Politecnico di Torino

Politecnico di Torino

Corso di Laurea magistrale in
Physics of Complex Systems

A.a. 2023/2034
Sessione di Laurea: Aprile 2024

Dynamical equilibrium selection by reinforcement learning algorithms: a study of two time-scale regimes in the El Farol Bar problem

Relatori
prof. Luca Dall'asta

Candidato
Giuseppe Citino

Acknowledgements

I would like to thank my supervisor, Prof. Luca Dall'Asta, for his availability and support throughout the journey of this thesis, both of which were far from granted. I extend my thanks to those who have been by my side, providing frontline or behind-the-scenes support during this academic journey filled with obstacles, to those who did not falter in the face of my periods of closure, and to those who experienced them with me and like me.

Abstract

Reinforcement learning (RL), a branch of machine learning inspired by behavioral psychology, focuses on training agents to make sequential decisions in dynamic environments. RL algorithms, exploit a reward system to learn optimal strategies and have found extensive applications in various fields, including robotics, finance, and game theory.
Game theory, on the other hand, provides a framework for understanding strategic interactions among rational decision-makers. One classic problem in game theory is the El Farol Bar problem, which models the dilemma faced by individuals trying to decide whether to attend a crowded bar, modeling any situation where individuals must make decisions based on limited information and competition for resources.
This thesis investigates the application of RL algorithms in the context of the El Farol Bar problem. By adapting an RL algorithm originally designed for potential stochastic games to a non-stochastic scenario, we uncover intriguing dynamics characterized by two distinct time scales. Through numerical analysis, we demonstrate that the relative speeds of these time scales critically influence convergence behavior.
Specifically, when the Q-function evolves in a fast regime, the convergence favors the unique symmetric mixed strategies Nash equilibrium. Conversely, if the fast scale governs the learning of strategies, the convergence shifts towards one of the pure strategies Nash equilibria. This observation highlights the intricate interplay between learning dynamics in RL algorithms and the stability of equilibria, offering valuable insights into their convergence properties.
This thesis aims at contributing to the understanding of RL algorithms in strategic decision-making contexts, shedding light on their adaptability and convergence behavior. It emphasizes the significance of considering the dynamics of learning processes in analyzing strategic interactions, paving the way for further exploration from a mathematical formalization perspective.

Contents

Chapter 1

Definitions, properties and notation in games

Let's start by introducing the main definitions and notations that will be used from now on.

1.1 Normal-form games

A normal form game $\mathcal{G}$ is a tuple $\big(I, A, u\big)$, where:

- $I = \{1, 2, \ldots, N\}$ is a finite set of players.
- $A = A_1 \times A_2 \times \ldots \times A_N$
- A_i is the set of strategies available to player i, for each $i \in I$. It can be both countable or uncountable.
- $u = \{u_i\}_{i \in I}$, where $u_i : A \to \mathbb{R}$, is the set of utility functions of player i, which assign a real-valued payoff to each combination of strategies chosen by the players.

This defines the normal-form game $\mathcal{G} = (I, A, u)$

Definition 1.1.1 (Pure strategy profile)**.** *A pure strategy profile is an element $a \in A$*

So utility functions map pure strategy profiles into real values. When considering utility function of player i the notation $a = (a_i, a_{-i})$ and $a_{-i} \in A_{-i}$ will also be adopted.

Definition 1.1.2 (Mixed strategy)**.** *Let $\Delta(X)$ denote the set of probability distributions defined over the set X. A mixed strategy for player i is an element $\pi_i \in \Delta(A_i)$. A mixed strategy profile will be an element in $\Pi := \Delta(A_1) \times \Delta(A_2) \times \cdots \times \Delta(A_N)$. So we will not consider cases in which there is correlation between different strategy sets.*

In the case of discrete set of strategies, the utility function u_i can be extended as a map defined over Π in the following way:

$$u_i(\pi) := \sum_{a \in A} \big(\prod_{j \in I} \pi_j(a_j) \big) \cdot u_i(a)$$

(The continuous case can be equivalently extended involving integrals instead of summations, but it will not be a case of interest for this thesis so we are not going to treat it

in detail). It follows that a pure strategy a_i can be seen as a mixed strategy with all probability weight on a_i. In this sense mixed strategies are an extension of pure ones that can be seen as a subset of the former. So when not specified, *strategy profile* will refer to a mixed strategy profile and it will be understood that also pure strategy profiles will be considered this way.

Definition 1.1.3 (Nash equilibrium)**.** *A strategy profile π^* is a Nash equilibrium iff $\forall i \in I$ it holds:*

$$u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i, \pi_{-i}^*) \quad \forall \pi_i \in \Delta(A_i)$$

Definition 1.1.4 (Best response)**.** *The best response of player i is a set-valued function $BR_i : \Delta(A\_i) \to A_i$ defined by:*

$$BR_i(\pi_{-i}) = \{b \in A_i \mid \ u_i(b, \pi{-}i) \geq u_i(c, \pi_{-i}) \ \forall c \in A_i\}$$

Proposition 1. *If for the normal-form game $\mathcal{G} = (I, A, u)$ is such that:*

- *$\forall i \in I$ is a non-empty, compact, convex subset of a Euclidean space*
- *$\forall i \in I$ u_i is continuous and its restriction to A_i is quasi-concave*

$\implies \exists$ a pure strategy Nash equilibrium of the game.

1 is a consequence of the well known Kakutani's fixed point Theorem, and the proof is based on showing that a Nash equilibrium is a fixed point of the set of best response funtions seen as a unique set-valued function, and in showing that the conditions in 1 are equivalent to the conditions of Kakutani's fixed point theorem for this function. Whatever game defined with discrete strategies sets does not fullfill the hypothesis of 1, because no discrete set is concave. It is important to notice that 1 gives only a sufficient contition for the existence of equilibria, not a necessary one. A strong consequenc of 1 is the following:

Theorem Every finite normal-form game admits a mixed strategies equilibrium.

This is because each set $\Delta(A_i)$ is a closed and convex euclidean space (it is by definition the set of convex combinations of the elements of A_i), and if A_i is finite then $\Delta(A_i)$ is also bounded. But a bounded and closed euclidean space is compact and so the extended game $\tilde{\mathcal{G}} := (I, \Pi, \{u_i\}_i)$ fullfills the hypothesis of 1.

1.1.1 Method of equating payoffs

When considering Nash equilibria in mixed strategies the following has a fundamental use from an operative point of view:

Proposition 2. *Let π_i and π_{-i} be such that:*

$$u_i(\pi_i, \pi_{-i}) \geq u_i(\pi_i', \pi_{-i}), \quad \forall \pi' i \in \Delta(A_i)$$

Let's denote the support of π_i by $\mathrm{Supp}(\pi_i)$*. Then $\forall a_i \in \mathrm{Supp}(\pi_i)$ it holds that $a_i \in BR_i(\pi_{-i})$*

Proof. $u_i(\pi_i, \pi_{-i}) = \sum_{b \in \text{Supp}(\pi_i)} \pi_i(b) u(b, \pi_{-i})$. If $c \in \text{Supp}(\pi_i)$ is not in $BR(\pi_{-i})$ then the weight of $\pi_i(c)$ could be moved on any $d \in BR(\pi_{-i})$ obtaining a greater payoff, which is a contradiction. □

Notice that if two elements are in $BR(\pi_{-i})$ then they must have the same payoff. So in the hypothesis of the proposition it holds:

$$u(a, \pi_{-i}) = u(b, \pi_{-i}), \quad \forall\ a, b \in \text{Supp}(\pi_i)$$

In particular this holds for Nash equilibria in which the condition of the proposition holds mutually for all players. This gives a condition to impose in order to fix the value of a mixed strategy in case of 2-player games or in case of symmetric mixed strategies with more than 2 players.

Examples

Let's present three prototipical 2-players normal-form game. 2-players settings allow games to be represented as tables in which rows represent player 1 and columns represent player 2. In each pair of numbers, the first one is the payoff obtained by the row player (player 1) and the second one is the payoff obtained by the column player (player 2). The first two games presented have pure strategies Nash equilibria, while the third one has no pure strategies Nash equilibria.

Prisoner's Dilemma

The Prisoner's Dilemma is a classic game in game theory involving two prisoners arrested for a joint crime. Each prisoner is offered the opportunity to cooperate with the other (by remaining silent) or betray them by confessing to the crime. The possible combinations of choices lead to different payoff outcomes for both prisoners. Despite cooperation being in the collective interest, each prisoner has a personal incentive to betray the other, resulting in a suboptimal outcome for both. Notable properties include the inefficiency of the Nash equilibrium and the absence of cooperation.

	Stay Silent	**Confess**
Stay Silent	$(2, 2)$	$(0, 3)$
Confess	$(3, 0)$	$(1, 1)$

Battle of the Sexes

The Battle of the Sexes is a game that represents a situation where two players must choose between two preferred activities differently. For example, a couple may have to decide whether to go to the opera or watch a football match. Although both prefer to spend time together, their divergent interests lead them to prefer different activities. The game can lead to three possible outcomes: one where both agree on a preference, one where they become "lonely followers," and one where they end up in different places. Important properties include the presence of multiple Nash equilibria and the possibility

coordination.

	Opera	**Football**
Opera	$(3,2)$	$(0,0)$
Football	$(0,0)$	$(3,2)$

Rock-Paper-Scissors

Rock-Paper-Scissors is a hand game where two players choose one of three symbols: rock, paper, or scissors. The winner is determined by the rules: paper beats rock, rock beats scissors, and scissors beat paper. The game is often played best out of three. There are no Nash equilibria in pure strategies in this game, as no player has an option that dominates the others in every circumstance. However, it can be approached using mixed strategies, where players choose their moves with certain probabilities.

	Rock	**Paper**	**Scissors**
Rock	$(0,0)$	$(-1,1)$	$(1,-1)$
Paper	$(1,-1)$	$(0,0)$	$(-1,1)$
Scissors	$(-1,1)$	$(1,-1)$	$(0,0)$

Let's use the method of equating payoff to determine the mixed strategies Nash equilibrium of the game: Let $\pi_1 = \big(p_1, p_2, 1-(p_1+p_2)\big)$ and $\pi_2 = \big(q_1, q_2, 1-(q_1+q_2)\big)$. Let's take player 1:

$$
\begin{aligned}
u_1(Rock, \pi_2) &= 0 \cdot q_1 - q_2 + 1 - q_1 - q_2 = 1 - q_1 - 2q_2 \\
u_1(Paper, \pi_2) &= q_1 - 0 \cdot q_2 - 1 + q_1 + q_2 = 2q_1 + q_2 - 1 \\
u_1(Scissors, \pi_2) &= -q_1 + q_2 + 0 \cdot (1 - q_1 - q_2) = q_2 - q_1
\end{aligned}
$$

$$
\begin{cases} u_1(Rock, \pi_2) = u_1(Scissors, \pi_2) \\ u_1(Paper, \pi_2) = u_1(Scissors, \pi_2) \end{cases}
\rightarrow
\begin{cases} 1 - q_1 - 2q_2 = q_2 - q_1 \\ 2q_1 + q_2 - 1 = q_2 - q_1 \end{cases}
\rightarrow
\begin{cases} q_1 = \dfrac{1}{3} \\ \\ q_2 = \dfrac{1}{3} \end{cases}
$$

So $\pi_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. One should do the same with player 2 in order to find p_1 and p_2 but by the symmetry of the table we get the same system of equations. So we conclude that the mixed Nash equilibrium is

$$\pi^* = \big((\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}), (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3})\big)$$

1.1.2 *maxmin* and *minmax* strategies

Let $\mathcal{G} = \big(I, A, u\big)$ be a normal-form game:

Definition 1.1.5 (*maxmin* strategy/value)**.** *A maxmin strategy for player i against players $-i$ is:*

$$\underline{a}_i \in \arg\max_{b\in A_i} \min_{c\in A_{-i}} u_i(b,c)$$

The corresponding maxmin value for players $-i$ is:

$$\underline{v}_{-i} := \max_{b\in A_i} \min_{c\in A_{-i}} u_i(b,c)$$

Definition 1.1.6 (Minmax strategy/value)**.** *A minmax strategy for player i against players $-i$ is:*

$$\overline{a}_i \in \arg\min_{c\in A_i} \max_{b\in A_{-i}} u_{-i}(b,c)$$

The corresponding minmax value for players $-i$ is:

$$\overline{v}_{-i} := \min_{c\in A_i} \max_{b\in A_{-i}} u_{-i}(b,c)$$

Proposition 3. *In a two-player normal-form game it holds:*

$$\forall \text{ player } i \quad \underline{v_i} \leq \overline{v_i}$$

Proof. The claim can be seen as a special case of the more general (and trivial) inequalities:

$$\forall a\colon \min_c u_i(a_i,c) \leq u_i(a_i,a_{-i}) \leq \max_b u_i(b,a_{-i})$$

□

Definition 1.1.7 (Value of the game/optimal strategies)**.** *If a two-player normal-form game admits*

$$\underline{v}_i = \overline{v}_i = v$$

- *v is called value of the game*
- *The corresponding strategies are called optimal strategies*

The very same definitions can be extended to the mixed strategy case, by minimizing (resp. maximising) with respect to $\pi_i \in \Delta(A_i)$

1.1.3 Some special classes of games

Definition 1.1.8 (Two player 0-sum game)**.** *A two player normal-form game is 0-sum if: $u_1(a) + u_2(a) = 0$, $\forall a \in \mathcal{A}$*

Example: Rock-Paper-Scissors.

Proposition 4. *In a two-player 0-sum game: $\underline{v}_i = -\overline{v}_j$, $\forall i \neq j$*

Proposition 5. *If a two-player game is zero-sum then $\exists v$ value of the game in mixed strategies.*

Definition 1.1.9 (Potential game)**.** *A normal-form game is potential game if there exists a function $\Phi : A \to \mathbb{R}$, called the potential function, such that for every player i and every pair of strategy profiles $a = (a_i, a_{-i})$, $a' = (a_i, a'_{-i}) \in A$ (so differing only in the strategy of player i), the following condition holds:*

$$u_i(a) - u_i(a') = \Phi(a) - \Phi(a')$$

It follows that Nash equilibria are local maxima of the potential function Φ.

Proposition 6. *Every finite potential game has a pure strategy Nash equilibrium*

The strong properties of this class of games are way more and some of them will be of central interest for this thesis: they will be introduced later in the context of learning dynamics.

Definition 1.1.10 (Congestion game)**.** *A congestion game is defined by a tuple* $\big(I, \mathcal{R}, A, c\big)$ *game where:*

- $I = \{1, 2, \dots, N\}$ *is a finite set of players.*
- $\mathcal{R} = \{1, 2, \dots, r\}$ *is the set of resources*
- $A = A_1 \times A_2 \times \ldots \times A_N$ *and* $\forall i \in I \ A_i \subseteq 2^{\mathcal{R}} \backslash \emptyset$*. So the available strategies for a given player are subsets of* $\mathcal{R}$*.*
- $c = \{c_r\}_{r \in \mathcal{R}}$ *where* $c_r : \{1, \dots, N\} \to \mathbb{R}$ *is the cost function relative to resource* r

The resulting congestion game is the normal-form game (I, A, u) *where each* u_i *is defined by:*

$$u_i(a) = u_i(a_i, a_{-i}) := -\sum_{r \in a_i} c_r\big(n_r(a)\big)$$

where $n_r(a)$ *is the number of players playing a strategy which contains* r*.*

Every congestion game admits a potential and is in fact a potential game. It is true also the converse: every potential game shares the potential function with a suitable congestion game.

Definition 1.1.11 (Symmetric game)**.** *A normal-form game* $(I, \tilde{A}, \tilde{u})$ *is symmetric if*

- $\tilde{A}_i = \tilde{A}_j \ \forall i, j \in I$
- *for every permutation of the set of players,* $\sigma : I \to I$*, we have* $\tilde{u}_i(a_1, a_2, \dots, a_n) = \tilde{u}_{\sigma^{-1}(i)}\big(a_{\sigma(1)}, s_{\sigma(2)}, \dots, s_{\sigma(n)}\big)$*.*

In low terms, the payoff is uniquely determined by the payoff of a single player, say player 1, and doesn't change when permuting all other players actions. Let $(a_1, ..., a_n)$ *be given. We have*

$$\tilde{u}_1(a_1, ...a_j, ...a_n) = \tilde{u}_j(a_j, ...a_1, ...a_n)$$

Since there's a unique set of strategies shared by all players a symmetric game is specified by the tuple (I, A, u) *and so the notation is slightly different from the one used for other normal-form games. [***?** *]*

1.2 Repeated games

Repeated games are a generalization of normal-form games, in which a given normal-form game is played repeatedly and the utility is defined in such a way to accumulate the utilities obtained at each repetition of the game. The repeated game is then a discrete time process. It can be given different equivalent definitions: the most formal way would be to define it in terms of extensive form games, which are not of interest for this thesis. So, to avoid unnecessarily burdening this work with definitions and notions that will not be used, the choice was made to introduce repeated games in a direct and more operational manner.

Definition 1.2.1 (Finitely repeated games)**.** *A finitely repeated game is a tuple* $(\mathcal{G}, T, U, \mathbb{A})$ *where:*

- $\mathcal{G} = (I, A, u)$ *is a normal-form game also denoted as stage game of the repeated game.*
- T *is a positive integer, denoting how many times the stage game is repeated*
- $\mathbb{A} := \times_{i=1}^{N}$ *where* $\mathbb{A}_i := \left\{a_i^{(\cdot)} : \{1, 2, \ldots, T\} \to A_i\right\}$*. So strategies are maps from the set of iterations (denoted times) to the strategies of the stage game, i.e. mapping each time* t *to an element of* A_i*:* $t \mapsto a_i^t \in A_i$*. We can extend the definition to mixed strategies straightforwardly: a mixed strategy for player* i *is a map:* $\pi_i^{(\cdot)} : \{1, 2, \ldots, T\} \to \Delta(A_i)$
- $U := \{U_i\}_{i=1}^{N}$*, where* $U_i(\pi) := \sum_{t=1}^{T} u_i(\pi^t)$

The class of repeated games that will be of interest for this thesis are infinitely repeated games, which can't be simply defined as the $lim_{T\to\infty}$ of a finitely repeated game, because the utility would give always infinities or zeros.

Definition 1.2.2 (Finitely repeated games)**.** *A finitely repeated game is a tuple* $(\mathcal{G}, U, \mathbb{A})$ *where:*

- $\mathcal{G} = (I, A, u)$ *is a normal-form game.*
- $\mathbb{A} := \times_{i=1}^{N}$ *as in the finitely repeated games.*
- $U := \{U_i\}_{i=1}^{N}$*, where* $U_i(\pi)$ *is a bounded function* $\forall$ *sequence* $\{u_i(\pi^t)\}_{t=1}^{\infty}$

Two forms of U_i often adopted are:

- $U_i(\pi) = \frac{1}{T}\sum_{t=1}^{\infty} u_i(\pi^t)$
- $U_i(\pi) = \sum_{t=1}^{\infty} \gamma^t \cdot u_i(\pi^t)$, where $0 < \gamma < 1$ is the discount factor (notational ambiguity: here γ^t means γ to the power of t, while in π^t it is just a label for the strategy profile at time t).

The latter will be taken as the standard choice from now on. This form introduces a new parameter, the discount factor, which represents a way to tune the weight or importance that players assign to future payoffs compared to immediate payoffs. A higher discount factor typically encourages more cooperative behavior in repeated games. This is because players are more inclined to consider the long-term consequences of their actions and are willing to cooperate to achieve mutually beneficial outcomes over time. Conversely, a lower discount factor may lead to more short-term, self-interested behavior. Cooperation can be sustained in infinitely repeated games when the discount factor is sufficiently high. This is because players have a stronger incentive to maintain cooperative strategies, as the future benefits of cooperation outweigh the short-term gains from defection.

We will focus on stationary strategies, i.e. such that $\pi^t = \pi \ \forall t$. We can extend to this class of strategies the notion of Nash equilibrium straightforwardly: the stationary strategy profile π^* is a Nash equilibrium of the repeated game iff

$$U_i(\pi_i^*, \pi_{-i}^*) \geq U_i(\pi_i', \pi_{-i}^*) \qquad \forall i \in I \text{ and } \forall \pi_i' \in \Delta(A_i)$$

Chapter 2

Learning in games

The study of learning in games plays a fundamental role in applying game theory to model diverse contexts such as behavioral economics and evolutionary biology. It elucidates how rational agents adapt their strategies over time in response to past experiences and environmental feedback.

2.1 Fictitious play

This section concerning Fictitious Play is mainly based on [4].

2.1.1 The model

The Fictitious Play approach involves the assumption that each player believes their opponent is using a fixed, yet unknown, steady strategy. The aim of the learner is to infer such (fictitious) strategy and to play the best response against it. At each time step the learner updates its belief about the opponent's strategy using statistics extracted from opponent's previous strategies history up to the preceiding time step. At timestep $t = 0$ a weight function $\kappa_i^0 : A_{-i} \longrightarrow \mathbb{R}^+$ is assigned to each player i. Then at each time step t such weights are updated as follows:

$$1.\ \kappa_i^t(a_{-i}) = \kappa_i^{t-1}(a_{-i}) + \delta(a_{-i}^{t-1}\,,\, a_{-i})$$

Essentially $\kappa_i^t(a_{-i})$ counts how many times strategy a_{-i} has been played up to time step $t-1$ and up to a constant term $\kappa_i^0(a_{-i})$ which plays the role of a prior belief. The belief $\delta_i^t \in \Delta(A_{-i})$ is then naturally defined as:

$$2.\ \delta_i^t(a_{-i}) = \frac{\kappa_i^t(a_{-i})}{\sum_{s \in A_{-i}} \kappa_i^t(s)}$$

The belief is interpreted by the learner as the best estimate of the opponent's steady (mixed) strategy in order to play the best response strategy against it.

Remark 1. *δ_i^t is not just the empirical average unless one sets $\kappa_i^0 = 0$. In fact also the empirical average itself will be taken into account, as specified later.*

Because the best response function BR_i: $\Delta(A_{-i}) \longrightarrow A_i$ is generally a set-valued function, specifying a Fictious Play model means (also) specifying a rule according to which choosing a specific element in $BR_i(a_{-i})$. Traditionally in case of indifference between different strategies the rule consists in choosing one of them and not in randomizing between them[Fudenberg].

2.1.2 Convergence to a steady state

We can define the state of the Fictitious Play at time t as the strategy profile played at time t. In case the process of the Fictitious Play converges to steady state then the assumption of the model is consistent. Let's consider the conditions under which the Fictitious Play converges to a steady state. First of all it is necessary to define it.

Definition 2.1.1 (Steady state)**.** *A state $\tilde{a}$ is steady iff it exists a finite time T after which $\tilde{a}$ is played in every time step.*

At this point a first result is achieved through the following proposition:

Proposition 7. *Let $\tilde{a}$ be a strict N.E..*

- *If strategy profile $\tilde{a}$ is played at time t then $\tilde{a}$ will be played at every later time*
- *If Fictitious Play reaches a steady state then it is a pure strategy N.E.*

Proof.

- If a strict N.E $\tilde{a}$ is played at time t it means that $\forall i$: $\tilde{a}_i \in BR^i(\delta_i^t)$ which means that $\forall$ player i and $\forall a_i \in A_i$: $u^i(\tilde{a}_i, \delta_i^t) > u^i(a, \delta_i^t)$.
  $\tilde{a}$ is a strict N.E. so $\forall$ player i and $\forall a_i \in A_i$: $u_i(\tilde{a}_i, \tilde{a}_{-i}) > u^i(a, \tilde{a}_{-i})$.
  According to 2 we see that $\delta^i_{t+1}(a_{-i}) = \alpha\delta(\tilde{a}_{-i}, a_{-i}) + (1-\alpha)\delta^i_t(a_{-i})$ with $\alpha \in [0,1]$ [1]. By linearity of the utility function it follows: $u^i(a_i, \delta_i^{t+1}) = \alpha u(a_i, \tilde{a}_{-i}) + (1-\alpha)u_i(s_i, \delta_t^i)$. So also for $t+1$ it is true that $\tilde{a}_i \in BR(\delta_i^{t+1})$ and it is unique because the equilibrium is strict. So $\tilde{a}$ will be definitely played.

- If a steady state $\tilde{a}$ is reached it means that $\forall$ player i: $\lim_{t\to\infty} \delta_i^t(a_{-i}) = \delta(a_{-i}, \tilde{a}_{-i})$. So if $\tilde{a}$ is not a N.E. $\Rightarrow \exists j$ and $\tilde{\tilde{a}}_j$: $u_j(\tilde{\tilde{a}}_j, \tilde{a}_{-j}) > u_j(\tilde{a}_j, \tilde{a}_{-j})$. So the Fictitious Play would deviate to such strategy, which is a contradiction. So $\tilde{a}$ is a N.E. (and it is a pure strategy one since states are pure strategy profiles).

□

It is important to highlight a direct consequence: since the only steady states the Fictitious Play can converge to are N.E. , it means that it won't converge in case of a game with only mixed strategy N.E.

Definition 2.1.2. *Let d_t^i be the simple frequency count of the strategies played by i, which can be defined using the opponents κ_t^{-i}:*

$$d_t^i(a_i) = \frac{\kappa^t_{-i}(a_i) - \kappa^0_{-i}(a_i)}{t}$$

Such frequency count will be referred to also as "empirical distribution", for obvious reasons. We can now state the second result:

Proposition 8. *If $\forall$ player i, d_t^i converges to d^i, then $\tilde{\sigma} := (d^1, d^2)$ is a mixed strategy N.E.*

[1] α is a parameter depending on t: at $t = 0$ it is just $\alpha_0 = \sum_{s\in A_{-i}} \kappa_i^0(s)$, then it can be simply updated as $\alpha_t = \alpha_0 + t$

Proof. In case all d_t^i converge to d^i, then it holds:

$$d^i = \lim_{t\to\infty} \frac{\kappa_{-i}^t - \kappa_{-i}^0}{t} = \lim_{t\to\infty} \frac{\kappa_{-i}^t}{t + \sum_{a\in A_i} \kappa_i^0(s)} = \lim_{t\to\infty} \delta_{-i}^t$$

where, in the second equality, two terms have been added or subtracted as they don't depend on t and are negligible with respect to the other terms which are $\mathcal{O}(t)$. So also δ_{-i}^t converges and it converges to d. Asymptotically, at each stage player i plays the best response to d_{-i} and the sequences of choices as a whole must be a best response to d_{-i}, unless one could want to deviate, which is a contradiction with the fact that d_i is stationary. As this holds for both players, then (d_1, d_2) must be a N.E. □

One may wonder if there are specific conditions that ensure convergence of the empirical distribution. A few results can be found in literature:

Proposition 9. *The empirical distributions converge if the game has generic payoffs*[2] *and holds at least one of the following:*

- *the stage game is* 2×2
- *the stage game is 0-sum*
- *the stage game is solvable by I.E.S.D.S*

A fourth more technical case can be found in the literature.

2.2 Reinforcement learning

[6] [8] Very generally speaking, reinforcement learning (RL) is a type of machine learning paradigm where an agent learns to make decisions by interacting with an environment in order to achieve certain goals or maximize some notion of cumulative reward (in the case of learning in games the goal is maximizing the utility function). Unlike supervised learning where the model learns from labeled input-output pairs or unsupervised learning where the model learns patterns in data without explicit supervision, reinforcement learning is based on trial and error learning through exploration and exploitation of the environment: exploitation means that the agent uses the information accumulated during the learning in order to "adapt" optimally its behaviour, while exploration means that the agent never remains rigidly constrained to the strategy/behavior currently considered optimal, but always seeks, at least in part, to explore new paths. In reinforcement learning, the agent observes the current state of the environment, takes an action, and then receives feedback in the form of a reward signal indicating how good or bad that action was in the given state. As already mentioned, in the case of interest for this thesis, the goal of the agent is to learn a policy (a mapping from states to actions) that maximizes the cumulative reward over time.
Key components of reinforcement learning include:

1. **Agent/agents**: The entity/entities that learn and make decisions based on the interactions with the environment.

[2] def of generic payoff

2. **Environment**: The external system or process with which each agent interacts. With multiple agents, the other agents are part of the environment.
3. **State**: A representation of the current situation or configuration of the environment.
4. **Action**: The decision or behavior chosen by the agent based on the current state.
5. **Reward**: The feedback signal provided by the environment to evaluate the action taken by the agent.
6. **Policy**: The strategy or rule that the agent follows to select actions based on states.
7. **Value Function**: A function that estimates the expected cumulative reward of being in a certain state or taking a certain action in a given state.

Multi-agent reinforcement learning is (not surpsisingly) more complex than single-agent, since: the environment, of which the other agents are part, is actively interacting and the resulting dynamics is more difficult to predict. What's more, it is not straightforward what the goal of a learning process should be as in general no univoque solution concept is defined: one could try to find an equilibrium, or to maximize the global utility and these are just two examples. Therefore, we will introduce the basic idea underlying all the RL approaches that will be explored in this thesis (Q-learning) in the context of single agent learning, then look naturally at the multi-agent versions as generalizations.

2.2.1 Markov decision process and Bellman equation

A natural way to introduce Q-learning is starting from a stochastic processes context, in which self-consistent equations concerning the optimal value of an agent arises. Let's start by briefly recalling what a Markov process is. A discrete time Markov process is a stochastic model used to describe a sequence of events or states over discrete time intervals, where the future state depends only on the current state and not on the sequence of events that preceded it.
Formally, let $\{s_t\}$ be a sequence of random variables representing the states of the process at discrete time instants $t = 0, 1, 2, \ldots$. The process is said to satisfy the Markov property if the conditional probability of transitioning to a future state s_{t+1} given the current state s_t and all previous states $s_{t-1}, s_{t-2}, \ldots$ depends only on s_t. This can be expressed as:

$$P(s_{t+1} = j | s_t = i, s_{t-1}, s_{t-2}, \ldots) = P(s_{t+1} = j | s_t = i)$$

for all states i and j, where P is the transition matrix. The following shorthand notation will be adopted: $P(s_{t+1} = j | s_t = i) = P_{s_t, s_{t+1}}$
In other words, the future behavior of the process is determined solely by its current state, and not by the history of how it arrived at that state. This property is also known as memorylessness.

A Markov decision process (MDP) is a generalization of a discrete time Markov process, in which at each timestep an agent has to choose an action, receiving a reward dependent

on the action chosen. From the point of view of the process, several transition matrices are defined and the agent's choices consist essentially in choosing on of such transition matrices at each time step. Formally a MDP is univocally defined by:

- A set of states $\mathcal{S}$
- An action set A
- A utility function u: $\mathcal{S} \times A \longrightarrow \mathbb{R}$
- A transition function T: $\mathcal{S} \times A \longrightarrow \Delta(\mathcal{S})$ which $\forall a \in A$ defines a transition matrix $P^a_{s,s'}$ with $s, s' \in \mathcal{S}$

In this context, a strategy for the agent can be seen as a policy:

Definition 2.2.1 (Policy)**.** *We define a policy to be a mapping from the set of states $\mathcal{S}$ to the set of probability distributions over the actions set A:*

$$\begin{aligned} \pi : \quad & \mathcal{S} \to \Delta(A) \\ & s \mapsto \pi_s \end{aligned}$$

Through π_s the definition of T and u on A can be extended to $\Delta(A)$:

$$\begin{aligned} P^{\pi_s}_{s,s'} &:= \sum_{a \in A} \pi_s(a) P^a_{s,s'} \\ u(s, \pi_s) &:= \sum_{a \in A} \pi_s(a) u(s, a) \end{aligned}$$

Given a policy π, at each possible sequence of states $\{s_t\}_{t=0}^{\infty}$ a probability measure $\mathcal{P}\big(\{s_t\}_{t=0}^{\infty}\big)$ can be associated. Given a discount factor $\gamma \in [0, 1]$, the purpose of the agent is to maximize the expected discounted payoff:

$$U\big(\pi\big) := \mathbb{E}_{\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot u(s_t, a_t)\right]$$

Let's define the optimal value $V(s)$ of the process as the maximum over the set of policies of expected discounted payoff given the initial state s:

$$V(s) := \max_{\pi} U(\pi)\big|_{s_0 = s}$$

$V(s)$ must obey the following self-consistent equation:

$$V(s) = \max_{\pi_s} \left[u(s, \pi_s) + \gamma \sum_{s' \in \mathcal{S}} P^{\pi_s}_{s,s'} V(s')\right]$$

in which the optimal value associated to state s has been decomposed into the immediate reward obtained and the discounted value of the future states reachable from that state. This equation is the so called Bellman equation and in the context of Markov decision processes has a very simple structure due to the Markov property.

The Bellman equation can be rewritten by explicitly expressing the average over the probability distribution of the policy:

$$V(s) = \max_{\pi_s} \left[\sum_{a \in A} \pi_s(a) \left(u(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{s,s'} V(s') \right) \right]$$

From which one can define a state-action quality function $Q(s,a) := u(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{s,s'} V(s')$

Hence the Bellman equation can be written in terms of Q

$$\begin{cases} Q(s,a) = u(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{s,s'} V(s') \\ V(s) = \max_{\pi_s} \left(\sum_{a \in A} \pi_s(a) Q(s,a) \right) \end{cases}$$

This is the formulation of the Bellman equation on which Q-learning is based.

2.2.2 Q-learning

From the still purely mathematical formulation of the optimal solution of the MDP there is an interest in developing effective numerical methods for optimizing decision-making. Q-learning, one of the main reinforcement learning algorithms, addresses this request: the algorithm's objective is to iteratively refine the state-action quality function Q enabling an agent to make decisions that maximize the expected payoff over time. It is important to underline that Q-learning does not refer to a single, specific algorithm: in fact, it represents a general framework for reinforcement learning methods that share the fundamental objective of refining the state-action quality function to optimize decision-making.
Given the MDP defined by $(\mathcal{S}, A, u, T)$, algorithm 1, taken from [6] is a possible specific but still general formulation of a Q-learning algorithm. The learning rate α tunes how much the algorithm trusts new observations with respect to past history: it can be fixed or another possible choice is to initialize $\alpha = 1$ and let it decrease up to 0, in a way specified by α_t. The *explore* parameter, tunes how often the algorithm will deviate from the current policy estimate, in order to ensure an adequate exploration of the policies space.

Remark 2. *1 does not use the Markov matrix of the process in the learning, only in the simulation of the process: it means that it does not suppose the agents to know anything about the probabilistic process, only the states. This kind of approach will be referred to as "model-free" from now on. This kind of models use the observation of the current state at time $t+1$ in order to update all the values concerning time t, and use $V(s_{t+1})$ as an estimator for $\sum_{s \in \mathcal{S}} P_{s_t,s} V(s)$*

Algorithm 1 Q-learning

Initialize

- Value function $V(s)$, $\forall s \in \mathcal{S}$
- Action-value function $Q(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in A$
- Policy $\pi(s,a) = \dfrac{1}{|A|}$, $\forall s \in \mathcal{S}$, $\forall a \in A$
- Initial learning rate $\alpha = 1$
- Deviation probability $explore \in [0,1]$
- Discount factor $\gamma \in [0,1]$

Repeat

- With probability *explore* choose uniformly an action a_t from A.
  Otherwise observe current state s_t and choose an action a_t with probability distribution $\pi(s_t, \cdot)$.
- Observe new state s_{t+1}
- Update Q, V, π and α:

$$Q(s_t,a_t) \longleftarrow \cdot Q(s_t,a_t) + \alpha \cdot \big(u(a_t,s_t) + \gamma V(s_{t+1}) - Q(s_t,a_t)\big)$$
$$V(s_{t+1}) \longleftarrow \max_{p \in \Delta(A)} \sum_{a \in A} p(a) \cdot Q(s_{t+1},a)$$
$$\pi(s_{t+1}, \cdot\ ) \longleftarrow \arg\max_{p \in \Delta(A)} \sum_{a \in A} p(a) \cdot Q(s_{t+1},a)$$
$$\alpha \longleftarrow \alpha_t$$

Notice that the maximization step can be very demanding from a numerical point of view when the cardinality of A grows: often the choice is to perform analitically the maximization by introducing a temperature parameter, that allows an easy computation of the arg max and at the same time induces exploration (in the version above the exploration is manually inserted through the *explore* probability parameter):

$$\pi(s_{t+1}, \cdot\ ) \longleftarrow \frac{exp\big(\frac{Q(s_{t+1},b)}{\tau}\big)}{\sum_{b \in A} exp\big(\frac{Q(s_{t+1},b)}{\tau}\big)}$$

This update is equivalent to modifying algorithm 1 in such a way that rhe third step in the *repeat* section is

$$\pi(s_{t+1}, \cdot\ ) \longleftarrow \arg\max_{p \in \Delta(A)} \left(\sum_{a \in A} p(a) \cdot Q(s_{t+1},a) - \tau \sum_{a \in A_i} p(a) \cdot logp(a) \right)$$

i.e. the maximization leads to selecting p that maximizes the expected Q but also maximize an entropy term tuned by τ. This can be obtained by simply adding the entropy term in the *argmax* step, or by adding it in the definition and update of V:

$$V(s_{t+1}) \longleftarrow \max_{p \in \Delta(A)} \left(\sum_{a \in A} p(a) \cdot Q(s_{t+1}, a) - \tau \sum_{a \in A} p(a) \cdot logp(a) \right)$$

2.2.3 Q-learning in repeated games

The same idea of optimizing a Q-function can also be applied in the context of repeated games. One perhaps needlessly convoluted way of viewing this is by noting that a repeated game can be seen as an MDP (Markov Decision Process) where the set of states is a singleton. However, unlike MDPs, repeated games are multiple agents decision processes, so the simplification introduced by no longer having a dependence on states (and effectively not having a stochastic process, except for the stochasticity associated with mixed strategies) is offset by the fact that there is no longer a unique criterion for optimizing the Q-function (a criterion which in the single-agent case was simply to maximize). Clearly, just as each agent is associated with a different utility function, each agent will be associated also with a different Q-function. But some attention has to be payed: RL for MDP allows a single agent to maximize a delayed reward in a stochastic but stationary environment. Convergence to the optimal policy is guaranteed with sufficient experimentation. Multi-agent reinforcement learning extends beyond MDPs, incorporating dependencies between agents' policies, so the environment is not stationary anymore.
Reinforcement learning enables single-agent optimal behavior through trial-and-error. However, when multiple learners act in a shared environment, traditional approaches not necessarily work. Assumptions for convergence are often violated in the multi-agent setting, requiring coordination even when objectives align. With opposing goals, achieving equilibrium becomes crucial.

Remark 3. *As already stated in the previous section, if not specified differently we are going to consider pure strategies as a subset of mixed strategies. So whenever a solution in terms of mixed strategies is searched, or a maximization over the possible mixed strategies is performed, the very same can be done in terms of pure strategies: it is sufficient to restrict the search or the maximization to the subset of mixed strategies which are delta distributed. When referring to a strategy, the term "mixed" will be used only to recall that in the most general case it is a probability distribution and not to exclude the pure strategies subset.*

Base case

One of the simplest algorithm incorporating the main features of RL is the one proposed in [5]:

$$\begin{aligned} q_i(a_i^t) &\longleftarrow q_i(a_i^t) + u_i(a^t) \\ \pi_i(a_i) &\longleftarrow \frac{q_i(a_i)}{\sum_{b \in A_i} q_i(b)} \qquad \forall a_i \in A_i \end{aligned}$$

At each time step an action is chosen using π, then the realized payoff is used in a very simple way to update the value function at the chosen action, finally the whole π is updated. It is evident that a positive payoff obtained when choosing a_i^t will lead to increase the corresponding probability. The exploration has to be "inserted" by choosing at random a strategy (instead of following π). Alternatively one could use a softmax update of π using the temperature parameter to tune exploration

Bellman equation inspired approach

Let's look at an analogous of the Bellman equation for this mult-agent stateless setting:

$$\begin{cases} Q_i(a) = u_i(a) + \gamma \cdot V_i \\ V_i = \mathcal{F}\big(\pi_i, Q_i\big) \end{cases}$$

Here $\mathcal{F}$ is a functional substituting (and in some sense generalizing) the maximization of the orginal Bellman equation, and is the one characterizing the different Q-learning approaches. In defining it, one has to keep in mind that for each player i it should lead to a maximization of the Q_i function (which is directly related to the utility function of player i), and possibly (not necessarily) using the structure of the game in doing that. Common choices are:

- $\mathcal{F}\big(\pi_i, Q_i\big) = \max_{\pi_i} \min_{a_{-i}} \sum_{a \in A} \pi_i(a_i) \cdot Q_i(a)$ i.e. each player maximizes over its own strategies after minimizing over opponent's pure strategies. This is a well suited approach in case of 2-player zero-sum games that, as shown in section 1.1 are strongly related with the *maxmin* structure.

- $\mathcal{F}\big(\pi_i, Q_i\big) = \max_{\pi_i} \sum_{a \in A} \pi_i(a_i) \cdot \pi_{-i}(a_{-i}) \cdot Q_i(a)$ of course this is only a theoretical definition but in practice it makes no sense for agent i to be aware of what the strategies of its opponents are: in practice instead of π_{-i} a belief μ_i (see fictitious play) about the opponents strategies is used. This version of Q-learning can be formulated in such a way that the belief μ_i is built up in the same way as in the classical fictitious play. This approach has the advantage to exploit the strenght of the fictitious play approach.

In both cases the considerations on the numerical difficulties of maximizing over probability distributions and of using a direct update introducing a temperature parameter hold.

Another class of approaches comes from the idea of studying the dynamics of an averaged version of the Q-function:

$$q_i(a_i) := E_{\pi_{-i}}[Q_i(a)] = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(a_{-i}) \cdot Q_i(a)$$

This way the Q-function does not depend anymore on opponent's action and the Bellman equation has the same form of the single agent setting (but with no states), and is straightforwardly obtained by averaging the equation for $Q(a)$:

$$\begin{cases} q_i(a) = E_{\pi_{-i}}[u_i(a)] + \gamma V_i \\ V_i = \max\limits_{\pi_i \in \Delta(A_i)} \sum\limits_{a_i \in A_i} \pi_i(a_i) \cdot q_i(a_i) \end{cases}$$

where again one can add the entropy term in the definition of V_i or only in the *argmax* step of the algorithm. Despite the formal definition, in practice the term $E_{\pi_{-i}}[u_i(a)]$ is replaced by the actually realized payoff.

A paradigmatic formulation of RL using this approach that will be set as reference point of this thesis from now on is:

$$\begin{cases} q_i(a_i^t) \longleftarrow q_i(a_i^t) + \alpha \left(u_i(a^t) + \gamma \big(\sum\limits_{b \in A_i} \pi_i(b) q_i(b) \big) - q_i(a_i^t) \right) \\ \pi_i(a_i) \longleftarrow \dfrac{exp(\frac{q_i(a_i)}{\tau})}{\sum\limits_{b \in A_i} exp(\frac{q_i(b)}{\tau})} \qquad \forall a_i \in A_i \end{cases} \tag{2.1}$$

Remark 4. *Notice that the update of π_i makes $\sum\limits_{b \in A_i} \pi_i(b) q_i(b)$ a noisy estimator of V_i*

Remark 5. *Notice that in (2.1) , q_i^t is updated only at the effectively realized action a_i^t. Clearly the same cannot be done for the π_i^t. This kind of update is known as asyncronous update, since different components of the q_i evolve at different speeds being updated less frequently. In order to compensate this asymmetry, often the learning rate α is also proportional to $\dfrac{1}{\pi_i(a_i^t)}$, because less chosen actions correspond to less frequently updated components and we want to have a q_i that evolves uniformly. In the case of learning rates decreasing in time (presented for the stochastic games generalization in section 3.1) the compensation can be obtained by having a different learning rate for each action and taking into account also for difference in the frequency of an action.*

Some insights on the convergence of such algorithms have been found analyzing them in the context of evolutionary game theory, that we are going to introduce in the next section.

2.3 Evolutionary game theory

[12][3] Evolutionary game theory is a game theory framework that studies the evolution of strategies in populations of interacting agents over time. It combines principles from game theory and evolutionary biology to understand how behaviors, represented as strategies, emerge and persist in dynamic environments. Evolutionary game theory undergoes three significant shifts from "traditional" game theory. Firstly, the notion of strategy is reinterpreted: while classical game theory involves players selecting strategies

from predefined sets, the idea is introduced that species possess sets of strategies, or genotypic variants, inherited or acquired through mutation. This concept can be extended to human culture, where society harbors a variety of cultural forms from which individuals choose or inherit.
Secondly, the evolutionarily stable strategy (ESS) was introduced as an alternative to the Nash equilibrium. An ESS refers to a strategy that, when adopted by an entire population, remains resistant to invasion by a minority group with a mutant genotype. This concept also applies to cultural forms within societies, ensuring stability against invasion by alternative cultural forms.

Consider a symmetric normal-normal form game (I, A, u) (recall that here A is the set of strategy available to each player). Essentially the re-interpretation of strategies consists in the following: in classical game theory, a mixed strategy π_i for player i is a probability distribution over A: so π_{a_i} is the probability that player i chooses action $a_i \in A$.

In evolutionary game theory we can start by considering the same distribution π_i over A, where now A represents the set of features available to each agent in a given population, population which is supposed to be very large (cardinality $>> 1$, we never specify a value) and well-mixed: more precisely, each agent can be of a certain type, and this type is specified by $a_i \in A$. For the moment the index i has no meaning: there is not a correspondence between agents (which are in a very large number and may be considered even uncountable) and the label i, so we take it just as part of the name. $\pi_(a_i)$ is the fraction of agents in this population having the feature a_i (or of the type a_i). Let for the moment consider (I, a, u) to be a 2-player symmetric normal form game: $u(a, b)$ is, in this interpretation, a measure of the fitness that an individual of the type a has when interacting with an individual of the type b. In a N-player setting so, $u(a_1, ..., a_N)$ is the fitness that an individual of the type a_1 has when interacting with other $N-1$ individuals of the type $a_2, .., a_N$. So in this context, the "old" number of player is now the grade of interactions. This explains also the choice of setting in a symmetric game.

The mathematical tools are the same, so the utility functions in this framework still have the linearity properties that allowed the extension of their definition from pure to mixed strategies (see section 1.1). Conversely the point of view has completely changed, since the randomization over pure strategies coming from rationality hypothesis has been replaced by a purely mechanistic process of selection. What's more, the game is in some sense internal: in its simplest formulation, the aim is to model an evolutionary process leading to a stable strategy from an evolutionary perspective, where it's important to note that strategy, in this context, has to be intended as the proportion of subpopulations with a specific feature.

So we can finally re-introduce labels i and $-i$ with much care: $u(a_i, a_{-i})$ has to be intended as the fitness of an individual of type spcified by a_i interacting with individuals specified by types $a_{-i} = (a_1, ..., \not{a}_i, ..., a_N)$. Being in a symmetric game, such fitness doesn't depend on the order of the latter. Even more attention has to be paid when dealing with pupulation distributions (the old mixed strategies):

Consider a symmetric mixed strategy $\tilde{\pi}$, i.e. such that $\pi_i = \pi_j = y \ \forall i, j \in I$. We adopt the notation $\pi = (y, y^{N-1})$ in order to highlight the symmetry.

- $$u(y, y^{N-1}) = \sum_{(a_1,...,a_N)} \left(\prod_{i \in I} y(a_i) \right) u(a_1, ..., a_N)$$

  is interpreted as the mean fitness of the individuals belonging to a population distributed according to y.

- Consider the same symmetric strategy y

  $$u\left(y, \epsilon\tilde{y}^{N-1} + (1-\epsilon)y^{N-1}\right) = \epsilon u(y,\ \tilde{y}^{N-1}) + (1-\epsilon)u(y, y^{N-1})$$

  is interpreted as the mean fitness of the individuals belonging to a population distributed according to y when a fraction ϵ of the population has a different distribution $\tilde{y}$

2.3.1 Evolutionarily stable strategy

The concept of evolutionarily stable strategy (ESS) is a concept, introduced in evolutionary game theory, that refers to a strategy that proves resistant when embraced by a population to adapt to a particular environment. In essence, it cannot be replaced by a different strategy, when the new strategy is initially sufficiently uncommon.

Definition 2.3.1. *A symmetric mixed strategy y for a symmetric game is evolutionarily stable iff $\exists \epsilon > 0$ sufficiently small such that $\forall \tilde{y}$*

$$u\left(y,\ (1-\epsilon)y^{N-1} + \epsilon\tilde{y}^{N-1}\right) > u\left(\tilde{y},\ (1-\epsilon)y^{N-1} + \epsilon\tilde{y}^{N-1}\right)$$

It can be characterized comparing the definition order by order in ϵ. Indeed by linearity the definition can equivalently be expressed as

$$(1-\epsilon)u(y, y^{N-1}) + \epsilon u(y, \tilde{y}^{N-1}) > (1-\epsilon)u(\tilde{y}, y^{N-1}) + \epsilon u(\tilde{y}, \tilde{y}^{N-1})$$

and one finds that y is evolutionarily stable iff either

$$u(y, y) > u(\tilde{y}, y^{N-1})$$

(i.e. strict Nash equilibrium condition) or $u(y, y^{N-1}) = u(\tilde{y}, y^{N-1})$ and $u(y, \tilde{y}^{N-1}) > u(\tilde{y}, \tilde{y}^{N-1})$

2.3.2 Replicator dynamics

The concept of mixed strategy, as we have seen, has an interpretation as a distribution of features or types within a population. It can therefore be viewed as a genuine state of the population. By introducing a temporal dependency, one can define a dynamics of this state.
Replicator dynamics is a concept in evolutionary game theory that describes such evolution of strategies within a population over time. It is often used to model the dynamics of biological populations, economic agents, or social groups engaged in strategic interactions. In its simplest form, replicator dynamics can be represented by the following differential equation:

$$\frac{dy(a_i)}{dt} = y(a_i)\big[u(a_i, y^{N-1}) - u(y, y^{N-1})\big] \tag{2.2}$$

This equation describes how the proportion of individuals having feature a_i changes over time ($\frac{dy_i}{dt}$) based on the difference between the payoff $u(a_i)$ and the average payoff of the population.
Replicator dynamics capture the idea that strategies with higher payoffs relative to the population average will increase in frequency, while strategies with lower payoffs will decrease. This dynamic process mirrors the principles of natural selection, where successful traits are favored and spread throughout a population over time.

- A fixed point of the replicator dynamics is not necessarily (actually generally is not) a maximizer of the average fitness of the population.
- $y(a_i) = 0$ and $y(a_i) = 1$ are both fixed points of the replicator dynamics: the former is trivial and the latter comes from the observation that $u(y(a_i) = 1, y^{N-1}) = u(a_i, y^{N-1})$
- a Nash equilibrium y^* is a fixed point since for any action a_i in the support of y^*, Nash equilibrium, it holds

$$u(a_i, y^{*N-1}) = u(y^*, y^{*N-1})$$

- Every Nash equilibrium that is an ESS is a stable fixed point of the replicator dynamics.

2.4 Replicator dynamics as continuous time limit of Q-learning dynamics

Using a procedure presented in [13] we can show that for a dynamics of the form

$$\begin{cases} q_i^{t+1}(a_i) = q_i^t(a_i) + \alpha\left(u_i(a_i, a_{-i}^t) + \gamma\big(\sum_{b\in A_i} \pi_i^t(b) q_i^t(b)\big) - q_i^t(a_i)\right) \\ \pi_i^t(a_i) = \dfrac{exp\big(\frac{q_i^t(a_i)}{\tau}\big)}{\sum_{b\in A_i} exp\big(\frac{q_i^t(b)}{\tau}\big)} \end{cases} \tag{2.3}$$

one can write a continuous time limit equation for $\pi_i^t(a_i)$:

$$\frac{d}{dt}\pi_i^t(a_i) = \alpha\left[\frac{1}{\tau}\cdot\pi_i^t(a_i)\cdot\big(u(a_i, \pi_{-i}^t) - u_i(\pi_i^t, \pi_{-i}^t)\big) + \pi_i^t(a_i)\cdot\sum_{b\in A_i}\pi_i^t(b)\cdot log\frac{\pi_i^t(b)}{\pi_i^t(a_i)}\right] \tag{2.4}$$

which has the form of the replicator dynamics plus an entropy term.

Let's see how:

$$\frac{\pi_i^{t+1}(a_i)}{\pi_i^t(a_i)} =$$

$$= \frac{exp\big(\frac{q_i^{t+1}(a_i)}{\tau}\big) \cdot \sum_{b\in A_i} exp\big(\frac{q_i^t(b)}{\tau}\big)}{\sum_{b\in A_i} exp\big(\frac{q_i^{t+1}(b)}{\tau}\big) \cdot exp\big(\frac{q_i^t(a_i)}{\tau}\big)} =$$

$$= \frac{exp\big(\frac{q_i^{t+1}(a_i)}{\tau}\big) \cdot exp\big(\frac{-q_i^t(a_i)}{\tau}\big) \cdot \sum_{b\in A_i} exp\big(\frac{q_i^t(b)}{\tau}\big)}{\sum_{b\in A_i} exp\big(\frac{q_i^{t+1}(b)}{\tau}\big) \cdot exp\big(\frac{-q_i^t(b)}{\tau}\big) \cdot exp\big(\frac{q_i^{t+1}(b)}{\tau}\big)} =$$

$$= \frac{exp\big(\frac{\Delta q_i^t(a_i)}{\tau}\big)}{\sum_{b\in A_i} \pi_i^t(a_i) exp\big(\frac{\Delta q_i^t(b)}{\tau}\big)} \qquad \text{where } \Delta q_i^t(b) := q_i^{t+1}(a_i) - q_i^t(a_i)$$

We just obtained that:

$$\pi_i^{t+1}(a_i) = \pi_i^t(a_i) \cdot \frac{exp\big(\frac{\Delta q_i^t(a_i)}{\tau}\big)}{\sum_{b\in A_i} \pi_i^t(a_i) exp\big(\frac{\Delta q_i^t(b)}{\tau}\big)}$$

so:

$$\pi_i^{t+1}(a_i) - \pi_i^t(a_i) = \pi_i^t(a_i) \cdot \left(\frac{exp\big(\frac{\Delta q_i^t(a_i)}{\tau}\big)}{\sum_{b\in A_i} \pi_i^t(a_i) exp\big(\frac{\Delta q_i^t(b)}{\tau}\big)} - 1 \right) =$$

$$= \pi_i^t(a_i) \cdot \left(\frac{exp\big(\frac{\Delta q_i^t(a_i)}{\tau}\big) - \sum_{b\in A_i} \pi_i^t(a_i) exp\big(\frac{\Delta q_i^t(b)}{\tau}\big)}{\sum_{b\in A_i} \pi_i^t(a_i) exp\big(\frac{\Delta q_i^t(b)}{\tau}\big)} \right)$$

Now we define the continuous time limit multiplying the discrete time t by a positive constant ξ and subsequently letting $\xi \to 0$. This way we define a continuous time $\tilde{t}$ such that as $\xi \to 0$ and $t \to \infty$ then $t \cdot \xi = \tilde{t}$. Let's build up the time derivative of $\pi_i^{\tilde{t}}$:

$$\lim_{\xi\to 0} \frac{\Delta \pi_i^{t\xi}(a_i)}{\xi} = \lim_{\xi\to 0} \frac{\pi_i^{t\xi}(a_i)}{\sum_{b\in A_i} \pi_i^{t\xi}(a_i) exp\big(\frac{\Delta q_i^{t\xi}(b)}{\tau}\big)} \cdot \left(\frac{exp\big(\frac{\Delta q_i^{t\xi}(a_i)}{\tau}\big)}{\xi} - \frac{\sum_{b\in A_i} \pi_i^{t\xi}(a_i) exp\big(\frac{\Delta q_i^{t\xi}(b)}{\tau}\big)}{\xi} \right)$$

The first term gives simply $\pi_i^{\tilde{t}}(a_i)$ as the term $\Delta q_i^{t\xi}(b)$ of the exponential clearly goes to 0 and the sum of π_i gives 1. The second term instead, if one collects the numerators, is a $\frac{0}{0}$ so we have to use the firs order approximation (consider only the numerator of the second term):

$$\lim_{\xi\to 0} exp\big(\frac{\Delta q_i^{t\xi}(a_i)}{\tau}\big) - \sum_{b\in A_i} \pi_i^{t\xi}(b) exp\big(\frac{\Delta q_i^{t\xi}(b)}{\tau}\big) =$$

$$= \lim_{\xi\to 0} 1 + \frac{\Delta q_i^{t\xi}(a_i)}{\tau} - \sum_{b\in A_i} \pi_i^{t\xi}(b) - \sum_{b\in A_i} \pi_i^{t\xi}(b)\frac{\Delta q_i^{t\xi}(b)}{\tau}$$

$$= \lim_{\xi\to 0} \frac{\Delta q_i^{t\xi}(a_i)}{\tau} - \sum_{b\in A_i} \pi_i^{t\xi}(b)\frac{\Delta q_i^{t\xi}(b)}{\tau}$$

so by noticing that:

$$\lim_{\xi\to 0} \frac{\Delta q_i^{t\xi}(a_i)}{\xi} = \frac{d}{d\tilde{t}} q_i^{\tilde{t}}(a_i)$$

we can collect everything to get the first result:

$$\frac{d}{d\tilde{t}}\pi_i^{\tilde{t}}(a_i) = \frac{1}{\tau}\cdot \pi_i^{\tilde{t}}(a_i)\cdot\left(\frac{d}{d\tilde{t}} q_i^{\tilde{t}}(a_i) - \sum_{b\in a_i} \pi_i^{\tilde{t}}(b)\cdot \frac{d}{d\tilde{t}} q_i^{\tilde{t}}(b)\right) \tag{2.5}$$

Let's do the same for q_i: The update rule already involves a relation between subsequent times so we can directly apply the continuous time limit to it (α is the scale of the time evolution so it has to be multiplied by ξ too):

$$\lim_{\xi\to 0} q_i^{(t+1)\xi}(a_i) = \lim_{\xi\to 0} q_i^{t\xi}(a_i) + \alpha\xi\left(u_i(a_i, a_{-i}^{t\xi}) + \xi\big(\sum_{b\in A_i} \pi_i^{t\xi}(b) q_i^{t\xi}(b)\big) - q_i^{t\xi}(a_i)\right)$$

$$\lim_{\xi\to 0} \frac{\Delta q_i^{t\xi}(a_i)}{\xi} = \lim_{\xi\to 0} \alpha\left(u_i(a_i, a_{-i}^{t\xi}) + \gamma\big(\sum_{b\in A_i} \pi_i^{t\xi}(b) q_i^{t\xi}(b)\big) - q_i^{t\xi}(a_i)\right)$$

$$\frac{d}{d\tilde{t}} q_i^{\tilde{t}}(a_i) = \alpha\left(u_i(a_i, a_{-i}^{\tilde{t}}) + \gamma\big(\sum_{b\in A_i} \pi_i^{\tilde{t}}(b) q_i^{\tilde{t}}(b)\big) - q_i^{\tilde{t}}(a_i)\right)$$

We can plug this last result in (2.5):

$$\frac{d}{d\tilde{t}}\pi_i^{\tilde{t}}(a_i) = \frac{1}{\tau}\cdot \pi_i^{\tilde{t}}(a_i)\cdot\alpha\Bigg(u_i(a_i, a_{-i}^{\tilde{t}}) + \gamma\big(\sum_{b\in A_i} \pi_i^{\tilde{t}}(b) q_i^{\tilde{t}}(b)\big) - q_i^{\tilde{t}}(a_i) -$$

$$- \sum_{b\in a_i} \pi_i^{\tilde{t}}(b)\cdot\big(u_i(b, a_{-i}^{\tilde{t}}) + \gamma\big(\sum_{c\in A_i} \pi_i^{\tilde{t}}(c) q_i^{\tilde{t}}(c)\big) - q_i^{\tilde{t}}(b)\big)\Bigg)$$

The average of q_i does not depend on a_i or b and so the corresponding terms cancel out because as already observed before, the sum of π_i gives 1. Moreover notice that $\sum_{b\in a_i} \pi_i^{\tilde{t}}(b)\cdot u_i(b, a_{-i}^{\tilde{t}}) = u_i(\pi_i^t, a_{-i}^t)$

$$\frac{d}{d\tilde{t}}\pi_i^{\tilde{t}}(a_i) = \frac{1}{\tau}\cdot \pi_i^{\tilde{t}}(a_i)\cdot\alpha\left(u_i(a_i, a_{-i}^{\tilde{t}}) - u_i(\pi_i^t, a_{-i}^t) - q_i^{\tilde{t}}(a_i) + \sum_{b\in A_i} \pi_i^{\tilde{t}}(b) q_i^{\tilde{t}}(b)\right)$$

Now, exploiting again the fact that $\sum_{b\in a_i}\pi_i^{\tilde{t}}(b)=1$ we can write

$$\begin{aligned}\frac{d}{d\tilde{t}}\pi_i^{\tilde{t}}(a_i) &= \frac{1}{\tau}\cdot\pi_i^{\tilde{t}}(a_i)\cdot\alpha\left(u_i(a_i,a_{-i}^{\tilde{t}})-u_i(\pi_i^t,a_{-i}^t)-\sum_{b\in a_i}\pi_i^{\tilde{t}}(b)q_i^{\tilde{t}}(a_i)+\sum_{b\in A_i}\pi_i^{\tilde{t}}(b)q_i^{\tilde{t}}(b)\right)\\ &= \frac{1}{\tau}\cdot\pi_i^{\tilde{t}}(a_i)\cdot\alpha\left(u_i(a_i,a_{-i}^{\tilde{t}})-u_i(\pi_i^t,a_{-i}^t)+\sum_{b\in a_i}\pi_i^{\tilde{t}}(b)\big(q_i^{\tilde{t}}(b)-q_i^{\tilde{t}}(a_i)\big)\right)\end{aligned}$$

Finally, noticing that $\dfrac{\pi_i^{\tilde{t}}(b)}{\pi_i^{\tilde{t}}(a_i)}=\dfrac{exp\left(\frac{q_i^{\tilde{t}}(b)}{\tau}\right)}{exp\left(\frac{q_i^{\tilde{t}}(a_i)}{\tau}\right)}$ then we have that

$$\frac{1}{\tau}\big(q_i^{\tilde{t}}(b)-q_i^{\tilde{t}}(a_i)\big)=log\frac{\pi_i^{\tilde{t}}(b)}{\pi_i^{\tilde{t}}(a_i)}$$

Now we can put everything together and substitute $\tilde{t}\to t$ since there is no longer a need to distinguish discrete and continuous time with different notations. Just recall that t is now a continuous parameter.

$$\frac{d}{dt}\pi_i^t(a_i)=\alpha\left[\frac{1}{\tau}\cdot\pi_i^t(a_i)\cdot\big(u(a_i,a_{-i}^t)-u_i(\pi_i^t,a_{-i}^t)\big)+\pi_i^t(a_i)\cdot\sum_{b\in A_i}\pi_i^t(b)\cdot log\frac{\pi_i^t(b)}{\pi_i^t(a_i)}\right]$$

Then to obtain equation (2.4) one has to average over π_{-i}^t

We can see that $\pi_i^t(a_i) = 0$ is a stationary state, since it corresponds to a vanishing time variation:

$$\frac{d}{dt}\pi_i^t(a_i) = \alpha\left[\frac{1}{\tau}\cdot 0\cdot\left(u(a_i,\pi_{-i}^t) - u_i(\pi_i^t,\pi_{-i}^t)\right) + 0\cdot\sum_{b\in A_i}\pi_i^t(b)\cdot log\frac{\pi_i^t(b)}{\pi_i^t(a_i)}\right] = 0$$

Similarly, also $\pi_i^t(a_i) = 1$ is a stationary state, since in that case $u(\pi_i^t,\pi_{-i}^t) = u(a_i,\pi_{-i}^t)$ and $\pi(b_i) = 0\ \forall b_i \neq a_i$, so again:

$$\frac{d}{dt}\pi_i^t(a_i) = \alpha\left[\frac{1}{\tau}\cdot 1\cdot(0) + 1\left(\cdot\sum_{b\neq a_i\in A_i} 0\cdot log(0) + 1\cdot log(1)\right)\right] = 0$$

What's more we do know that a symmetric mixed strategy Equilibrium that is an ESS is a stable fixed point for π^t, as long as τ is is sufficiently small. Already from (2.3) one can see that in the limit of large temperature $\tau >> |q^t|$, π tends to uniform distribution. So in a first approximation we can conclude that this kind of learning is characterized by stable strategies which are the ESS strategies of the underlying game, absorbing extreme states that are difficult to reach when the temperature is sufficiently high and a competing effect between the replicator dynamics that tends to fix in its fixed points and the entropy term that tends to lead to uniform distributions.

Chapter 3

Stochastic games

[11] [6] Stochastic games (or Markov games) are a generalizations of repeated games, in which there is not only one single stage game, but instead at each stage possibly a different game is played. To be more precise, the stochastic game consists in a discrete time Markov process in which each state corresponds to a different stage game. Differently from a simple Markov process, the transition probabilities depend also on the strategy profile played at the current state/game. In the following, it will be understood that at a state corresponds a game. Formally a (infinite horizon) Stochastic Game is defined by:

- A finite set of players, with cardinality k, which can be represented by the interval $I = [1, k] \subset \mathbb{Z}$
- A set of states $\mathcal{S}$
- A collection of action sets $A_1, ..., A_k$ which are the sets of pure strategies available for each player in each state. As usual A will denote $\times_{j \in I} A_j$, the set of strategy profiles.
- $\forall$ player i a utility function u_i: $A \times \mathcal{S} \longrightarrow \mathbb{R}$
- A transition function T: $A \times \mathcal{S} \longrightarrow \Delta(\mathcal{S})$ which $\forall a \in A$ defines a transition matrix $P^a_{s,s'}$ with $s, s' \in \mathcal{S}$

This defines the stochastic game $\mathcal{SG} = \big(\mathcal{I}, \mathcal{S}, T, A, u\big)$

Remark 6. *This is not the most general formulation of Stochastic Game as in this setting the difference between two games lies only in the payoff and not in the structure. A more general formulation could take into account also state-dependent action sets:* $A^s_1, ...A^s_k$ *with* $s \in \mathcal{S}$.

Let s_t denote the state occurring at time t. At each timestep t each player i chooses an action $a_i \in A_i$ and receives a reward $r^i_t := u_i(s_t, a_i, a_{-i})$. A strategy for player i can be defined as a function associating at each timestep and each current state an action or, more generally a mixed strategy: π_i: $\mathbb{Z}^+ \times \mathcal{S} \longrightarrow \Delta(A_i)$. Given a strategy profile π, at each possible sequence of states $\{s_t\}^\infty_{t=0}$ a probability measure $\mathcal{P}\big(\{s_t\}^\infty_{t=0}\big)$ can be associated. Given a discount factor $\gamma \in [0, 1]$, the purpose of each player i is to maximize the expected discounted payoff:

$$U_i\big(s_0, \pi\big) := \mathbb{E}_{\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t r^i_t\right]$$

A special class of strategies, particularly relevant in this infinite horizon formulation of Stochastic Game, is the time independent strategies (or steady strategies) class, for which the action selected at each stage depends only on the current state: π_i: $\mathcal{S} \longrightarrow \Delta(A_i)$.

3.1 Reinforcement learning in stochastic games

Several $Q-$learning approaches, suited for the diverse classes of stochastic games can be found in the literature [6] [9] [7]. The structure of the learning is always the one presented in subsection 2.2.3 with the state dependence incorporated (in this sense it resembles more the MDP $Q-$learning). As we will see with some examples, in order to ensure the convergence in this highly non-stationary environment, the different models exhibit more complicated features and stronger convergence hypothesis. Even though there is no unique general formulation, since each formulation is suited for the specific class of games considered, some common features are often found:

- Decreasing learning rates: the learning rates are often decreasing functions of the iteration time or of the state-action counters. The way such functions decrease is also a matter of the convergence hypothesis.

- Double time-scale induced by two different learning rates: a double time-scale evolution is often present, induced by two learning rates that decrease at different speeds. One of the learning rates corresponds to the evolution of the Q-function while the other corresponds to the evolution of the strategies or of a value-function as in the case of 0-sum games.

3.1.1 Two-player 0-sum stochastic games

Definition 3.1.1. *A stochastic game is 0-sum iff each stage game is 0-sum*

The concept of the value of a game can be extended to this framework. In a zero-sum stage game, it is possible to arbitrarily choose a player and use only its payoff matrix, privileging its perspective. A two-player zero-sum stochastic game is then uniquely specified by a single payoff function: $u^s : a_1 \times a_2 \longrightarrow \mathbb{R}$ will denote the payoff matrix of Player 1 at stage game s.
If M is the payoff matrix of a 0-sum game: $Val(M)$, $\mathcal{X}(M)$, $\mathcal{Y}(M)$ will denote respectively: the value of the game, the set of optimal strategies for the first and the second player.
Because having different stage games implies having a starting state s_0, one may expect to have a different value V_{s_0} for each $s_0 \in \mathcal{S}$. So in the following, value of the game V will refer to the vector whose components are all the $\left|\mathcal{S}\right|$ (possibly) different values.

Using this notation, it is now possible to introduce a definition of the solution to the stochastic game $\mathcal{SG} = \big(\mathcal{I}, \mathcal{S}, T, A, u\big)$, as proposed by Shapley in [11]:

Definition 3.1.2. *The value of a 0-sum stochastic game is defined to be the unique solution of the system:*

$$V_s = Val\big[u_s(a)\big) + \sum_{s' \in \mathcal{S}} P^a_{s,s'} \cdot V_{s'}\big]$$

Proposition 10. *Given a two-player 0-sum stochastic game:*

- $\exists! \; V$ *value of the game.*
- $\exists \; \pi^*$ *steady solution of the game.*

Proof. [11] □

3.1.2 Q-learning in two-player 0-sum stochastic games

The main idea of Q-learning can be exploited also in this framework to estimate optimal strategy profiles of the stochastic game. Extending the Q-learning concept to stochastic games involves addressing multi-agent interactions and trying to exploit the peculiar structure of the game considered. Transitioning from single to multi-agent introduces complexities, necessitating assumptions and simplifications based on the specific context. In case of 0-sum games a possible choice is, for each player, to assume that its opponent employs a minmax strategy. This simplification allows agents to plan while considering the worst-case scenario, easing the computational burden in multi-agent environments. This approach, employed in [6], essentially consists in replacing the *max* operator in algorithm 1 with the *maxmin*:

Algorithm 2 Multiagent Q-learning (player i)

Initialize

- Value function $V^i(s)$, $\forall s \in \mathcal{S}$
- Action-value function $Q^i(s, a, o)$, $\forall s \in \mathcal{S}$, $\forall a \in A_i$ and $\forall o \in A_{-i}$
- Policy $\pi_i(s, a) = \dfrac{1}{|A_i|}$, $\forall s \in \mathcal{S}$, $\forall a \in A_i$
- Initial learning rate $\alpha = 1$
- Deviation probability $explore \in [0, 1]$
- Discount factor $\gamma \in [0, 1]$

Repeat

- With probability *explore* choose uniformly an action a_t from a_i.
  Otherwise observe current state s_t and choose an action a_t with probability distribution $\pi_i(s_t, \cdot)$.
- Observe new state s_{t+1}
- Update Q^i, V^i, π and α:

$$Q^i(s_t, a_t) \longleftarrow \cdot Q^i(s_t, a_t) + \alpha \cdot \left(u^i(a_t, s_t) + \gamma V^i(s_{t+1}) - Q^i(s_t, a_t)\right)$$

$$V^i(s_{t+1}) \longleftarrow \max_{p \in \Delta(A_i)} \min_{o \in A_{-i}} \sum_{a \in A_i} p(a) \cdot Q(s_{t+1}, a, o)$$

$$\pi_i(s_{t+1}, \cdot\ ) \longleftarrow \arg\max_{p \in \Delta(A_i)} \arg\min_{o \in A_{-i}} \sum_{a \in A_i} p(a) \cdot Q(s_{t+1}, a)$$

$$\alpha \longleftarrow \alpha_t$$

Notice that this algorithm is the maxmin generalization to a multi-agent setting of algorithm 1.

3.1.3 Decentralized Q-learning in two-player 0-sum stochastic games

In [9] a different approach is proposed, which has the advantage of not necessiting the knowledge of the opponent's action set nor of the 0-sum structure of the game. To present it, some definitions and specifications are necessary. Such definitions will be "theoretical", in the sense that they will have a correspondence in the algorithm, but such correspondence will not use such definitions neither in the initialization nor in the updates.
In this framework, the value function is not defined as the solution of the game, so a different notation will be employed to avoid ambiguities (v instead of V):

Definition 3.1.3 (Value function)**.** *Given the stochastic game* $\mathcal{SG} = \left(\mathcal{I}, \mathcal{S}, T, A, \{u_i\}_{i \in \mathcal{I}}\right)$

and strategy profile π, the value function of player i is a solution of the self-consistent equation, $\forall s \in \mathcal{S}$:

$$v^i(s) := \mathbf{E}_\pi \left[u_i(s,a) + \gamma \sum_{s' \in \mathcal{S}} v^i(s') \cdot P^a_{s,s'} \right]$$

Remark 7. *The average is taken only on the probability measure of the mixed strategy profile, but the term averaged is in turn an explicit average with respect to the probability measure of the stochastic process, so it is equivalent to the averages presented in the previous sections.*

Definition 3.1.4 (Q-function and local Q-function)**.** *Given the stochastic game $\mathcal{SG} = \big(\mathcal{I}, \mathcal{S}, T, A, \{u_i\}_{i \in \mathcal{I}}\big)$, the Q-function and local Q-function of player i are respectively, $\forall$ $a \in A$ and $\forall$ $s \in S$*

$$Q^i(s,a) := u_i(s,a) + \gamma \sum_{s' \in \mathcal{S}} v^i(s') \cdot P^a_{s,s'}$$

$$q^i(s,a_i) := \mathbb{E}_{\pi_{-i}} \Big[Q^i(s,a) \Big]$$

The key element is the local Q-function, which is independent on the opponent's actions. Indeed, as anticipated, the strength of the algorithm lies in the fact that each player only needs to know its own action set, and observe the current state and the payoff obtained. Let's list other parameters and peculiarities of the model:

- All players are assumed to know the bounds of their utility functions: $\forall$ player i, $\big|\big|u_i\big|\big| \leq R \in \mathbb{R}^+$. Therefore they will ititialize q^i and v^i so that $\big|\big|q^i\big|\big|_\infty \leq \frac{R}{1-\gamma}$ and $\big|\big|v^i\big|\big|_\infty \leq \frac{R}{1-\gamma}$

- The players, as already stated, observe the current state and record them, with perfect recall. In particular, they record the number of visits of each state s denoting it with $\#s$.

- In each state s, each player i chooses an action according to a smooted best response function:

$$\mathcal{BR}^i\big(q^i(s,\,\cdot), \tau_{\#s}\big) := \underset{p \in \Delta(A_i)}{\arg\max} \sum_{a_i \in A_i} \Big(p(a_i) \cdot q^i(s,a_i) + \tau_{\#s} \cdot \nu^i_s\big(p(a_i)\big) \Big)$$

  and

$$\pi_i := \mathcal{BR}^i\big(q^i(s,\,\cdot), \tau_{\#s}\big)$$

  where $\tau_{\#s} > 0$ is a decreasing function of s satisfying some assumptions (specified below) in order to ensure convergence to the value of the game. It plays the role of a temperature parameter which has the purpose of tuning the probability of deviating from observations, in a sense that will be clarified by the specific form of ν^i_s.
  ν^i_s indeed, is a smooth function of the probability distribution p. If it takes the form of an entropy:

$$\nu_s^i\big(p(a_i)\big) := -\sum_{a_i \in A_i} p(a_i) \cdot logp(a_i)$$

then

$$\pi_i(s, a_i) = \frac{exp\left(\dfrac{q^i(s, a_i)}{\tau_{\#s}}\right)}{\displaystyle\sum_{a' \in A_i} exp\left(\dfrac{q^i(s, a')}{\tau_{\#s}}\right)}$$

and the interpretation of $\tau_{\#s}$ as a temperature is straightforward.

- For each player i two different learning rates: $\alpha_i(\#s)$ for the learning of q^i and $\beta_i(\#s)$ for the learning of v^i. This learning rates together with the temperature parameter must satisfy some assumptions (specified below) in order to ensure convergence to the value of the game.

Assumption 1-i The sequences α_c and β_c are non-increasing and satisfy $\sum_c \alpha_c = \infty$, $\sum_c \beta_c = \infty$, and $\lim_{c\to\infty} \alpha_c = \lim_{c\to\infty} \beta_c = 0$.

Assumption 1-ii Given any $M \in (0, 1)$, there exists a non-decreasing polynomial function $C(\cdot)$ (which may depend on M) such that for any $\lambda \in (0, 1)$, if

$$\{\ell \in \mathbb{Z}^+ | \ell \leq c \text{ and } \frac{\beta_\ell}{\alpha_c} > \lambda\} \neq \emptyset \Longrightarrow$$
$$\max\{\ell \in \mathbb{Z}^+ | \ell \leq c \text{ and } \frac{\beta \ell}{\alpha_c} > \lambda\} \leq Mc, \quad \forall c \geq C(\lambda^- 1).$$

Assumption 2-i Given any pair of states (s, s'), there exists at least one sequence of actions such that s' is reachable from s with some positive probability within a finite number, n, of stages.

Assumption 2-ii The sequence τ_c is non-increasing and satisfies $\lim_{c\to\infty}(\tau_{c+1} - \tau_c)/\alpha_c = 0$ and $\lim_{c\to 1} \tau_c = \delta$ for some $\delta > 0$. What's more $\sum_c \alpha_c^2 < \infty$.

Assumption 2'-i Given any pair of states (s, s') and any infinite sequence of actions, s' is reachable from s with some positive probability within a finite number, n, of stages.

Assumption 2'-ii The sequence τ_c is non-increasing and satisfies $\lim_{c\to\infty}(\tau_{c+1} - \tau_c)/\alpha_c = 0$ and $\lim_{c\to\infty} \tau_c = 0$. α_c satisfies $\sum_{c=1} \alpha_c^{2-\rho} < \infty$, for some $\rho \in (0, 1)$. There exists $C, C' \in (0, \infty)$ such that $\alpha_c^\rho \exp(4D/\tau_c) \leq C'$ for all $c \geq C$.

Under assumptions 1 and 2 the convergence to a ϵ-Nash equilibrium is ensured while under assumptions 1 and 2′ a convergence to a Nash equilibrium is instead ensured. Note that also this one is model-free.

Remark 8. *The learning dynamics involves only the the local Q-function q^i, which is a function defined on $\mathcal{S} \times a_i$. In this sense, from the perspective of player i it is a single player Q learning with a softmax update of the strategy.*

Algorithm 3 Decentralized Q-learning (player i)

Initialize

- $q^i(s, a_i)\ \forall\ s \in \mathcal{S}$ and $\forall\ a_i \in A_i$
- $\pi_i(s, a_i)\ \forall\ s \in \mathcal{S}$ and $\forall\ a_i \in A_i$
- $v^i_\pi(s)\ \forall\ s \in \mathcal{S}$
- Smooth function $\nu^i_s(a_i)\ \forall\ s \in \mathcal{S}$ and $\forall\ a_i \in A_i$
- Fix the discount factor γ

Repeat

- Observe and record current state s_t
- Update learning rates, local Q-function and strategy:

$$\alpha_i \longleftarrow \min\left\{1, \frac{\alpha_i(\#s_{t-1})}{\pi_i(s_{t-1}, a_i^{t-1})}\right\}$$

$$q_i(s^{t-1}, a_i^{t-1}) \longleftarrow q_i(s^{t-1}, a_i^{t-1}) + \alpha_i \cdot \left(u_i(s^{t-1}, a^{t-1}) + \gamma v(s^t) - q_i(s^{t-1}, a_i^{t-1})\right)$$

$$\pi_i(s_t, \cdot) \longleftarrow \underset{p \in \Delta(A_i)}{\arg\max} \sum_{a_i \in A_i} \left(p(a_i) \cdot q_i(s^t, a_i) + \tau_{\#s} \cdot \nu^i_s\big(p(a_i)\big)\right)$$

- Choose and record a_i^t according to $\pi_i(s^t, \cdot)$
- Observe and record current payoff $u_i(s^t, a^t)$
- Update value function:

$$v_i(s^t) \longleftarrow v_i(s^t) + \beta_i(\#s^t) \cdot \left(\sum_{a_i \in A_i} \pi_i(s^t, a_i) \cdot q_i(s^t, a_i) - v_i(s^t)\right)$$

This is the first algorithm encountered in which a time-scales separation appears: β decreases faster than α and this induces a dynamics in which v is almost stationary from the point of view of q. The idea is that v is in some sense a long term belief on the quality of a state and it has to be more difficult to modify it than the q, which has to be intended as a signal reflecting the natural fluctuations and the contingency of the outcomes received.

3.1.4 Potential stochastic games

Definition 3.1.5 (Stochastic potential game)**.** *A stochastic game* $\mathcal{SG} = \big(\mathcal{I}, \mathcal{S}, T, A, \{u_i\}_{i \in \mathcal{I}}\big)$ *is potential if there exists a state-dependent potential function* $\Phi : S \times \big(\times_{i \in I} \Delta(A_i)\big) \to \mathbb{R}$ *such that for every (initial state)* $s \in \mathcal{S}$*,*

$$\Phi(s, \pi'^i, \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = U_i(s, \pi'^i, \pi_{-i}) - U_i(s, \pi_i, \pi_{-i}) \tag{3.1}$$

for any player i, any $\pi_i, \pi'^i \in \Delta(A_i)$, and any $\Phi_{-i} \in \times_{j\in I\setminus i}\Delta(A_j)$

In other words it is potential if the associated game $\mathcal{G}'\big(\mathcal{I}, \times_{i\in I}\Delta(A_i), \{V_i\}_{i\in\mathcal{I}}\big)$ is potential. [7] Unlike conventional stochastic games, where the focus often lies on finding equilibrium strategies, potential stochastic games offer a broader perspective by incorporating notions of stability and convergence associated with potential functions. These games serve as a valuable tool for analyzing strategic behavior in dynamic systems with complex interdependencies and evolving objectives. In [7] a model-free algorithm, similar to the decentralized Q-learning of the previous section, is presented. Once more, the learning involves directly a local Q-function q^i and the update of the strategy π_i is a softmax with temperature parameter and once more there are two different decreasing time-scales $\alpha_i(c)$, $\beta_i(c)$, where as always i labels different players. This time, the two time-scales decouple the learning of the various quantities involved in a different way: the fast time-scale (α_i) governs the evolution of q^i while v^i is updated directly from q^i, evolving consequently at the same speed. The slow time-scale governs the evolution of the strategy π_i. While $\beta_i(c)$ is again a function of the occurrencies of a state ($c = \#s$), $\alpha_i(c)$ is function of the occurrence of a couple action-state ($c = \#(s, a_i)$). This time the temperature parameter τ is a fixed parameter that tunes the convergence to an ϵ-Nash equilibrium in the following sense: the algorithm converge to an ϵ-Nash equilibrium π^*_τ and $\lim_{\tau\to 0}\pi^*_\tau = \pi^*$, Nash equilibrium of the game. It is important to underline that generally in a potential (stochastic) game, there may be more than a single equilibrium: so in this case the dynamics will select one of such equilibria.
As in the previous case, some assumptions on the time-scales are needed in order to ensure the convergence:

Assumption $\alpha(c)$, $\beta(c)$ are non-increasing sequences and satisfy (the label of player i is omitted as these conditions must hold for all the players):

1. $\sum_c \alpha(c) = \infty$, $\sum_c \beta(c) = \infty$, and $\lim_{c\to\infty}\alpha(c) = \lim_{c\to\infty}\beta(c) = 0$

2. For some p, $p' \geq 2$ $\sum_c \alpha(c)^{1+p/2} < \infty$ and $\sum_c \beta(c)^{1+p'/2} < \infty$

3. For any $x \in (0,1)$, $\sup_c \alpha(\lfloor xc\rfloor)/\alpha(c) < \infty$, $\sup_c \beta(\lfloor xc\rfloor)/\alpha(n) < \infty$

4. $\lim_{c\to\infty}\beta(c)/\alpha(c) = 0$.

One example of stepsizes that satisfy the assumption is: $\alpha_c = (c+1)^{-h}$ and $\beta_c = (c+1)^{-g}$ with $0 < h < g < 1$. Let's define the algorithm:

Algorithm 4 Decentralized Q-learning for potential games (player i)

Initialize

- $q_i(s, a_i)\ \forall\ s \in \mathcal{S}$ and $\forall\ a_i \in A_i$
- $\pi_i(s, a_i)\ \forall\ s \in \mathcal{S}$ and $\forall\ a_i \in A_i$
- Entropy-like function $\nu(p(\cdot)) := \sum_{a' \in A_i} -p(a')logp(a')$
- Fix the discount factor γ and temperature factor τ
- Initialize the state and action-state counters $n_0(s) = 0$ and $n_0^i(s, a_i) = 0$, $\forall$ player i , $\forall s \in \mathcal{S}$, $\forall a_i \in A_i$
- At time $t = 0$ observe initial state s^0, choose action a_i^0 according to $\pi_i\ \forall i$. Observe $u_i(s^0, a^0)$.

Repeat $(t \geq 1)$

- Observe and record current state s^t
- $n(s^{t-1})$ += 1 and $n_i(s^{t-1}, a_i^{t-1})$ += 1
- Update local Q-function and strategy:

$$q_i(s^{t-1}, a_i^{t-1}) \longleftarrow q_i(s^{t-1}, a_i^{t-1}) + \alpha_i\big(n_i(s^{t-1}, a_i^{t-1})\big) \cdot \Bigg(u_i(s^{t-1}, a^{t-1}) + \tau \cdot \nu(\pi_i(s^{t-1}, \cdot)) + \gamma \sum_{a' \in A_i} \pi_i(s^t, a') \cdot q_i(s^t, a') - q_i(s^{t-1}, a_i^{t-1})\Bigg)$$

$$\pi_i(s^{t-1}, a_i^{t-1})$$

$$\longleftarrow \pi_i(s^{t-1}, a_i^{t-1}) + \beta_i\big(n(s^{t-1})\big) \cdot \left(\frac{exp\left(\frac{q_i(s^{t-1}, a_i^{t-1})}{\tau}\right)}{\sum_{a' \in A_i} exp\left(\frac{q_i(s^{t-1}, a')}{\tau}\right)} - \pi_i(s^{t-1}, a_i^{t-1})\right)$$

- Choose and record a_i^t according to $\pi_i(s^t, \cdot)$
- Observe and record current payoff $u_i(s^t, a^t)$

Also in this case there is a time-scale separation: this time the two time-scales affect the learning of q and of the strategy π: in particular again q evolves faster while π instead plays the role of a fixed belief of the agent, needing more than a simple fluctuation to be modified.

In the next chapter we will see that in a stateless setting tuning this time-scales difference (also reverting their $>$ relation) will lead to different equilibrium selection.

Chapter 4

Learning the El Farol bar problem repeated game

It is quite evident that when the space of states of a stochastic game $\mathcal{SG}$ is a singleton, it is equivalent to a repeated game whose stage game is the only stage game of $\mathcal{SG}$ corresponding to its unique state. This naive consideration brings to a maybe still naive but non-negligible conclusion: every result in the framework of stochastic games can be exploited and must hold also for repeated games. Stochastic games represent usually a more complex class of problems compared to repeated ones, but still they brought to the development of different computational approaches in order to face the new challenges they offered. Consider potential games: a best response dynamic based on the potential function is a standard approach that allows to find Nash equilibria. Conversely, algorithm 4 uses a dynamics which is more inspired to Q-learning, never using explicitly the form of the potential function. What kind of results can be found by applying this approach to the potential repeated game? What kind of effect do the two time-scales induce on the repeated game?

4.1 El Farol bar problem

El Farol bar problem represents a typical problem in the context of decision-making, where patrons must decide whether to attend a bar, taking into account its limited capacity. Formulated originally in [1] its main idea is the following: this Farol bar is a beautiful bar offering very likeable shows once a week and every week each member of a population has to choose whether to go to the bar or stay home, considering that there is no room for everyone and that it is preferable to stay home than to stay in a too crowded bar. This scenario exemplifies coordination dilemmas prevalent in complex systems, where individual actions impact collective outcomes. By exploring the dynamics of this problem it is possible to get insights into emergent behavior, strategic interactions, and the role of expectations in social settings. El Farol Bar Problem applications of course go further its simple premise, offering valuable lessons applicable to economics, sociology, and artificial intelligence giving a simple but still not trivial model for phenomena like market dynamics, traffic congestion, and social network dynamics.

Here we consider one of its possible formulations: the bar has limited seats M and the population is composed by N elements, both fixed. Let $\mathcal{A}$ be the number of people choosing to attend the bar, $x > 0$ be a real number and G and H respectively denote the

actions *Go to the bar* and *stay Home*. The utility function takes the form:

$$\begin{aligned} u_i(a_i = G, a_{-i}) &= x \quad \text{if } \mathcal{A} \leq M \\ u_i(a_i = G, a_{-i}) &= -1 \quad \text{if } \mathcal{A} > M \\ u_i(a_i = H, a_{-i}) &= 0 \quad \forall \mathcal{A} \end{aligned} \tag{4.1}$$

Some considerations:

- It is a congestion game, so it is also potential.
- There are of $\binom{N}{M}$ pure strategy Nash equilibria, that are the ones in which exactly M agents choose G and the rest H. If one of the "G" agents deviates from this strategy its reward decreases from x to 0. On the other hand, if one of the "H" agents deviates its reward decreases from 0 to -1 because in this case $A = M + 1$.
- There is a unique mixed strategies symmetric Nash equilibrium (determined below).
- There are countably many Asymmetric Nash equilibria in which part of the players choose a pure strategy and part play a mixed strategy.

Symmetric equilibrium One can find the symmetric mixed strategy equilibrium using the *method of equating payoffs*: suppose such equilibrium exists and all players (but i) play it. Let's denote by $\pi|p$ a symmetric mixed strategy such that $\pi_i|_p(G) = p \; \forall i$ and by $\mathcal{A}_{-i}$ the number of players choosing GO other than player i:

$$\begin{aligned} &u_i(H, \pi_{-i}|_p) = 0 \\ &u_i(G, \pi|_p^{-i}) = xPr(\mathcal{A}_{-i} \leq M - 1 \mid p) - Pr(\mathcal{A}_{-i} > M - 1 \mid p) \end{aligned}$$

But $Pr(\mathcal{A}_{-i} \geq M) = 1 - Pr(\mathcal{A}_{-i} \leq M - 1)$ so

$$u_i(G, \pi_p^{-i}) = (x + 1)Pr(\mathcal{A}_{-i} \leq M - 1 \mid p) - 1$$

and by equating $u_i(G, \pi_p^{-i}) = u_i(H, \pi_p^{-i})$, one finds that at the equilibrium it must hold

$$Pr(\mathcal{A}_{-i} \leq M - 1 \mid p^*) = \frac{1}{1 + x}$$

where $Pr(\mathcal{A}_{-i} \leq M - 1 \mid p) = \sum_{\ell=0}^{M-1} \binom{N-1}{\ell} p^\ell (1-p)^{N-1-\ell}$

The value p^* exists as long as $x \geq 0$ indeed, denoting for simplicity $Pr(\mathcal{A}_{-i} \leq M - 1 \mid p) := f(p)$:

$$f(0) = 1 \text{ and } f(1) = 0$$

$$\frac{d}{dp} f = \sum_{\ell=1}^{M-1} \binom{N-1}{\ell} \ell(\ell + 1 - N) p^{\ell-1} (1-p)^{N-2-\ell}$$

each term in the sum is ≥ 0 provided that $p \in [0, 1]$ (and is zero only at the extremes) excluding $(\ell + 1 - N)$ which is ≤ 0 (it is zero only if $\ell = N - 1$ which is a trivial case, but still the argument works). So $f(p)$ is a non increasing continuous function that spans from 1 to 0, implying that p^* exists.

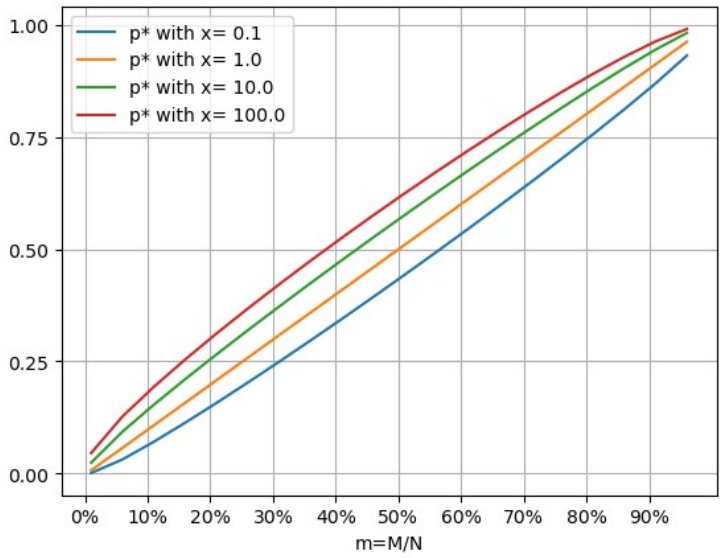


Figure 4.1: Numerical estimation of p^* for different values of the ratio thershold/players

Asymmetric equilibrium From the computation just made it is possible to extend the same reasoning to the asymmetric equilibrium. Suppose part of the players, say n, choose a pure strategy and in particular n_G choose G, while n_H choose H. Suppose the remaining part of the players play the same mixed strategy, and consider one of them: since we do know that n players will play deterministically, the situation from the point of view of players playing mixed strategies is equivalent to the case in which all players play the same mixed strategy but having a different threshold $\tilde{M} := M - n_G$ and a reduced population $\tilde{N} := N - n$. So a value $\tilde{p}^*$ can be found as before by imposing (here $\tilde{\mathcal{A}}_{-i}$ is the number of players other than i choosing GO among the population playing the mixed strategy):

$$Pr(\tilde{\mathcal{A}}_{-i} \leq \tilde{M} - 1 \mid \tilde{p}) = \sum_{\ell=0}^{M-1} \binom{\tilde{N}-1}{\ell} p^{\ell}(1-p)^{\tilde{N}-1-\ell} = \frac{1}{1+x}$$

which is the equilibrium mixed strategy of the players playing a mixed strategy. One should still show that the players choosing a pure strategy have no convenience in deviating: for whatever player coming from the pure strategy population, the stochasticity in the outcome depends exlusively on the stochasticity induced by the mixed strategy population. Thus, for any player from the pure strategy population:

$$\begin{aligned} &u_i(H, \pi_{-i}|_{\tilde{p}}) = 0 && \text{as before} \\ &u_i(G, \pi|_{\tilde{p}}^{-i}) = (1+x)Pr(\tilde{\mathcal{A}} \leq M - 1 \mid p) - 1 && \text{notice } \tilde{\mathcal{A}} \text{ instead of } \tilde{\mathcal{A}}_{-i} \end{aligned}$$

But it turns out that it is always more convenient H than G as $u_i(G, \pi_i|_{\tilde{p}}) < 0$ if $Pr(\tilde{\mathcal{A}} \leq M - 1 \mid p) < \frac{1}{1+x}$. But this is always the case, because:

- $Pr(\tilde{\mathcal{A}}_{-i} \leq M - 1 \mid p) = \frac{1}{1+x}$
- $Pr(\tilde{\mathcal{A}} \leq M - 1 \mid \tilde{p}) < Pr(\tilde{\mathcal{A}}_{-i} \leq M - 1 \mid \tilde{p})$

The last inequality can be intuitively deduced from the fact that $Pr(\tilde{\mathcal{A}} \leq M-1 \mid \tilde{p})$ is the probability that among $\tilde{N}$ players, at most $\tilde{M}-1$ choose action G, while $Pr(\tilde{\mathcal{A}}_{-i} \leq M-1 \mid \tilde{p})$ is the probability that among $\tilde{N}-1$ players, at most $\tilde{M}-1$ choose action G; the latter is obviously larger. This fact results even more evident considering the expected values of these probabilities: $E[\mathcal{A}_{-i}] = (\tilde{N}-1)\tilde{p}$, while $E[\mathcal{A}] = \tilde{N}\tilde{p}$. In any case it is a general result holding for binomial distributions and a formal proof can be given.

Returning to the problem at hand, we can conclude that $u_i(H, \pi_{-i}|_{\tilde{p}}) > u_i(G, \pi_i|_{\tilde{p}})$ so the asymmetric equilibria are those with $n_G = 0$ and so those having only a sub-population playing the pure strategy H and a sub-population playing the mixed strategy $\tilde{p}$.

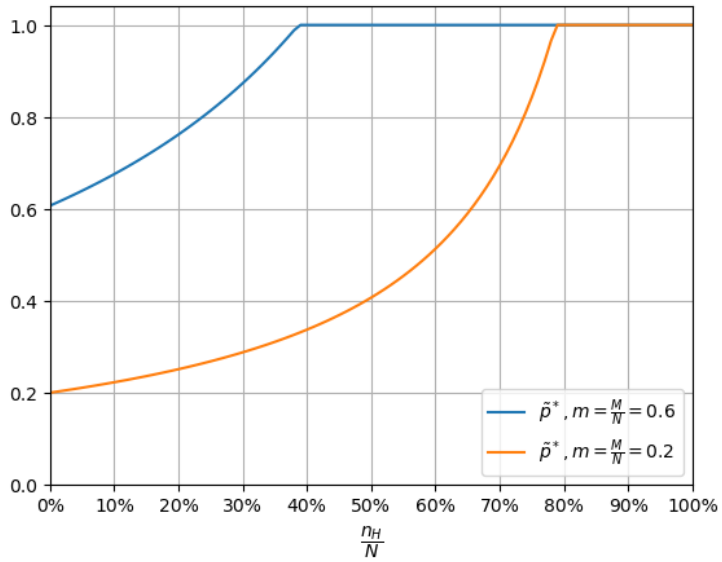


Figure 4.2: Numerical estimation of $\tilde{p}^*$ as a function of the ratio $\dfrac{n_H}{N}$ for two different values of the ratio thershold/players $m = \dfrac{M}{N}$. Clearly when $\dfrac{n_H}{N} > 1 - m$ $\tilde{p}^*$ is exactly 1 as there's probability 1 that $\mathcal{A} \leq M$

For $n_H = 0$ instead we recover the value of p^*

4.1.1 Evolutionary stability

Now we check if the symmetric Nash equilibrium found in the El Farol bar problem is an ESS.

$$u(\pi_i|_{p^*}, \pi_{-i}|_{p^*}) = (1-p^*)u(H, \pi_{-i}|_{p^*}) + p^* u(G, \pi_{-i}|_{p^*}) = p^* u(G, \pi_{-i}|_{p^*}) = p^*\big[(1+x)f(p^*) - 1\big] = 0$$

as $f(p^*) = \dfrac{1}{1+x}$ by definition of p^*

$$u(\pi_i|_q, \pi_{-i}|_{p^*}) = (1-q)u(H, \pi_{-i}|_{p^*}) + qu(G, \pi_{-i}|_{p^*}) = q\big[(1+x)f(p^*) - 1\big] = 0$$

so we are in the second case and have to check the second inequality:

$$u(\pi_i|_{p^*}, \pi_{-i}|_q) = p^*\big[(1+x)f(q) - 1\big] := p^* \cdot g(q)$$

$$u(\pi_i|_q, \pi_{-i}|_q) = q\big[(1+x)f(q) - 1\big] := q \cdot g(q)$$

So the condition is simply $p^* \cdot g(q) > q \cdot g(q)$ i.e. $(p^* - q)g(q) > 0$
We know that

- $g(p^*) = 0$
- $f(q)$ (and consequently $g(q)$) is non increasing and in $(0,1)$ it is strictly decreasing

Which implies $q < p^* \Rightarrow g(q) > 0$ and $q > p^* \Rightarrow g(q) < 0$. So the inequality is always satisfied.

4.2 Numerical results for the El Farol bar problem

Algorithm 4 is employed in the repeated El Farol bar problem (4.1). Notice that the learning rate of q should be a decreasing function of the number of times a player chooses a specific action (in the stochastic games setting also of the visits of a given state, but in the repeated game this counter coincides with the iterations). So possibly at each time step there is a different learning rate for each player and for each of the two actions. This implies that we have not total control on the learning rate of q because now it may change in different simulations and there isn't a biunivocal correspondence between learning rates and exponents h and g (we would like for example to have equal learning rates if $h = g$). Anyway, this correspondence is asymptotically valid in the cases in which both strategies are played with comparable probability at least at the beginning of the dynamics: also for this reason the choice has been that of initializing the π's as uniform distributions. Let's re-write this algorithm 4 in a dynamical system form and without the states dependence (here $n_i(a_i^t)$ is the number of times player i has played action a_i^t up to time t):

$$\pi_i^{t+1}(a_i) - \pi_i^t(a_i) = t^{-g} \cdot \left[\frac{exp\left(\frac{q_i^t(a_i)}{\tau}\right)}{\sum_{a' \in A_i} exp\left(\frac{q_i^t(a')}{\tau}\right)} - \pi_i^t(a_i) \right] \qquad \forall a_i \in \{H, G\}$$

$$q_i^{t+1}(a_i^t) - q_i^t(a_i^t) = n_i(a_i^t)^{-h} \cdot \left[u_i(a^t) + \tau \cdot \nu(\pi_i^t(\cdot)) + \gamma \sum_{a' \in A_i} \pi_i^t(a') \cdot q_i^t(a') - q_i^t(a_i^t) \right]$$

In the simulations the following quantities are computed (here T denotes the number of iterations so $t = T$ is the time at the last iteration):

- The number of players ending up in a pure strategy H and G respectively n_H and n_G (as expected the latter will turn out to be always 0)
- Sample average probability of choosing G:

$$\mu_G := \frac{1}{N} \sum_{i \in I} \pi_i^T(G)$$

- Sample average probability of choosing G among the sub-population ending up in a mixed strategy:

$$\mu_G|_{mix} := \frac{1}{n_G + n_H} \sum_{i \in I|_{mix}} \pi_i^T(G)$$

- Sample mean squared error of the probability of choosing G:

$$\sigma_G := \sqrt{\frac{1}{N}\sum_{i\in I}\left(\pi_i^T(G) - \mu_G\right)^2}$$

- Sample mean squared error of the probability of choosing G among the sub-population ending up in a mixed strategy:

$$\sigma_G|_{mix} := \sqrt{\frac{1}{n_G + n_H}\sum_{i\in I|_{mix}}\left(\pi_i^T(G) - \mu_G|_{mix}\right)^2}$$

- The symmetric Nash equilirium probability p^* and the asymmetric mixed strategy equilibrium probability of the sub-population $\tilde{p}$, both defined in section 4.1

In some of the simulations also the quantity $Err := \mu_G - p^*$ is considered. Notice that such quantity is a good measure of the error or quality of the numerical estimate only in the case in which, for some reason (see below), one is expecting the algorithm to compute the symmetric Nash equilibrium. Interesting results come out by exploring the different behaviour of the algorithm with different learning rates. The two parameters controlling the learning rates are the two exponents h and g. Conversely the discount factor is kept fixed in all the simulations (as it doesn't affect qualitatively the discussion below if not in extreme cases) $\gamma = 0.6$. Also x is kept fixed ($x = 1$), also because with values of x around 1 the results may differ in the absolute magnitude of the parameters but not qualitatively. The population is fixed at $N = 1000$ in all the simulations and the threshold is specified as a ratio m of the population:

$$m := \frac{M}{N}$$

.

The choice of the different regimes explored are to be interpreted keeping in mind that the convergence hypothesis for algorithm 4 require that $0 < h < g < 1$, which means that when the learning rates fullfill such requirements the algorithm is provably convergent to a Nash equilibrium of any suitable stochastic potential game, where suitable refers to the Markovian structure of the stochastic game: so in the case of repeated games, any potential game is a suitable potential game.

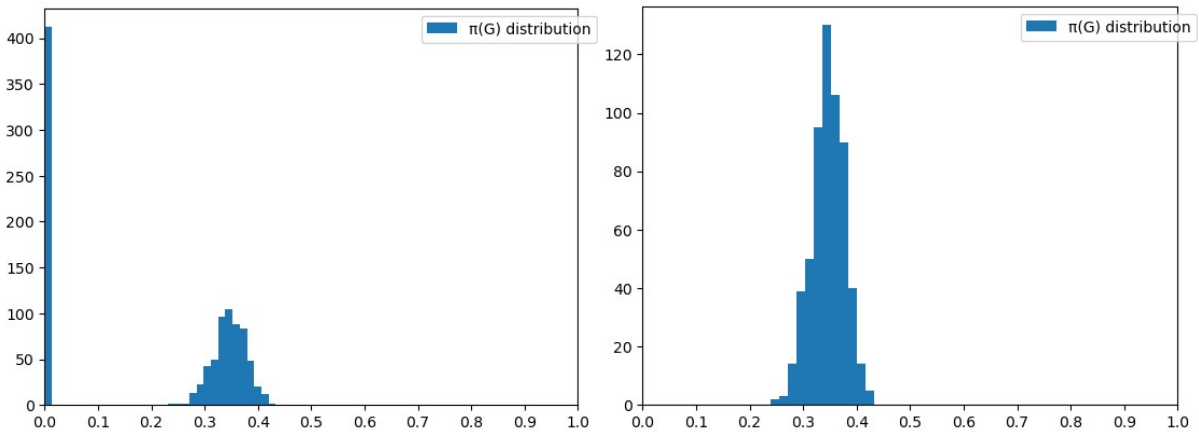
As a last remark, the algorithm is provably convergent to an ϵ-Nash equilibrium where, from an analytical point of view, $\epsilon \to 0$ as $\tau \to 0$. In the simulations τ is finite so there will be often some dispersion around the equilibria reached (exept in some cases and regimes). In the interpretation of the results, *converges to an equilibrium* will be used indifferently to say that the state reached is actually a Nash equilibrium or just an ϵ-Nash equilibrium.

Low temperature $\tau = 10^{-4}$

4.2.1 $0 < h < g < 1$

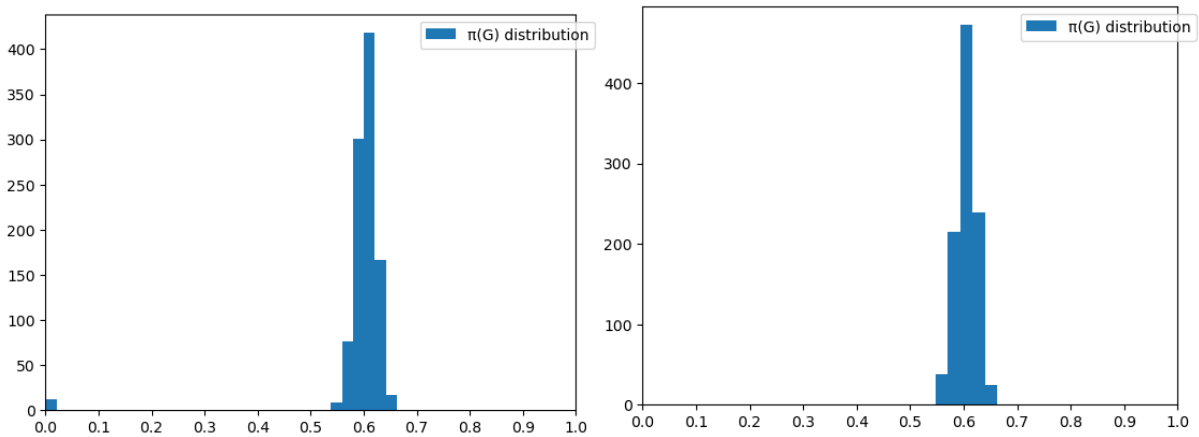
By setting in the hypothesis of convergence of the algorithm, the strategies converge to an asymmetric Nash equilibrium in which part of the players set on the strategy H and

the other part set on the mixed strategy $\tilde{p}$. When $m > 0.5$ the population choosing H is very small (Figure 4.3), while in the opposite case they represent the biggest part of the population (Figure 4.4). Numerical results resumed in Table 4.1. Notice that in this regime, the dynamics of q is faster than the one of π and asimpotically ($t \to \infty$), π is steady from the point of view of the evolution of q.



(a) *Distribution in the population* (b) *Distribution in the sub-population*

Figure 4.3: Distribution of $\pi^T(G)$ among the players, after $5 \cdot 10^3$ iterations. $h = 0.1$ and $g = 0.9$ and $m = 0.2$. Values in Table 4.1



(a) *Distribution in the population* (b) *Distribution in the sub-population*

Figure 4.4: Distribution of $\pi^T(G)$ among the players, after $5 \cdot 10^3$ iterations. $h = 0.1$ and $g = 0.9$ and $m = 0.6$. Values in Table 4.1

	$\tilde{p}$	$\mu_G\|_{mix}$	$\sigma_G\|_{mix}$	n_H	n_G
$m = 0.2$	0.339	0.343	0.147	411	0
$m = 0.6$	0.617	0.612	0.025	28	0

Table 4.1: $h = 0.1$ and $g = 0.9$. In both cases, part of the population chooses one the pure strategies H, while the other part accumulates around a single mixed strategy. $\mu_G|_{mix}$ approximates well $\tilde{p}$ but it makes sense to interpret it as an estimate of p^* only in the $m = 0.6$ case, where the H population is very small and the difference between this asymmetric equilibrium and the symmetric one is negligible.

The two thresholds $m = 0.2$ and $m = 0.6$ have been chosen as two non-extreme representatives of two corresponding classes: very intuitively this two classes are the set of cases with $m > 0.5$ and those with $m < 0.5$. Indeed, corresponding to the threshold 0.5 an abrupt decrease of the n_H population can be observed (Figure 4.5):

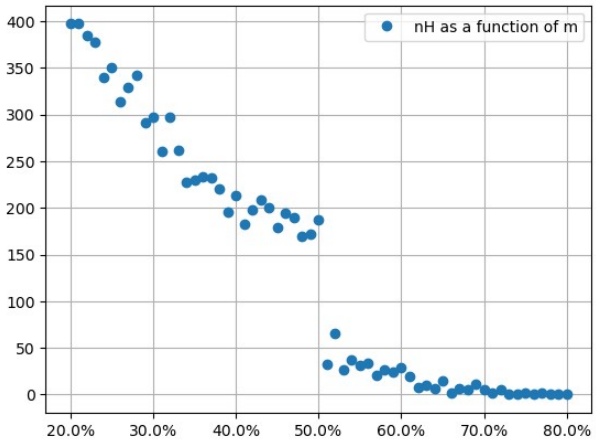
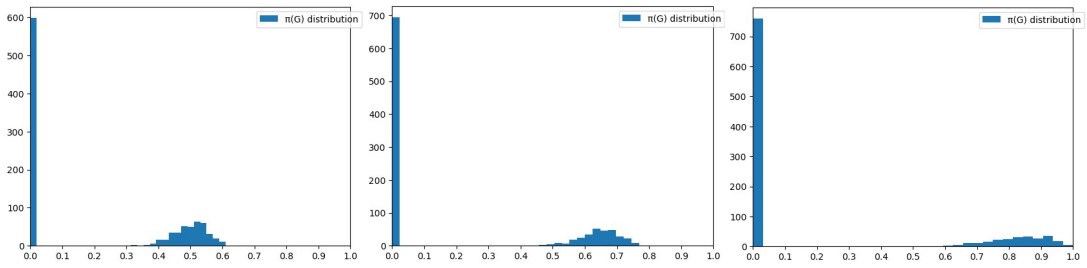
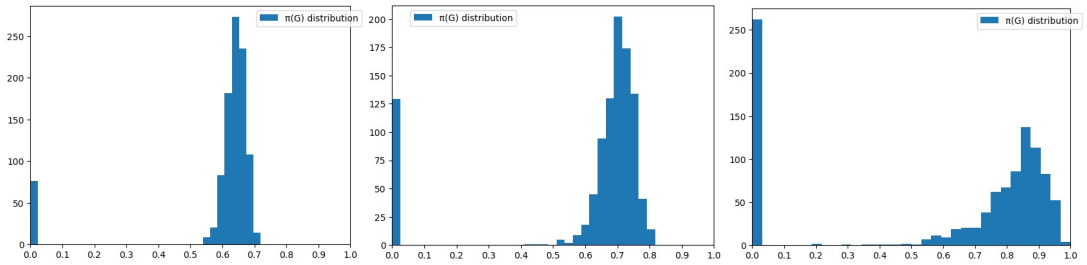


Figure 4.5: n_H population as a function of m after $5 \cdot 10^3$ iterations, with a stepsize of 1%

When the difference between h and g decreases two things happen: the n_H population in the case $m = 0.6$ grows, becoming predominant (as already happened in the $m = 0.2$ case) and in both cases the mean of the *mix* sub-population approaches 1, while still approximating well $\tilde{p}$ (which approaches 1 as well), but the mean squared error grows too (Figure 4.6).



(a) $m = 0.2$, $h = 0.2$, $g = 0.8$ (b) $m = 0.2$, $h = 0.3$, $g = 0.7$ (c) $m = 0.2$, $h = 0.4$, $g = 0.6$



(d) $m = 0.6$, $h = 0.2$, $g = 0.8$ (e) $m = 0.6$, $h = 0.3$, $g = 0.7$ (f) $m = 0.6$, $h = 0.4$, $g = 0.6$

Figure 4.6: Distribution of $\pi^T(G)$ among the players, after $5 \cdot 10^3$ iterations. $h = 0.1$ and $g = 0.9$ and $m = 0.6$

Looking at the dynamics of a generic player, two types of temporal evolution can be observed: players ending up in a mixed strategy are characterized by a q that does not converge to a specific value but oscillates until the end of the simulation. The π also exhibits oscillations but of a milder nature right from the start, stabilizing towards a value towards the end of the simulation. Players ending up in a pure strategy (in this case only the pure strategy H, but as we will see later, this is a characteristic of players converging to pure strategies in general) have a q that, after some oscillations, rapidly converges to a fixed value, just like the corresponding π.

• Let's consider the first type of dynamics (Figure 4.7): we have players playing mixed strategies who, during the temporal evolution, try one or the other strategy more or less frequently (depending on their value of $\pi^t(G)$). This, combined with the statistical fluctuations of the outcome of each stage game (i.e., the realization of $u(a^t)$), clearly leads to having a q^t that reflects these fluctuations. The fluctuations of q^t do not significantly decrease over time because the learning rate decreases like $\sim (\pi^t(G) \cdot t)^{-0.1}$, and after 5000 iterations, it will still be greater than 0.4 (it does not have a certain value because, as mentioned at the beginning, the learning rate of q varies from simulation to simulation, but a lower bound can still be determined). In contrast, the learning rate of $\pi^t(G)$ decreases according to $t^{-0.9}$ and it is already below 0.02 after 100 iterations and below 0.002 after 1000 iterations. This explains its much more stable dynamics and the fact that this stability increases as the simulation progresses. After a sufficently long time for the same reason also the value of q^t will stabilize but at that point the dynamics of π will be already fixed.

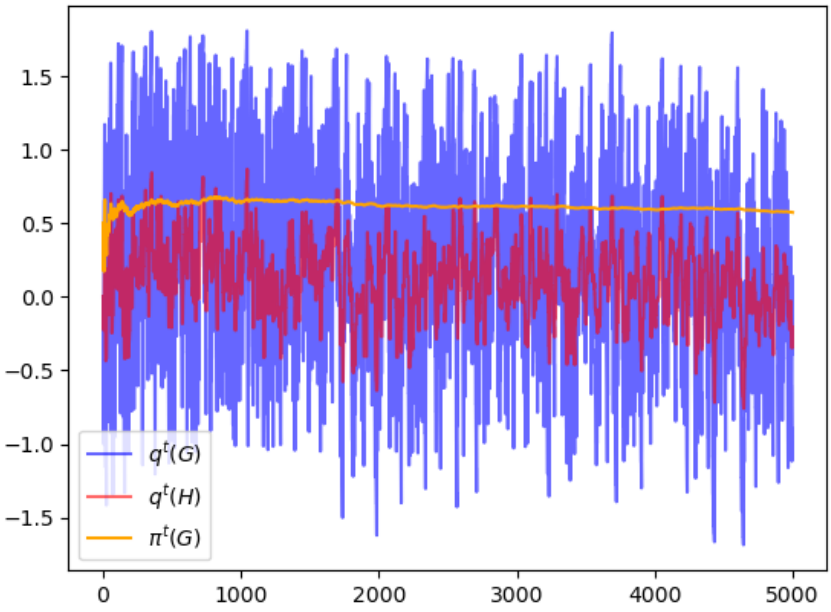
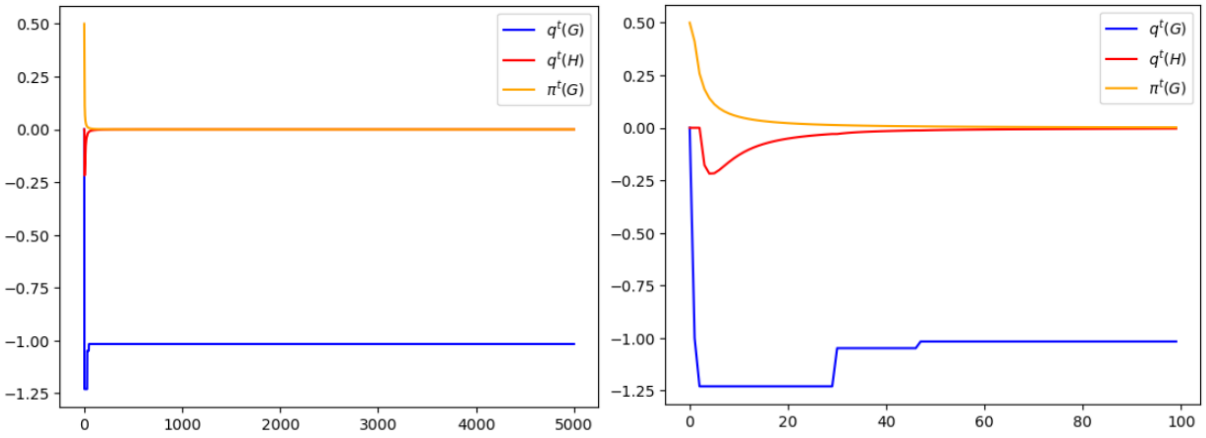


Figure 4.7: Dynamics of $\pi^t(G)$ and q^t for a player ending up in mixed strategies, as function of the iterations t

• Concerning the second type: the dynamics is characterized by a very rapid settling of the initial phases of both π and q (Figure 4.8).



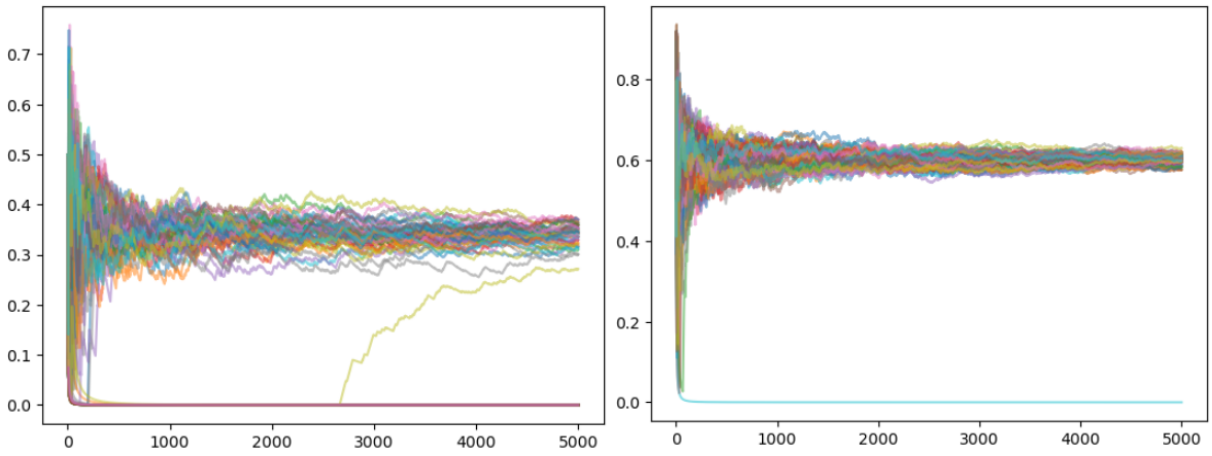
(a) *Complete dynamics*

(b) *Zoom out to the first 100 iterations, where all three quantities already stabilize*

Figure 4.8: Dynamics of $\pi^t(G)$ and q^t for a player ending up in the pure strategy H

Finally it's instructive to track the dynamics concerning the only $\pi^t(G)$ for a portion of the population taken as a sample. In Figure 4.9 one can see both types of players. Furthermore, a third type of players can be observed, much less frequent: those who initially exhibit a pure strategy dynamic and eventually transition into a mixed strategy. Remarkably it never happens the converse. It means that a fluctuation is able to induce a player, fixed in the H pure strategy to select a mixed strategy, which has to be more stable in some sense: a player j using the pure strategy H indeed, receives a payoff 0, independently on the other players choices. But due to the asyncronous update,

when playing strategy H, only the value corresponding to $q_j(H)$ is updated. The $q_j(H)$ converges rapidly and the $q(G)$ is never updated unless strategy G is chosen due to a flcuctuation induced by temperature. When this happens, it may happen that for that particular stage, G reveals the winning choice, and since $n(G)$ is still small, the corresponding learning rate is large, causing an abrupt increasing in the value of $q_j(G)$ and consequently in that of $\pi_j(G)$. A player already playing in mixed strategies is instead already tuning $\pi_j(G)$ playing both strategies and it results less sensible to fluctuations.



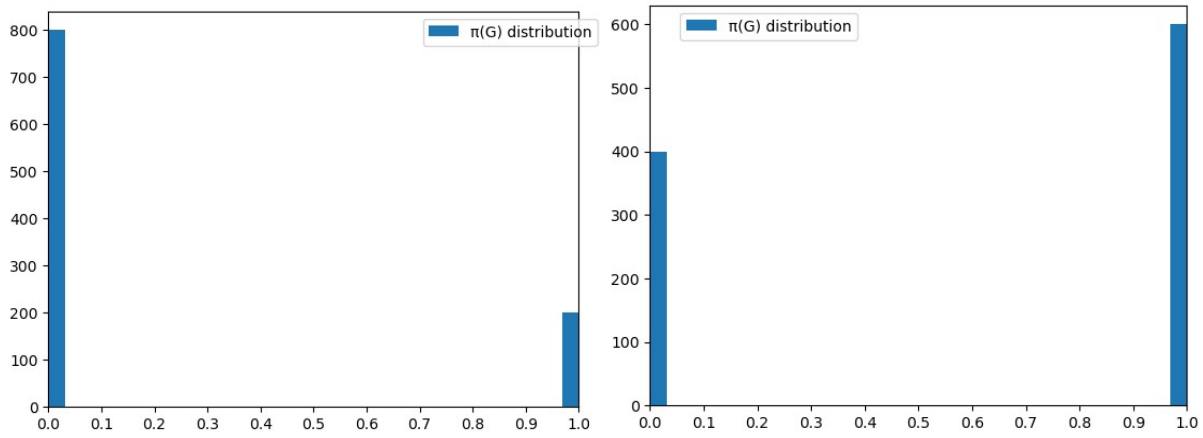
(a) *Sample of 100 players. m=0.2*

(b) *Sample of 100 players. m=0.6*

Figure 4.9: Dynamics of $\pi^t(G)$ for a sample of 100 players, as function of the iterations

4.2.2 $0 < g < h < 1$

When $h < g$ a new regime is reached: in this regime, all players end up in a pure strategy. The algorithm still converges to an equilibrium, but this time converges to one of the pure strategies Nash equilibria of the stage game, i.e. the ones in which $n_G = m$ and $n_H = 1 - m$ (Figure 4.10). Notice that also in this regime, the dynamics of π is faster than the one of q and asimpotically ($t \to \infty$) q is steady from the point of view of the evolution of π.

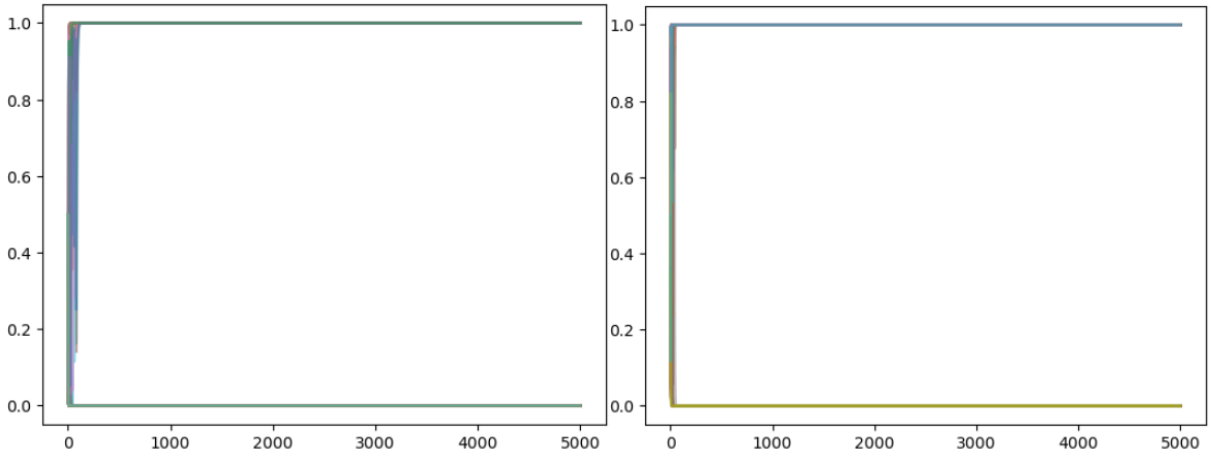


(a) $m = 0.2$, $h = 0.6$, $g = 0.4$ (b) $m = 0.6$, $h = 0.6$, $g = 0.4$

Figure 4.10: Distribution of $\pi^T(G)$ among the players, after $5 \cdot 10^3$ iterations. In both cases the convergence to the pure strategy equilibrium is perfect.
Case $m = 0.2$: $n_G = 200$ and $n_H = 800$
Case $m = 0.6$: $n_G = 600$ and $n_H = 400$

In all other cases: $h = 0.7$ and $g = 0.3$, $h = 0.8$ and $g = 0.2$, $h = 0.9$ and $g = 0.1$ a pure strategies state is reached: depending on the specific value of m and of the population N the actual equilibrium can be reached in some of these cases, increasing the number of iterations when needed. It turns out that a specific choice (or set of choices) for h and g leads to a pure Nash equilibrium. When the equilibrium is not reached, all players still fix on a pure strategy giving a strategy profile characterized by $n_G < M$. So in general the state reached is a pure strategies state with $n_G \leq 0$, where the case $= 0$ is a Nash equilibrium.

All the players are characterized by the same dynamics as the ones ending up in pure strategies in the other time-scales regime (Figure 4.11):

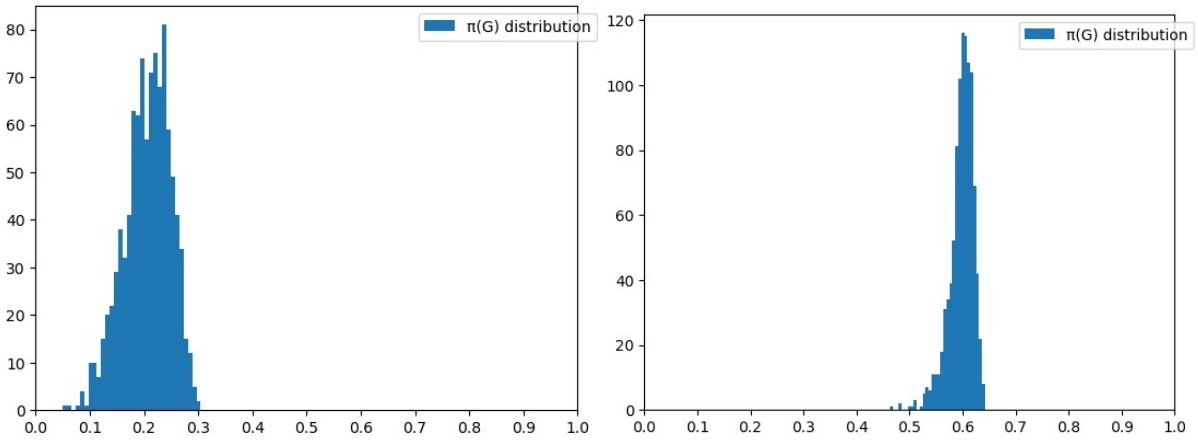


(a) *Sample of 100 players. m=0.2* (b) *Sample of 100 players. m=0.6*

Figure 4.11: Dynamics of $\pi^t(G)$ for a sample of 100 players, as function of the iterations

4.2.3 $0 < h < g$

By relaxing the request $g < 1$, in order to create a wider separation in the time-scales, the algorithm converges to the symmetric mixed strategies Nash equilibrium of the stage game (Figure 4.12). Numerical results resumed in Table 4.2. In the following $h = 0.1$ and $g = 1.2$.



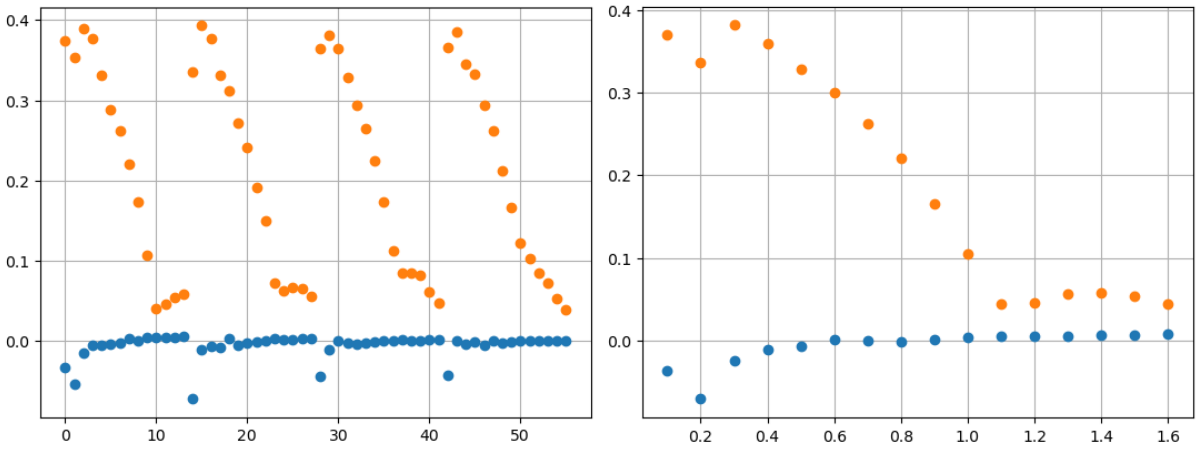
(a) $m = 0.2$ (b) $m = 0.6$

Figure 4.12: Distribution of $\pi^T(G)$ among the players, after 10^4 iterations. $h = 0.1$ and $g = 1.2$. Values in Table 4.2

	p^*	μ_G	σ_G	n_H	n_G
$m = 0.2$	0.200	0.206	0.043	0	0
$m = 0.6$	0.600	0.598	0.023	0	0

Table 4.2: $h = 0.1$ and $g = 1.2$. In both cases the distribution is peaked around the mixed strategy symmetric equilibrium and the whole population ended up in a mixed strategy.

In some range of the two time-scales, the values Err and σ_G^2 depend weakly on the absolute value of the single h and g, while depending strongly on their difference $g - h$. In order to see the dependence of the estimate on the time-scales, different ranges of h and g have been tested Figure 4.13. Below are some results results concerning this fact: for each value of h in a range between 0.1 and 0.4 (stepsize 0.1), the values of g in a range between h and $h + 1.3$ (stepsize 0.1) have been tested. The figures repeat periodically evereytime h changes value, meaning that in such ranges, the performance depends mostly on the difference $g - h$. The case $m = 0.2$ has been used because is the one presenting a higher σ_G. With much greater values of h, Err begins to grow. A similar simulation has been made, this time choosing only a range of g relative to $h = 0.1$, in order to determine an optimal distance $g - h$. This way, the values $h = 0.1$ and $g = 1.3$ of 4.12 have been selected.



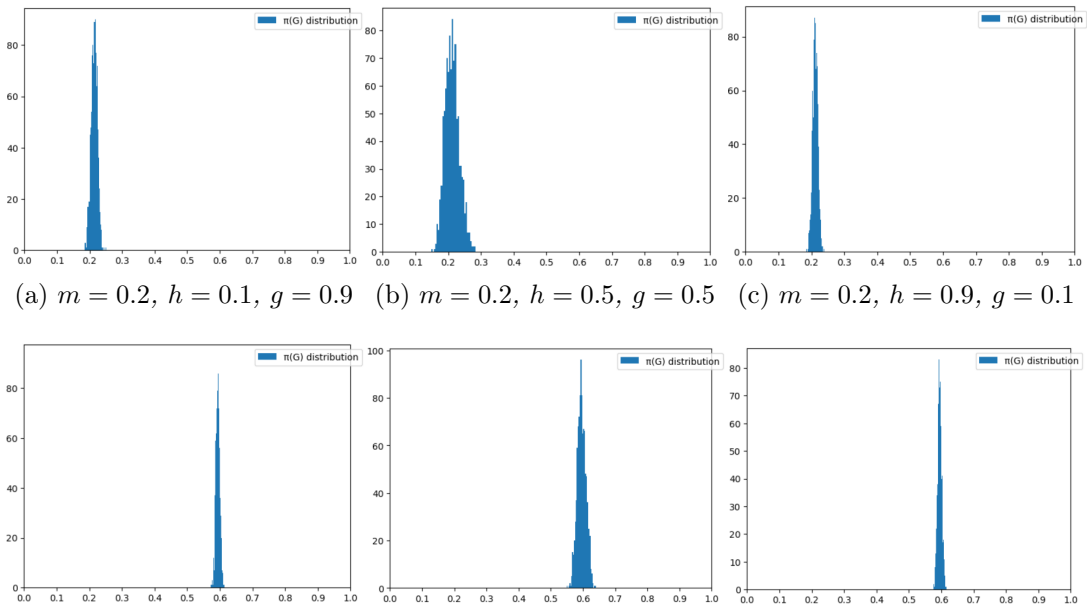
(a) h *between 0.1 and 0.4,* g *from* h *to* $h + 1.3$ (b) h *= 0.1,* g *from* 0.1 *to* 1.6

Figure 4.13: Err (in blue) is in most cases small but σ_G (in orange) decreases with increasing $g - h$.

The decreasing in σ_G^T is a sign of the fact that the distribution is becoming more and more dense around the mixed strategy. So separating enough the time-scales, the algorithm converges to the symmetric mixed strategies Nash equilibrium.

High temperature $\tau = 0.5$

When the temperature is sufficiently high, it can be observed that independently on the the time-scale regime, all players tend to converge to the symmetric mixed strategies equilibrium (Figure 4.14): the distribution is sharper as the difference $|h - g|$ is larger.

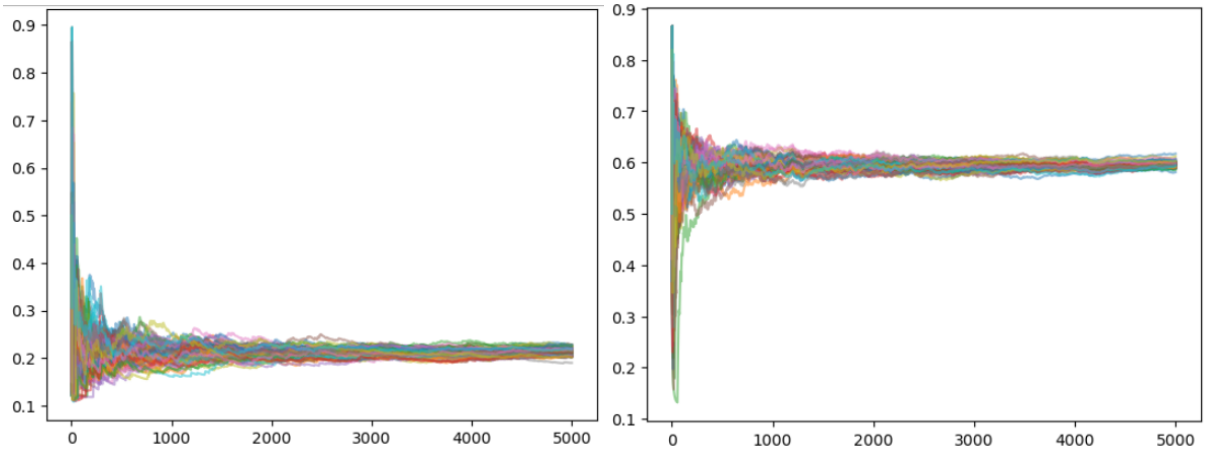


(a) $m = 0.2$, $h = 0.1$, $g = 0.9$ (b) $m = 0.2$, $h = 0.5$, $g = 0.5$ (c) $m = 0.2$, $h = 0.9$, $g = 0.1$

(d) $m = 0.6$, $h = 0.1$, $g = 0.9$ (e) $m = 0.6$, $h = 0.7$, $g = 0.3$ (f) $m = 0.6$, $h = 0.9$, $g = 0.1$

Figure 4.14: Distribution of $\pi^T(G)$ among the players, for different values of h and g after $5 \cdot 10^3$ iterations.

In Figure 4.15 one can see that during the dynamics the pure stategies are never systematically adopted, neither in the $m = 0.2$ case where at low temperature there was a great fraction of the players fixed in the strategy H, nor even in the $g < h$ regime where at low temperature all players adopted pure strategies.



(a) $m = 0.2$,*h=0.9,* $g = 0.1$ (b) $m = 0.6$,*h=0.1,* $g = 0.9$

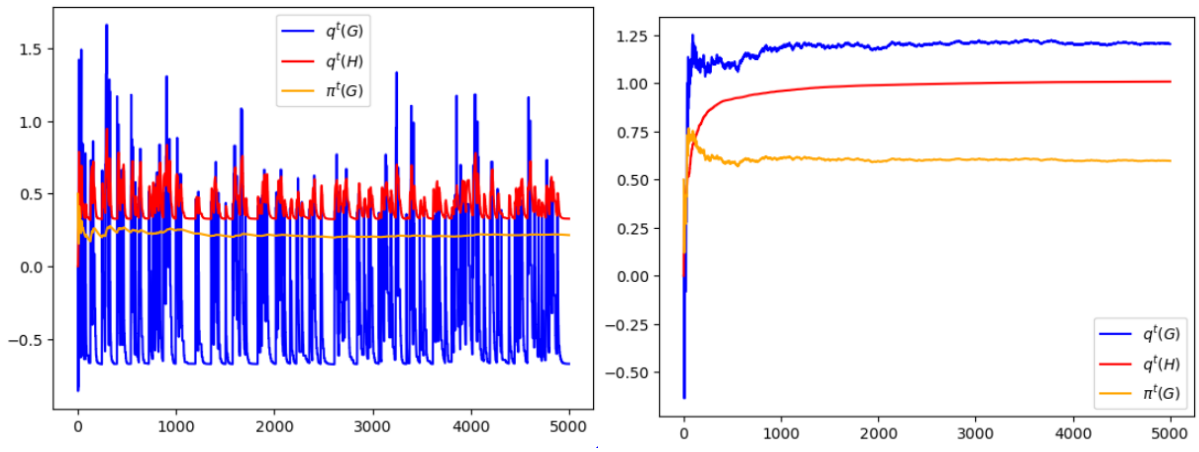
Figure 4.15: Two examples of dynamics of a 100 players sample in the high temperature regime $\tau = 0.5$

The fluctuations induced by the temperature are sufficently strong to make all the pure strategies player deviate enought to become mixed strategies players, as it happended rarely in the low temperature regime. This is not surprising since, with a temperature of the order of q it is very difficult to reach values of q that can lead the smoothed update of $\pi(G)$ towards the extremes 0 or 1. Moreover, clearly the symmetric mixed equilibrium in both time-scale regimes has some stability. This was evident in low temperature when players deviating from a pure strategy tended to approach the symmetric equilibrium strategy but never happened the converse and it is evident here, where temperature fluctuations are able to avoid players getting stuck in pure strategies but don't affect much players ending up in a strategy sufficiently close to the symmetric mixed strategy equilibrium.

Even if the two time-scale regimes lead to the same equilibrium convergence, the dynamics is still different (Figure 4.16):

- in the $h < g$ regime, q is still oscillating when π has already reached equilibrium: as already observed it has a very slowly decreasing learning rate so it is very sensible to the fluctuations of the single realizations of the stage game.

- in the $h > g$ regime instead, also q has a stable dynamics.



(a) $m = 0.2$,h=*0.1*, $g = 0.9$ (b) $m = 0.6$,h=*0.9*, $g = 0.1$

Figure 4.16: Typical dynamics in the high temperature regime

4.2.4 Summarizing numerical results

From the simulations at low temperature, two distinct regimes that govern the behavior of the system can be identified: a first regime, that can be referred to as fast-q regime (where $h < g$), and the opposite regime, fast-π regime ($h > g$). The very same regimes at sufficiently high temperature have the same behaviour.

- **fast-q**: at sufficiently high temperature the dynamics of π converges to the symmetric mixed strategy equilibrium. q keeps oscillating also after π fixes. At low temperature, the dynamics converges to states characterized by the presence of players ending up in mixed strategies. Moreover, by appropriately selecting the parameters, it is possible to ensure that the algorithm converges to:
  - one of the asymmetric Nash equilibria
  - the symmetric mixed Nash equilibrium (in some cases g has to be set greater than one, exiting the hypothesis of convergence of the original algorithm)
- **fast-π**: at sufficiently high temperature the dynamics of π converges to the symmetric mixed strategy equilibrium and also q stabilizes. At low temperature, the dynamics converge to states in which all players end up in pure strategies. Also in this case, by appropriately selecting the parameters, it is possible to ensure that the algorithm converges to one of the pure strategies Nash equilibria: the players learn to coordinate in the most efficient way, i.e. the total sum of the payoffs among the population is maximized.

4.2.5 The fast-π regime

Algorithm 4 has been proven to be convergent to a Nash equilibrium of the game in [7]. However, the regime in which $h > g$ is out of the convergence hypothesis of the algorithm. So on one hand it is not surprising that, at least in some settings, it gives different results. On the other hand, one may wonder why it still converges to Nash equilibria in most of the cases.

Let's consider first the limit of vanishing temperature $\tau \to 0$ and assume g to be sufficiently close to 0 to consider t^{-g} as 1: in this limit the update of π in (4.2) is simply given by (let's take the update corresponding to $a_i = G$):

$$\pi_i^t(G) = \begin{cases} 1 & \text{if} \quad q_i^t(G) > q_i^t(H) \\ 0 & \text{if} \quad q_i^t(G) < q_i^t(H) \\ \frac{1}{2} & \text{if} \quad q_i^t(G) = q_i^t(H) \end{cases} \tag{4.2}$$

So we have that whenever it happens that a player reinforces an action obtaining positive payoff or conversely the action is discouraged we have that the strategy instantaneously becomes a pure one and consequently fixed for the replicator dynamics. Let's inspect when this happens:

- at $t = 0$ $\pi_i(G) = \pi_i(H) = \frac{1}{2}$ and $q_i(G) = q_i(H) = 0$ for any player i.
- players choosing H as first action don't change q since $u(H, a_{-i}) = 0$ and the other terms in the update are also 0 (so they remain with $\pi_i = \frac{1}{2}$). Thus, the only ones changing π_i and q_i are those choosing G.

So we have two possible scenarios at the first time step:

$\star\star\star$

1. The fraction of players choosing G is bigger than M. In this case they all get a payoff -1, and their strategy becomes $\pi(G)_i = 0$ for the rest of the dynamics. On the other hand the other fraction will still randomize between the two actions. For this fraction of players the situation is the same as the beginning, with a smaller number of opponents.

2. The fraction of players choosing G is smaller than M. In this case they all get a payoff $+1$, and their strategy becomes $\pi(G)_i = 1$ for the rest of the dynamics. As in the other case, the other fraction will still randomize between the two actions. For this fraction of players the situation is the same as it was at the beginning, with a smaller number of opponents and a smaller M.

This process reiterates for the rest of the dynamics until all players will have tried once action G, settling in one or the other pure strategy.

Interestingly, it can't happen that the still randomly playing agents reinforce action G if at that stage the ones choosing G has exceeded M. So at the end there will be fraction $\leq M$ fixed in the strategy $\pi(G) = 1$ and the other fixed in the strategy $\pi(G) = 0$.

The Nash equilibrium is reached for values of h and g which are not extreme (in the simulations presented it was $h = 0.6$ and $g = 0.4$). Considering a still fast but finite speed dynamics also for π_i, the reasoning is qualitatively unchanged if not for the fact that the update (4.2) is now weighted by a learning rate, so when an action a_i is reinforced we have that $\pi_i(a_i)$ goes towards 1 but it doesn't become exactly 1. To be more precise, looking at (**??**) we see that at $t = 1$ reinforcing an action a_i means actually setting $\pi_i(a_i) = 1$, but this becomes less and less accurate as t grows. We can then conclude that at the first time step we are in the very same scenario as in $\star\star\star$. For $t > 1$

the reinforcing or discouraging of action G is smoother than before and so also players who played G and exceeded M will have non-zero probability of trying again action G. As long as the number of players fixed in the strategy G is smaller than M there will be a chance of reinforcing action G. When the population fixed in G is exactly M, all randomizing players will necessarily discourage action G.
Tuning properly h and g is crucial in order to obtain the Nash equilibrium, since we need π_i not to have a hard update, in orther to induce a trial and error behaviour. On the other hand we need also the learning rate of q_i to be sufficiently large to ensure that the information relative to a reinforcement or a discouragement arrives to π_i.

Concerning the *high-τ regime*, we can consider again an extremely fast π regime and consider (**??**): in this regime we observed that no player ends up in a pure strategy, so clearly the finite temperature prevents π_i from being updated in a hard way as in the low-τ case. On the other hand we also observe that the entropy term is not yet predominant since, in that case, the system would end up in a uniform state $\pi_i(G) = \pi_i(H) = \frac{1}{2}$. We observe instead that the system sets around a stable fixed point of the replicator dynamics, i.e. the ESS $\pi_i(G) = p^*$. We can conclude that in this regime the temperature prevents players from fixing in the pure strategies and the dynamics resulting is a perturbed replicator dynamics that lead the players around the stable fixed point which which is the symmetric mixed Nash equilibrium.

4.2.6 The fast-q regime

In [7], where algorithm 4 is proposed, a formal proof of the convergence of such algorithm is given. From a "mathematician" perspective this would be enough. This in not the case, so let's try to analyze this regime in the light of the considerations made for the other one, in order to interpret the numerical results. Clearly the considerations made to approximate the dynamics with a replicator equations cannot hold. Anyhow, we can say something about the low τ regime, exploiting the approach used in the other regime. Referring to $\star\star\star$ we can again conclude that at $t = 1$ a fraction of the players will choose G:

1. If this fraction is bigger than M, such players will discourage G and being in the low *tau* regime they will adopt the strategy $\pi_i(G) = 0$. As we already noticed, players choosing H get deterministically a payoff equal to 0, so such players will be fixed in such strategy. The probability of this case occuring is clearly decreasing with increasing M. This justifies why equilibria reached with lower values of M were characterized by a bigger n_H population. Still it is not sufficient to justify the abrupt decrease in such population occurring around $\frac{M}{N} = 0.5$ shown in figure 4.5.

2. If this fraction is bigger than M, such players will reinforce G and, being in the low τ regime, they will adopt the strategy $\pi_i(G) = 1$. Such strategy is not a fixed point of the dynamics, so in the following time steps, whenever it happens that one of such players chooses G resulting in an exceeding population, its strategy will change. If this happens in the earlier phase of the dynamics, it will probably discourage G enough to adopt the strategy $\pi_i(G) = 0$, becoming part of the n_H population. Once again this is more likely to happen with lower values of M, justifying in part the

difference in the populations between the different equilibria varying M but does not justify the discontinuity observed.

4.3 Analytical approach through Stochastic Approximation theory techniques

The mathematical framework in which the proofs of the convergence of algorithms applied in stochastic games is the so called Stochastic Approximation Theory, first introduced in [10]. Essentially, it comprises a set of techniques that rigorously formalize the approach to approximating a dynamic system in the presence of stochasticity, providing tools to describe a corresponding deterministic equation for the average values of the involved quantities. The set of convergence hypothesis of the stochastic games algorithms of chapter 3 are due to this approach. We shall now exploit some of the principles of this framework in order to analyze more formally the two regimes, primarily relying on [2]. Let's begin this time with the fast-q regime.

Consider two dynamical equations describing algorithm 4:

$$\pi_i^{t+1}(a_i) - \pi_i^t(a_i) = \beta_i^t \cdot \left[\frac{exp\left(\frac{q_i^t(a_i)}{\tau}\right)}{\sum_{a' \in A_i} exp\left(\frac{q_i^t(a')}{\tau}\right)} - \pi_i^t(a_i) \right]$$

$$q_i^{t+1}(a_i) - q_i^t(a_i) = \alpha_i^t \cdot \left[u_i(a^t) + \tau \cdot \nu(\pi_i^t(\cdot)) + \gamma \sum_{a' \in A_i} \pi_i^t(a') \cdot q_i^t(a') - q_i^t(a_i) \right]$$

where $\beta_i^t \propto t^{-g}$ and $\alpha_i^t \propto \left(\pi_i^t\left(a_i\right) t\right)^{-h} \propto t^{-h}$ and u_i is defined by (4.1). From the point of view of player i, the received payoff depends on her action and on the unobserved actions of the other players. It is therefore convenient to express the stage payoff $u_i\left(a_i, a_{-i}\right)$ received by player i as the sum of two terms: the average payoff value when the other players behave according to their current mixed strategies $\bar{u}_i\left(a_i, \pi_{-i}^t\right)$ and a random variable $\eta_i^t\left(a_i\right)$ with zero mean and a variance that can be also specified.[1] The original equations become

$$\pi_i^{t+1}(a_i) - \pi_i^t(a_i) = \beta_i^t \cdot \left[\frac{exp\left(\frac{q_i^t(a_i)}{\tau}\right)}{\sum_{a' \in A_i} exp\left(\frac{q_i^t(a')}{\tau}\right)} - \pi_i^t(a_i) \right]$$

$$q_i^{t+1}(a_i^t) - q_i^t(a_i^t) = \alpha_i^t \cdot \left[\bar{u}_i(a_i, \pi_i^t) + \eta_i^t\left(a_i\right) + \tau \cdot \nu(\pi_i^t(\cdot)) + \gamma \sum_{a' \in A_i} \pi_i^t(a') \cdot q_i^t(a') - q_i^t(a_i^t) \right]$$

[1] In fact, only the noise term $\eta_i^t\left(G\right)$ is non zero, still it has zero mean and fluctuations equal to

$$\left\langle \eta_i^t\left(G\right) \eta_j^t\left(G\right) \right\rangle = \Omega \delta_{ij} \delta\left(t - t'\right)$$

with

$$\Omega = \left(1 - \omega\right) \omega \left(1 + x\right)^2$$

Under the already specified conditions $h, g \in (0,1)$, it is now possible to apply stochastic approximation theory to derive a system of ODE for the two quantities under consideration (we now write explicitely ν defined in chapter 3),

$$\begin{aligned}
\epsilon_i^t \frac{d}{dt}\pi_i^t(a_i) &= \frac{exp\left(\frac{q_i^t(a_i)}{\tau}\right)}{\sum_{a_i'} exp\left(\frac{q_i^t(a_i')}{\tau}\right)} - \pi_i^t(a_i) \\
\frac{d}{dt}q_i^t(a_i) &= \left\{ \bar{u}_i\left(a_i, \pi_{-i}^t\right) - \tau \sum_{a_i'} \pi_i^t(a_i') \log \pi_i^t(a_i') + \gamma \sum_{a_i'} \pi_i^t(a_i')\, q_i^t(a_i') - q_i^t(a_i) \right\}
\end{aligned}$$

where $\epsilon_i^t \propto \alpha_i^t/\beta_i^t$. Introducing $\pi_i(t) = \pi_i^t(G)$ (and $1 - \pi_i(t) = \pi_i^t(H)$) and $\delta q_i(t) = q_i^t(G) - q_i^t(H)$, we get the following simplified ODEs

$$\begin{aligned}
\epsilon_i^t \frac{d}{dt}\pi_i(t) &= \frac{exp\left(\frac{\delta q_i(t)}{\tau}\right)}{1 + exp\left(\frac{\delta q_i(t)}{\tau}\right)} - \pi_i(t) \\
\frac{d}{dt}\delta q_i(t) &= \left\{ \bar{u}_i\left(G, \pi_{-i}^t\right) - \bar{u}_i\left(H, \pi_{-i}^t\right) - \delta q_i(t) \right\}
\end{aligned}$$

We first consider the regime $0 < g < h < 1$, in which $\epsilon_i^t \to 0$ as $t \to \infty$. In this regime, we expect that π_i rapidly converges to a local fixed point

$$\pi_i = \frac{exp\left(\frac{\delta q_i}{\tau}\right)}{1 + exp\left(\frac{\delta q_i}{\tau}\right)}$$

which is slave with respect to the current value of δq_i, while the latter has to be determined considering its dynamics in which π_i^t appears as $\pi_i^t = \pi_i[\delta q_i^t]$, i.e.

$$\begin{aligned}
\frac{d}{dt}\delta q_i(t) &= \left\{ \bar{u}_i\left(G, \pi_{-i}^t\right) - \bar{u}_i\left(H, \pi_{-i}^t\right) - \delta q_i(t) \right\} \\
&= \left\{ \bar{u}_i\left(G, \pi_{-i}\left[\delta q_{-i}^t\right]\right) - \bar{u}_i\left(H, \pi_{-i}\left[\delta q_{-i}^t\right]\right) - \delta q_i(t) \right\}.
\end{aligned}$$

It is convenient to retrieve an equation for the mixed strategy profiles

$$\begin{aligned}
\frac{d}{dt}\pi_i(t) &= \frac{d}{dt}\frac{exp\left(\frac{\delta q_i}{\tau}\right)}{1 + exp\left(\frac{\delta q_i}{\tau}\right)} \\
&= \frac{1}{1 + exp\left(\frac{\delta q_i(t)}{\tau}\right)} \frac{d}{dt} exp\left(\frac{\delta q_i(t)}{\tau}\right) - \frac{exp\left(\frac{\delta q_i(t)}{\tau}\right)}{\left(1 + exp\left(\frac{\delta q_i(t)}{\tau}\right)\right)^2} \frac{d}{dt} exp\left(\frac{\delta q_i(t)}{\tau}\right) \\
&= \frac{1}{\tau}\pi_i(t)\left(1 - \pi_i(t)\right)\frac{d}{dt}\delta q_i(t) \\
&= \pi_i(t)\left(1 - \pi_i(t)\right)\left\{\tau^{-1}\left[\bar{u}_i\left(G, \pi_{-i}(t)\right) - \bar{u}_i\left(H, \pi_{-i}(t)\right)\right] - \delta q_i(t)/\tau\right\} \\
&= \pi_i(t)\left(1 - \pi_i(t)\right)\left\{\tau^{-1}\left[\bar{u}_i\left(G, \pi_{-i}(t)\right) - \bar{u}_i\left(H, \pi_{-i}(t)\right)\right] - \log\frac{\pi_i(t)}{1 - \pi_i(t)}\right\}
\end{aligned}$$

We have obtained a generalized version of replicator-equations, also known as Sato-Crutchfield equations. The average payoff difference depends on the current mixed strategies of the other players, which makes an explicit calculation of the fixed points of the dynamics non-trivial. If we assume that all players behave symmetrically, i.e. $\pi_j(t) = \pi_i(t) = \pi(t)$, then the average payoff difference can be easily estimated. In fact

$$\begin{aligned}\bar{u}_i(G, \pi_{-i}(t)) - \bar{u}_i(H, \pi_{-i}(t)) &= \bar{u}_i(G, \pi_{-i}(t)) \\ &= x\omega\left[\pi^t\right] - 1\left(1 - \omega\left[\pi^t\right]\right) = (1+x)\,\omega\left[\pi^t\right] - 1\end{aligned}$$

where $\omega[\pi^t] = \Pr\left[\mathcal{A}_{-i} \leq M-1 | \{\pi_j(t)\}_{j\in I}\right]$. The latter quantity can be approximated as follows

$$\begin{aligned}\omega\left[\pi^t\right] &= \Pr\left[\mathcal{A}_{-i} \leq M-1 | \{\pi_j(t)\}_{j\in I}\right] \\ &\approx \Pr\left[\mathcal{A}_{-i} \leq M-1 | \pi(t) = \frac{1}{N}\sum_j \pi_j(t)\right] \\ &\approx \sum_{\ell=0}^{M-1} \binom{N-1}{\ell} \pi(t)^\ell (1-\pi(t))^{N-1-\ell}\end{aligned}$$

Considering this mean-field version of the Sato-Crutchfield equations,

$$\frac{d}{dt}\pi_i(t) = \pi_i(t)(1-\pi_i(t))\left\{\tau^{-1}\left[(1+x)\,\omega\left[\pi^t\right] - 1\right] - \log\frac{\pi_i(t)}{1-\pi_i(t)}\right\}$$

for small values of τ, the dynamics of $\pi_i(t)$ rapidly fix in one of the two absorbing states $\pi_i^t = 0$ or $\pi_i^t = 1$ depending on $\omega[\pi]$: if $\omega[\pi] > 1/(1+x)$, then the strategy π_i increases until it reaches the pure strategy G, otherwise it rapidly converges to the pure strategy H. Only when τ is large the entropic term associated to the effects of the "thermal" noise promotes the survival of purely mixed strategies. In the dynamics we also observe that, in which regime, the relative proportion of pure strategies played when τ is small is consistent with the already determined pure-strategy equilibria. This is due to the fact that the drift associated with the average payoff term, and determining the convergence of individual mixed strategies, vanishes when $\omega[\pi^t] = 1/(1+x)$, which corresponds to the equilibrium condition at the population level.

Then we consider the opposite regime $0 < h < g < 1$, in which $\epsilon_i^t(a_i) \to \infty$ as $t \to \infty$. It is convenient to rescale time in order to write the ODE system as

$$\begin{aligned}\frac{d}{dt}\pi_i(t) &= \left(\frac{exp\left(\frac{\delta q_i(t)}{\tau}\right)}{1 + exp\left(\frac{\delta q_i(t)}{\tau}\right)} - \pi_i(t)\right) \\ \frac{1}{\epsilon_i^t}\frac{d}{dt}\delta q_i(t) &= \bar{u}_i\left(G, \pi_{-i}^t\right) - \bar{u}_i\left(H, \pi_{-i}^t\right) - \delta q_i(t)\end{aligned}$$

Now the second equation converges faster than the first one and we can use again the separation of scales to state that δq_i can be considered a slave variable of π_{-i}^t by the relation

$$\begin{aligned}\delta q_i &= \bar{u}_i\left(G, \pi_{-i}^t\right) - \bar{u}_i\left(H, \pi_{-i}^t\right) \\ &= (1+x)\,\omega\left[\pi^t\right] - 1.\end{aligned}$$

Moreover, we expect that the slow variable $\pi_i(t)$ follows the equation

$$\frac{d}{dt}\pi_i(t) = \frac{exp\left(\frac{\omega[\pi^t](1+x)-1}{\tau}\right)}{exp\left(\frac{\omega[\pi^t](1+x)-1}{\tau}\right)+1} - \pi_i(t).$$

We notice that, for small values of τ,

$$\begin{cases} \pi_i^t \to 1 & \text{for } \omega[\pi^t] > 1/(1+x) \\ \pi_i^t \to 0 & \text{for } \omega[\pi^t] < 1/(1+x) \end{cases}$$

Using again the mean-field approximation for $\omega[\pi^t]$, we can analyze more in detail this relation. In the mean-field approximation, the function $\omega[\pi^t]$ is a decreasing function of π^t, therefore when $\pi_i^t \to 1$ we have $\omega[\pi^t] < 1/(1+x)$, which means that $\pi_i(t)$ decreases. On the contrary then $\pi_i^t \to 0$ then $\omega[\pi^t] > 1/(1+x)$ and certainly $\pi_i(t)$ increases. It means that, for all indivuals i the mixed-strategy dynamics undergoes a negative feedback mechanism that stops when π_i reaches an equilibrium point π^* at which

$$\pi^* \approx \frac{exp\left(\frac{\omega[\pi^*](1+x)-1}{\tau}\right)}{exp\left(\frac{[\omega[\pi^*](1+x)-1]}{\tau}\right)+1}.$$

Since the r.h.s. is almost a step function at small temperature (when $\tau \ll 1$), in this regime the fixed point is very close to the mixed Nash equilibrium, whereas it deviates towards $\pi^* \to 0.5$ for large temperatures.

4.4 Minority games

Let's consider a slight (but crucial) modification of the El Farol bar problem. Again, consider a game in which all players have a binary choice. A common choice is to assign an explicit numerical value to the choices: $A_i := \{-1, 1\}$. Let $\mathcal{A} := \sum_{i \in I} a_i$. The utility function of this symmetric game is given by

$$u(a_i, a_{-i}) = -a_i \mathcal{A}$$

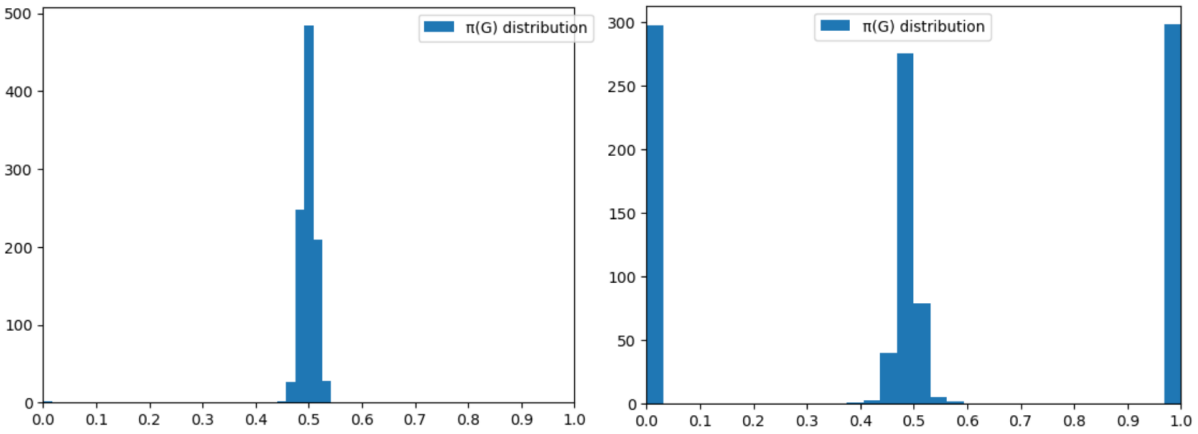
It is evident that player i gets a positive reward whenever his choice is 1 and $\mathcal{A} < 0$ or -1 and $\mathcal{A} > 0$, so whenever his choice is in the minority. In other words one way to see this game is that the purpose of each player is that of predicting what will be the sign of $\mathcal{A}$. This in just one possible formulation of a minority game: one possible modification/generalization would be that of choosing another antisymmetric function of $\mathcal{A}$. There can be identified several Nash-equilibria: a symmetric mixed strategy equilibrium in which all players use a strategy $\pi_i(+1) = \pi_i(-1) = \frac{1}{2}$ and $E[\mathcal{A}] = 0$. Then, if N is even there are $\binom{N}{N/2}$ pure strategy strict Nash equilibria in which the players split perfectly between the two strategies and $\mathcal{A} = 0$. If N is odd, there are $\binom{N}{1+N/2}$ pure strategy weak Nash equilibria with $\mathcal{A} = \pm 1$. Finally, always in the odd case it can be proved that there are countably many equilibria in which $N/2 - m$ of the players choose $+1$, $N/2 - m$ of the players choose -1 and the others play a mixed strategy.

4.5 Numerical results for the minority game

The minority game exhibits a simpler stucture with respect to the the El Farol bar problem game: indeed here tere is no asymmetry between the two strategies, and there is no neutral choice such as the H strategy of the El Farol bar problem, that gave deterministically a 0 reward. Indeed by performing the same simulations as in the other game, the same results come out with the difference that in the low-τ fast-π regime, there is no preferred choice, so the populations can be unbalanced in both the actions.

Low temperature $\tau = 0.1$

In this formulation, the scale of payoffs has the order of the population, so we rescale the two temperature regimes by $N = 1000$. In the fast-q regime the results are analogous to the El Farol bar problem ones, and the strategy profile converge either to the symmetric Nash equilibrium (figure 4.17a) or to one of the asymmetric Nash equilibria (figure 4.17b), depending on the specific choice of h and g.



(a) h=*0.1*, $g = 0.9$ (b) h=*0.4*, $g = 0.6$

Figure 4.17: Depending on the specific choice of the exponents, one between the symmetric or an asymmetric equilibrium is selected

In the fast-π regime analogously, the same results as in the El Farol bar problem are found again, and by tuning h and g one may recover the pure strategy equilibrium (figure 4.18). In the other cases, as already anticipated, there is no preferred choice, so the population may result unbalanced in one or the other (fgures 4.19a 4.19b)

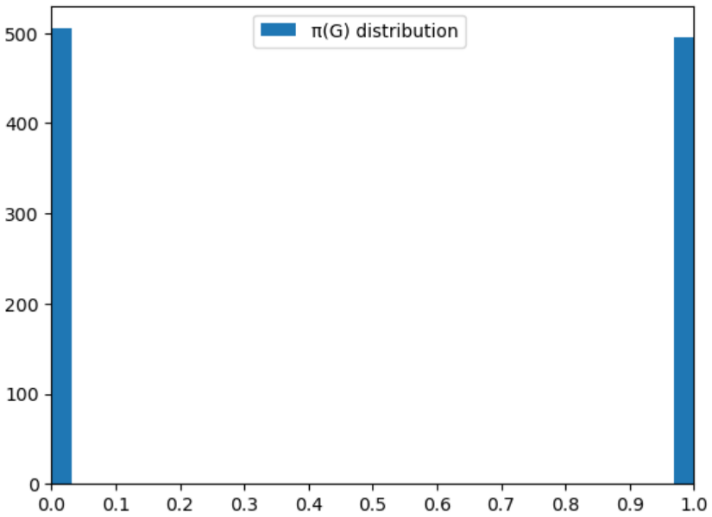
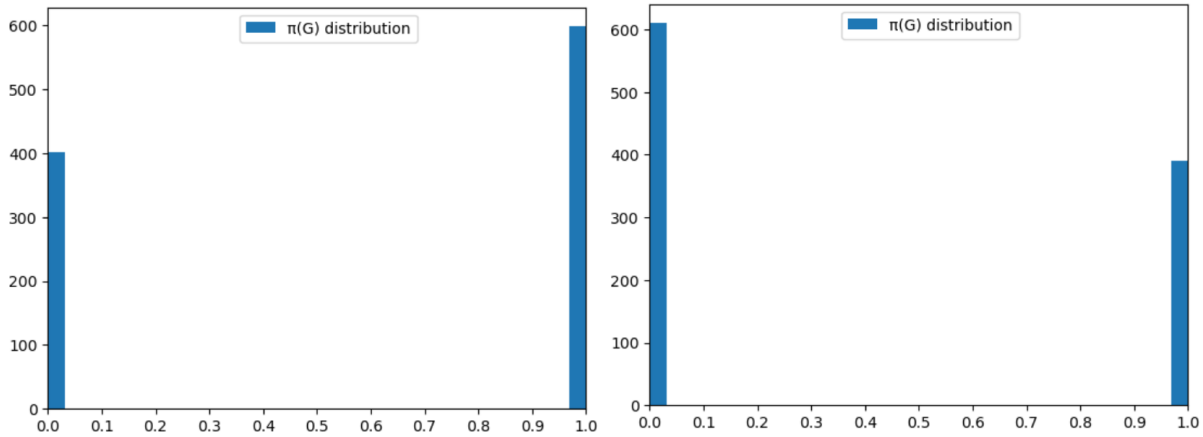


Figure 4.18: By tuning the exponents the pure Nash equilibrium is reached

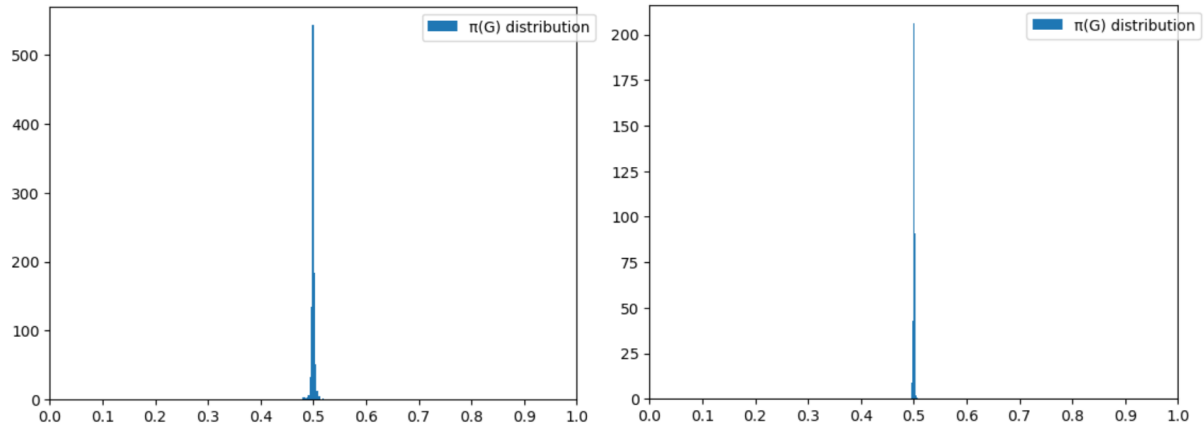


(a) h=*0.9*, $g = 0.1$ (b) h=*0.9*, $g = 0.1$

Figure 4.19: Only pure states are reached and there is no preferred choice

High temperature regime $\tau = 500$

In both time-scale regimes the symmetric Nash equilibrium is reached (figure 4.20)



(a) h=*0.9*, $g = 0.1$ (b) h=*0.1*, $g = 0.9$

Figure 4.20: In both cases the symmetric Nash equilibrium is reached

Chapter 5

Conclusion

This study applied a reinforcement learning algorithm initially designed for stochastic games to the repeated El Farol bar problem. This algorithm featured two temporal scales governing the evolution of player strategies (π) and a Q-function averaged over opponent strategies (q). The original algorithm exhibited a fast temporal evolution for q, treating π as stationary. However, by modifying the temporal scales to allow for rapid evolution of π, we observed a shift in equilibrium outcomes. Specifically, while the original regime led to convergence to symmetric mixed strategies, the reversed regime resulted in the emergence of pure strategy states, including pure strategy equilibria.
This work highlights the role of learning dynamics in equilibrium selection and stability, confirming how equilibrium stability is not just an intrinsic property of the game but is significantly influenced by the learning dynamics. Moreover, focusing on the Farol's Bar problem, we find that learning dynamics dictate whether players converge to pure coordination equilibria or resort to selfish strategies that yeld the same result as a coordination in the long term.
These findings offer valuable insights also considering how reinforcement learning dynamics mirror human learning and interpretation of real-world phenomena and events. By investigating the interplay between learning dynamics and equilibrium outcomes, it is possible to get a deeper understanding of strategic interactions in complex environments. In essence, this thesis underlines the intricate relationship between learning dynamics and equilibrium selection, shedding light on the adaptive nature of decision-making processes in dynamic environments.

Bibliography

[1] W. Brian Arthur. Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, 1994. URL https://www.jstor.org/stable/2117868.

[2] B. Bharath and V.S. Borkar. Stochastic approximation algorithms: Overview and recent trends. *Sadhana*, 24:425–452, 1999. doi: 10.1007/BF02823149. URL https://doi.org/10.1007/BF02823149.

[3] Maciej Bukowski and Jacek Miękisz. Evolutionary and asymptotic stability in symmetric multi-player games. *Institute of Economics, Warsaw School of Economics, Aleje Niepodległości 162, 02-554 Warsaw, Poland*, 2003. Received October 2001/Revised May 2003.

[4] D. Fudenberg and D. K. Levine. *The theory of learning in games*. MIT Press, Cambridge, MA., 1998.

[5] Jonathan Levin. *Learning in Games*. Stanford University, May 2006.

[6] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. *Brown University / Bellcore, Department of Computer Science*, 1994.

[7] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. Independent and decentralized learning in markov potential games. 05 2022. doi: 10.48550/arXiv.2205.14590.

[8] Yew-Soon Ong, editor. *Adaptation, Learning, and Optimization*, volume 12 of *Series Editor-in-Chief: Meng-Hiot Lim*. Nanyang Technological University, Singapore.

[9] Asuman Ozdaglar, Muhammed O. Sayin, and Kaiqing Zhang. Independent learning in stochastic games, 2021.

[10] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

[11] Lloyd S. Shapley. Stochastic games. *Mathematics*, 1953. Communicated by J. von Neumann, July 17, 1953.

[12] Karl Tuyls and Ann Nowe. Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review*, 20:63–90, 03 2005. doi: 10.1017/S026988890500041X.

[13] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. pages 693–700, 07 2003. doi: 10.1145/860575.860687.