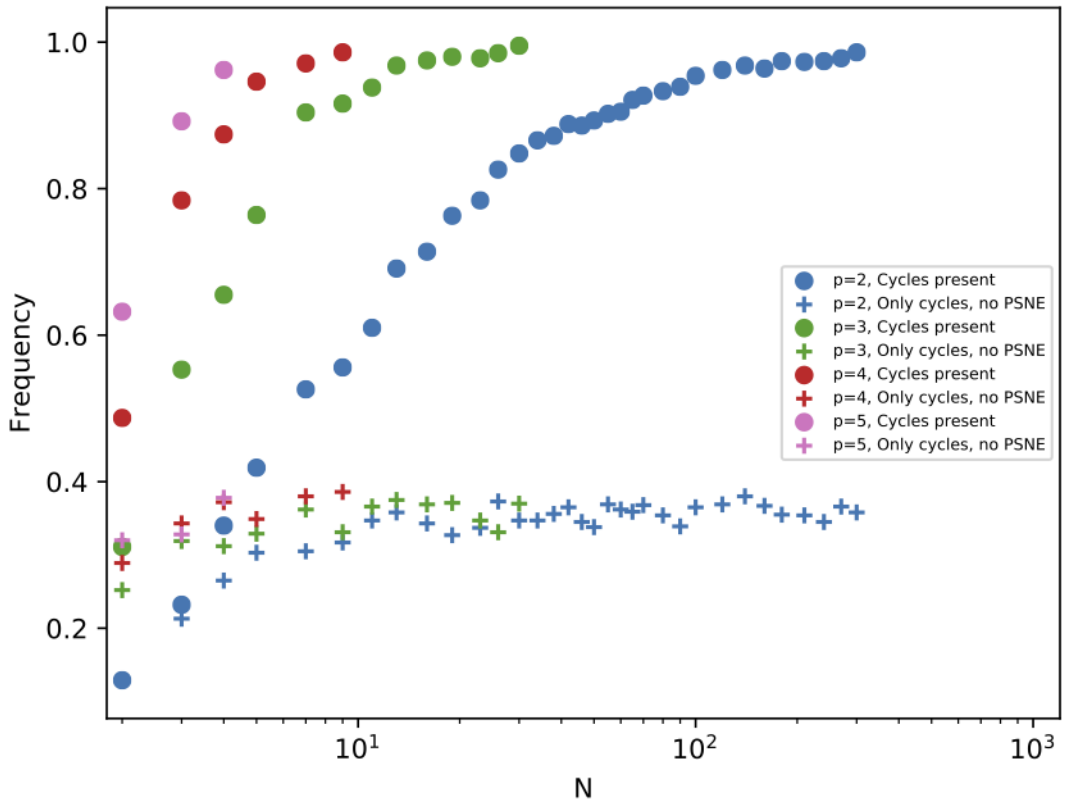
As we already know, the properties of this last monopartite graph can be studied by looking at its adjacency matrix, that will still be a permutation one. This seems to suggest that, if one looks at the best reply structure, the only difference between a p-player and a 2-player game will be the size of the adjacency matrix. This would allow us to exploit the analysis carried on for 2p games in [19] even for a larger number of players.

Let's see if our intuition works on an easy example. In the following plot we show the frequency of occurrence of cycles as a function of the number of strategies for games with different numbers of players.

[5]Meaning with that, a fixed order in which players make their moves, e.g. Player 1 chooses first, then Player 2 chooses,...

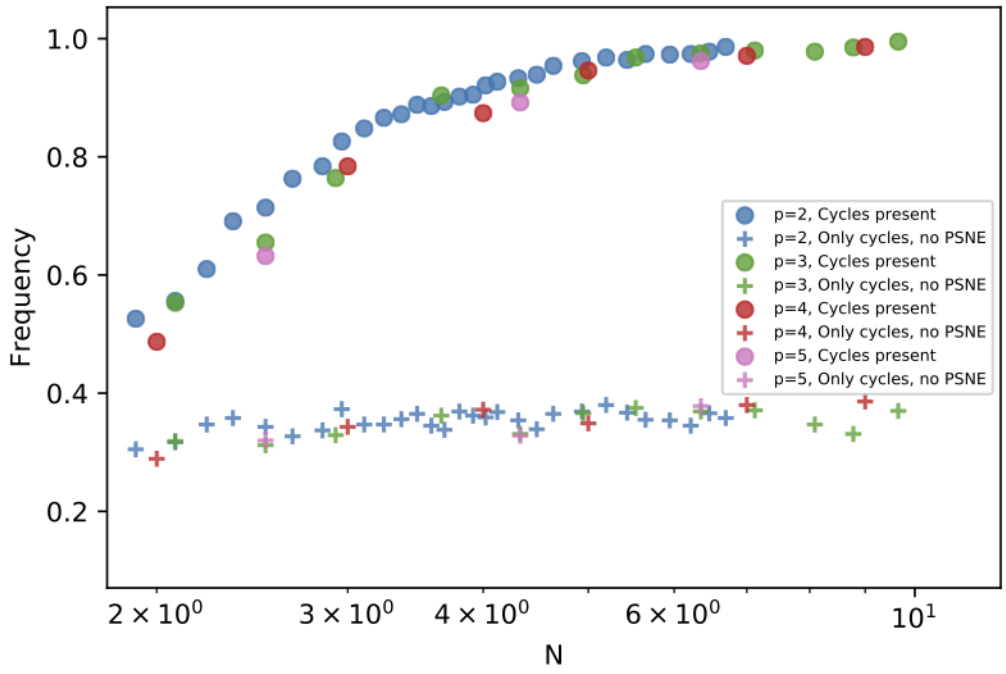
[6]If one assumes the point of view of player 1, each node would correspond to a possible sequence of moves played by all the other players, with the link between two nodes being a connection between two sequences established via the choice of player 1

Figure 3.2: Occurrence of cycles as a function of the number of strategies



[7]The presence of cycles is only related to the best reply structure of a game and can hence be investigated by simply looking at the adjacency matrices. Our guess is that the only difference between a p-players, N strategies game and a 2-players one with the same N lies in the size of these matrices (again, $N^{p-1} \times N^{p-1}$ for the p-players game and $N \times N$ for the 2 players one). This means that, for example, that a 4 players game with N strategies would have the same frequency of occurrence of cycles of a 2-players game with N^3 strategies, or of a 3-players game with $N^{\frac{3}{2}}$ strategies and so on. The curves in Fig. 3.2 should then collapse in a single curve if one appropriately rescales the x axis for each value of p.

[7]Original plot by courtesy of Dr. Torsten Heinrich



If the x axis is properly rescaled the five curves collapse in a single one. The 4p curve was taken as a reference in this example

This is indeed what we observe in the simulations. The result might be important because it tells us that, if a correlation between the best reply structure and the frequency of convergence of learning algorithms exists also for higher values of p, than all the analytical results obtained for 2p games in [19] can be extended to generic values of p.

Up to now, a similar investigation has not been carried on, so every enthusiasm would be premature. Moreover, as we said, the reasoning only holds as one assumes a fixed order of choice for p players games, while the dynamics of learning algorithms is anyway synchronous and doesn't imply any order of choice. A new definition of the best reply structure that takes this problem into account could be realized, but it's not sure whether the scaling results would still hold for it.
If it works anyway, with a reasoning not uncommon in science, the fixed order choice could anyway still be assumed as a crude yet useful approximation.

Bibliography

[1] A. R. Mele, P. Rawling, *The Oxford Handbook of Rationality*, Oxford University Press (2004).

[2] J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press (1944).

[3] L. Walras, *Elémènts d'économie politique pure*, Lausanne (1984)

[4] K. Leyton-Brown, Y. Shoham, *Essentials of Game Theory*, Morgan & Claypool Publishers (2008).

[5] J.M. Smith,*Evolution and the Theory of Games*, Cambridge University Press (1982).

[6] H. Gintis,*Game theory evolving: A rpoblem-centered introduction to modeling strategic behavior*, Princeton University Press (2000)

[7] R. Axelrod, W.D. Hamilton *The evolution of cooperation*, Science 211(4489):1390–1396 (1981).

[8] M.A. Nowak, D.C. Krakauer *The evolution of language.*, Proceedings of the National Academy of Sciences 96(14):8028–8033 (1999).

[9] J. Nash,*Equilibrium points in n-person games*, Proceedings of the National Academy of Sciences 36(1):48-49 (1950).

[10] D. Fudenberg, D.K. Levine *The theory of learning in games.* MIT press (1998)

[11] C. Daskalakis, P. W. Goldberg, C. H. Papadimitriou *The Complexity of Computing a Nash Equilibrium*, SIAM Journal on Computing, VOl. 39 (2009)

[12] J.H. Nachbar *"evolutionary" selection dynamics in games: Convergence and limit properties.* International journal of game theory 19(1):59–89. (1990)

[13] D.P. Foster, H.P. Young *On the nonconvergence of fictitious play in coordination games.* Games and Economic Behavior 25(1):79–96. (1998)

[14] D. Monderer, L.S. Shapley *Fictitious play property for games with identical interests.* Journal of economic theory 68(1):258–265. (1996)

[15] P. Milgrom, J. Roberts *Rationalizability, learning, and equilibrium in games with strategic complementarities.* Econometrica: Journal of the Econometric Society pp. 1255–1277. (1990)

[16] I. Arieli, H.P. Young *Stochastic learning dynamics and speed of convergence in population games.* Econometrica 84(2):627–676. (2016)

[17] S. Bowles, H. Gintis *A Cooperative Species: Human reciprocity and its evolution*, Princeton University Press (2011)

[18] T. Galla, J.D. Farmer, *Complex dynamics in learning complicated games*, Proceedings of the National Academy of Sciences 110(4) 1232-1236 (2013).

[19] M. Pangallo, T. Heinrich, J.D. Farmer *Best reply structure and equilibrium convergence in generic games*, arXiv:1704.05276
[20] C. Camerer, T. Ho *Experience-weighted attraction learning in normal form games*, Econometrica 67(4):827–874 (1999)
[21] R.B. Myerson *Game theory*, Harvard university press (2013).
[22] C.W. Gardiner *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Science*, 2nd Ed, Springer-Verlang New York (1985)
[23] C. H. Taubes *Lecture Notes on Probability, Statistics and Linear Algebra* (2010)
[24] Oxford Dictionaries *Markov chain | Definition of Markov chain in US English by Oxford Dictionaries*. (2017)
[25] R. Selten *A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games*. International Journal of Game Theory. 4 (1): 25–55. (1975).
[26] L. Page, S. Brin, R. Motwani, T. Winograd *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab (1999)
[27] S.H. Strogatz *Nonlinear Dynamics and Chaos : with Applications to Physics, Biology, Chemistry, and Engineering*, Westview Press (2015)
[28] J.B.T. Sanders, J.D. Farmer, T. Galla *The prevalence of chaotic dynamics in games with many players*, Scientific Reports 8:4902 (2018)
[29] M. Pangallo, J.B.T. Sanders, T. Galla, J.D. Farmer *A taxonomy of learning dynamics in 2 x 2 games*, arXiv:1701.09043
[30] J. Berg, A. Engel, *Matrix Games, Mixed Strategies and Statistical Mechanics*, Phys. Rev. Lett. 81, 4999-5002 (1998)
[31] T. Tao, V. Vu *Random Matrices: The circular Law* arXiv:0708.2895
[32] G. Livan, M. Novaes, P. Vivo *Introduction to Random Matrices*, SpringerBriefs in Mathematical Physics (2018)
[33] J. Najnudel, A. Nikeghbali. *The distribution of eigenvalues of randomized permutation matrices*. Annales de L'Institut Fourier, 63(3):773–838, 2013.