Bifurcation Analysis of Reinforcement Learning Agents in the Selten's Horse Game

Alessandro Lazaric, Enrique Munoz de Cote, Fabio Dercole, and Marcello Restelli

Politecnico di Milano, Department of Electronics and Information, piazza Leonardo da Vinci 32, I-20133 Milan, Italy
{lazaric,munoz,dercole,restelli}@elet.polimi.it

Abstract. The application of reinforcement learning algorithms to multiagent domains may cause complex non-convergent dynamics. The replicator dynamics, commonly used in evolutionary game theory, proved to be effective for modeling the learning dynamics in normal form games. Nonetheless, it is often interesting to study the robustness of the learning dynamics when either learning or structural parameters are perturbed. This is equivalent to unfolding the catalog of learning dynamical scenarios that arise for all possible parameter settings which, unfortunately, cannot be obtained through "brute force" simulation of the replicator dynamics. The analysis of bifurcations, i.e., critical parameter combinations at which the learning behavior undergoes radical changes, is mandatory. In this work, we introduce a one-parameter bifurcation analysis of the Selten's Horse game in which the learning process exhibits a set of complex dynamical scenarios even for relatively small perturbations on payoffs.

1 Introduction

Game Theory (GT) [9] provides formal models (i.e., games) for the study of the interaction between self-interested rational agents, whose goal is the maximization of the return (i.e., payoff). In particular, GT identifies the conditions for the existence of equilibria (e.g., Nash equilibria), i.e., strategic configurations in which no agent can change her strategy without worsening her payoff. Nonetheless, the computation of equilibria requires each agent to have a complete knowledge of the game (actions available to other agents and their payoffs).

On the other hand, Reinforcement Learning (RL) [11] enables autonomous agents to learn the optimal strategy that maximizes the return through a direct interaction with an unknown environment. Multiagent Reinforcement Learning [8] extends the traditional single-agent RL approach to game theoretic problems in which several agents interact. Although RL algorithms are guaranteed to find the optimal (Nash) strategy in problems with stationary environments, they may fail to converge in environments where other learning agents are involved. As a result, the learning process may exhibit very complex non-convergent (periodic or aperiodic) dynamics [6,10] that are often difficult to study by stochastically simulating single runs of execution.

K. Tuyls et al. (Eds.): Adaptive Agents and MAS III, LNAI 4865, pp. 129–144, 2008.

Evolutionary Game Theory (EGT) [4] studies the evolution of populations of agents as dynamical systems, notably with the replicator dynamics equation. The translation of Q-learning [13], one of the main RL algorithms, into suitable replicator dynamics [1,12] makes possible the study of the dynamics of the learning processes as the study of nonlinear dynamical systems. The simulation (numerical integration) of the replicator dynamics therefore provides an alternative approach to study the behavior of learning agents, which is however effective only when all parameter values are assigned. In fact, as better explained in Sec. 4.2, how robust the observed learning dynamics are, when either learning parameters or parameters defining the structure of the game change because of noise or system perturbations, cannot be assessed by simply organizing extensive simulations.

Bifurcation analysis [7] provides strong theoretical foundations and effective numerical techniques to study the robustness of a dynamical system to parameter perturbations. In particular, robustness, called *structural stability* in the dynamical system jargon, is lost at the critical parameter combinations, called *bifurcations*, at which arbitrarily small parameter perturbations induce radical qualitative, other than quantitative, changes in the system dynamics.

In this paper, we introduce bifurcation theory and we apply it to the analysis of the dynamics of the learning process in a three agents representative extensive form game: the Selten's Horse. We investigate the problem characteristics and the learning solutions through a bifurcation analysis with respect to one of the payoffs of the game. In particular, we show that the dynamical system can repeatedly loose structural stability even in relatively small payoff intervals, that multiple stationary and non-stationary (periodic) attractors can be present, and that several bifurcations regulate their appearance, disappearance, and the catastrophic transitions between them.

The rest of the paper is organized as follows. In Section 2 we introduce definitions of normal and extensive form games. In Section 3 we briefly review Q-learning and how its dynamics can be translated into replicator-like dynamics. An introduction to bifurcation analysis is provided in Section 4 and, finally, in Section 5 we analyze the Selten's Horse game as a case study for bifurcation analysis of multiagent reinforcement learning systems.

2 Game Theory Background

2.1 Normal Form Games

In Game Theory, games are defined as conflict situations between agents. In a normal form game, agents execute actions simultaneously according to their strategies and the outcome of the game is a payoff for each agent. Formally:

Definition 1. *A normal form game Γ is defined by the tuple $\langle \mathcal{N}, \mathcal{A}, \mathcal{R} \rangle$, where:*

- $\mathcal{N} = \{1, \ldots, n\}$ *is the set of agents in the game*
- $\mathcal{A} = A_1 \times \ldots \times A_i \times \ldots \times A_n$ *is the set of joint actions* $\mathbf{a} = (a_1, \ldots, a_i, \ldots, a_n)$, *where* a_i *is an element of the set* $A_i = \{a_{i1}, \ldots, a_{ij}, \ldots, a_{im_i}\}$ *of the* m_i *actions available to agent* i *(*$m_i = m$ *in the following)*
- $\mathcal{R} = \{R_1, \ldots, R_n\}$ *is the set of payoff functions, where* $R_i : \mathcal{A} \to \Re$ *is the payoff function for agent* i *that maps each joint action to a numerical payoff*

Furthermore, we define:

- $\mathcal{X} = X_1 \times \ldots \times X_i \times \ldots \times X_n$ as the set of joint strategies $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n)$, where strategy $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im})$ is a probability distribution over the action set A_i, so that $\mathbf{x}_i \in \Sigma_m = \{\mathbf{x}_i : 0 \leq x_{ij} \leq 1, \sum_{j=1}^{m} x_{ij} = 1\}$, where Σ_m is the m-dimensional simplex
- $\rho = \{\rho_1, \ldots, \rho_n\}$ as the set of expected payoff functions, where $\rho_i : \mathcal{X} \to \Re$ is the expected payoff function for agent i that maps each joint strategy to a numerical payoff, that is the sum of the payoffs for all the possible joint actions weighted by their probabilities according to the joint strategy

At each round of the game, each agent chooses an action a_i, a joint action $\mathbf{a}$ is executed, and a payoff $R_i(\mathbf{a})$ is returned. When an agent plays deterministically one action (say a_{ij} with $x_{ij} = 1$), then the strategy is *pure*, otherwise is a *mixed* strategy. The joint action of all agents but agent i is usually denoted as $\mathbf{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \in \mathcal{A}_{-i} = A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n$. Similarly, the joint strategy of all the agents but i is defined as $\mathbf{x}_{-i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$. In the following, we refer to *matrix* games, in which the payoff functions R_i are matrices P_i with dimensions $|A_i| \times |\mathcal{A}_{-i}|$, i.e. $\rho_i(\mathbf{x}) = \mathbf{x}_i P_i \mathbf{x}_{-i}$.

The main solution concept in a normal form game is the *Nash equilibrium.*

Definition 2. *Given a normal form game* $\Gamma = \langle \mathcal{N}, \mathcal{A}, \mathcal{R} \rangle$, *the joint strategy* $\mathbf{x}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)$ *is a Nash equilibrium when:*

$$\rho_i(\mathbf{x}_1^*, \ldots, \mathbf{x}_i^*, \ldots, \mathbf{x}_n^*) \geq \rho_i(\mathbf{x}_1^*, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n^*), \quad \forall i \in \mathcal{N}, \forall \mathbf{x}_i \in \Sigma_m. \tag{1}$$

In a Nash equilibrium none of the agent can improve her expected payoff by changing her strategy while all other agents keep playing the same strategies. In other words, each strategy $\mathbf{x}_i^*$ is the best response to $\mathbf{x}_{-i}^*$.

2.2 Extensive Form Games

In contrast with normal form games, extensive form games describe the sequential structure of decision making explicitly, and therefore allow the study of situations in which agents play one after the other and possibly several times at different stages of the game round [9]. An extensive form game is represented by a tree (Fig. 1). Each node represents a *state* of play of the game. The game begins at a unique initial node, and flows through the tree along a path determined by the actions taken by the agents until a terminal node is reached, where the game ends and payoffs are assigned to agents. At each non-terminal node only

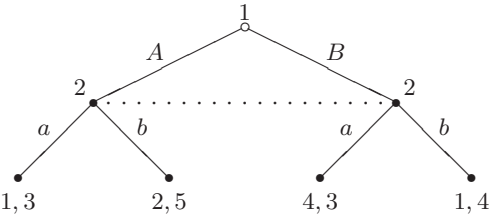


Fig. 1. A two agents, two actions, extensive form game with imperfect information. Dotted lines represent information sets. Labels at decision nodes identify the agent that plays, labels on edges agents' actions, and values at leaves agents' payoffs.

one agent plays by choosing among a set of available actions, each action being represented by an edge leading from a node to another. Games in which each agent knows exactly the node in the tree where she plays are games with *perfect* information, otherwise information is *imperfect*. The agents' uncertainty about the state is represented by *information sets* that group the states that cannot be distinguished by the agents. Formally:

Definition 3. *An extensive form game is a tuple* $\Gamma = \langle \mathcal{N}, \mathcal{G}, \{R_i\}, \iota, \{\mathcal{H}_i\}, \{\mathcal{A}_i\}\rangle$*, where:*

- $\mathcal{N}$ *is the set of the agents*
- $\mathcal{G} = \langle \mathcal{S}, s^0, \mathcal{T}\rangle$ *is a finite tree with a set of decision nodes* $\mathcal{S}$*, a unique initial node* $s^0 \in \mathcal{S}$ *and, a set of terminal nodes* $\mathcal{T}$
- $R_i : \mathcal{T} \to \Re$ *is the payoff function for agent* i *that maps each terminal node to a numerical payoff*
- $\iota : \mathcal{S} \to \mathcal{N}$ *is the agent function that maps decision nodes to the agent that plays at that node*
- *let* $\mathcal{H}$ *be the set of information sets* $h \subset \mathcal{S}$ *that partitions the set of decision nodes:* $\mathcal{S} = \bigcup_{h\in\mathcal{H}} h$ *and* $\forall h, h' \in \mathcal{H}, h \cap h' = \emptyset$*;* $\mathcal{H}$ *is partitioned into sets of information sets which belong to the same agent:* $\mathcal{H}_i = \{h \in \mathcal{H}, \forall s \in h, \iota(s) = i\}$
- $\mathcal{A}_i(h)$ *is the set of actions available to agent* $i = \iota(s)$ *in each information set* $h \in \mathcal{H}_i$*, such that* $s \in h$

Unlike normal form games, in the extensive form the strategies are defined as functions of the information set perceived by the agent, i.e., $\mathbf{x}_i(h) = (x_{i1}(h), \ldots, x_{im}(h)), h \in \mathcal{H}_i$. This is due to the fact that agent i may play more than once at different stages of the game. Thus, in the following, we denote by $\mathbf{x}_i$ the functional strategy over the information sets, while the joint strategy $\mathbf{x}$ is called *strategy profile* of the game.

In extensive form games some refinements of the Nash equilibrium are usually adopted as solution concepts. In the following, we focus only on the *sequential equilibrium* of Kreps and Wilson [5], which is the most suitable equilibrium for extensive form games with imperfect information. In fact, the sequential equilibrium takes into account not only the strategies, but also the agents' *beliefs* about

the state of the game. A belief for agent i is defined as a probability distribution $\boldsymbol{\mu}_i(h) = (\mu_{i1}, \ldots, \mu_{ij}, \ldots, \mu_{i|h|})$ over the states in the perceived information set $h \in \mathcal{H}_i$, where μ_{ij} is the probability for agent i to be in the j-th state of h. The set of beliefs $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_i, \ldots, \boldsymbol{\mu}_n)$ is called *system of beliefs*. The expected payoff $\rho_i(\mathbf{x}|\boldsymbol{\mu}_i)$ for agent i, given her belief $\boldsymbol{\mu}_i$ and a joint strategy $\mathbf{x}$, is defined as the expected payoff when the probability to be in the states of her information sets is exactly given by her belief. The system of beliefs together with the strategy profile define an *assessment* $\sigma = \langle \boldsymbol{\mu}, \mathbf{x} \rangle$. A sequential equilibrium is an assessment $\sigma^* = \langle \boldsymbol{\mu}^*, \mathbf{x}^* \rangle$ such that the strategies in $\mathbf{x}^*$ are mutual best responses (*sequential rationality*) and the beliefs in $\boldsymbol{\mu}^*$ are consistent with the probability distribution induced by $\mathbf{x}^*$ on the states of the game (*Bayesian consistency*). Finally, the notion of consistency in the sense of Kreps and Wilson also requires the existence of a sequence of assessments $\sigma_k = \langle \boldsymbol{\mu}_k, \mathbf{x}_k \rangle$, each with fully mixed $\mathbf{x}_k$ and Bayesian consistent $\boldsymbol{\mu}_k$, that converges to σ^*. Technically, this latter condition avoids that beliefs on information sets never visited at the sequential equilibrium remain undetermined. More formally:

Definition 4. *An assessment $\sigma^* = \langle \boldsymbol{\mu}^*, \mathbf{x}^* \rangle$ is a sequential equilibrium of an extensive form game Γ if:*

- (sequential rationality)*:*

$$\rho_i(\mathbf{x}_1^*, \ldots, \mathbf{x}_i^*, \ldots, \mathbf{x}_n^* | \boldsymbol{\mu}_i^*) \geq \rho_i(\mathbf{x}_1^*, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n^* | \boldsymbol{\mu}_i^*), \quad \forall i \in \mathcal{N}. \tag{2}$$

- (Bayesian consistency)*: the joint strategy $\mathbf{x}^*$ induces a probability distribution on states equal to the system of beliefs $\boldsymbol{\mu}^*$*
- (Kreps and Wilson consistency)*: there is a sequence $\sigma_k = \langle \boldsymbol{\mu}_k, \mathbf{x}_k \rangle$, such that*

$$\mathbf{x}_k \to \mathbf{x}^*, k \to \infty \tag{3}$$

being $\mathbf{x}_k$ fully mixed and $\boldsymbol{\mu}_k$ consistent with $\mathbf{x}_k$

2.3 From Extensive Form to Normal Form Games

Sometimes it is convenient to transform a game from its extensive form to a normal form, so as to benefit from the results coming from the normal form representations. The transformation from extensive to normal form can be done as follows. The set of agents $\mathcal{N}$ remains the same. For any agent i, the set of actions A_i in the normal form game contains one action for each possible sequence of choices that the agent takes at decision nodes s such that $\iota(s) = i$. Finally, payoff functions are such that for each joint action the payoff is defined as that obtained at the termination node reached in the extensive form game.

It can be shown [9] that sequential equilibria of the extensive form game are always preserved as Nash equilibria of the normal form game. Nonetheless, other Nash equilibria could be generated, and this may prevent learning algorithms designed for normal form games, that are generically aimed at converging to Nash equilibria, from successfully solving extensive form games.

3 Reinforcement Learning and Q-Learning Dynamics

RL is a learning paradigm that enables an agent to learn the optimal strategy to solve a given task through a trial-and-error process of direct interaction with an unknown environment. At each time instant, the state of the environment evolves in response to the action taken by the agent and a reward is returned. The goal of a reinforcement learning agent is to learn the strategy $\mathbf{x}^*$ that maximizes the rewards through time. More formally, a strategy $\mathbf{x}(s) = (x_1(s), \ldots, x_i(s), \ldots, x_m(s))$ is defined as a mapping from a state s to a probability distribution over actions, where $x_i(s)$ is the probability of taking action i in state s. The quality of a strategy $\mathbf{x}$ can be measured by the action value function $Q^{\mathbf{x}}(s, a)$, defined as the expected sum of discounted rewards obtained by taking action a in state s and following $\mathbf{x}$ thereafter:

$$Q^{\mathbf{x}}(s, a) = E\left[\sum_{k=0}^{\infty} \delta^k r_k | a(0) = a\right]$$

where $\delta \in [0, 1)$ is the discount factor, and r_k is the reward returned at time k. The optimal action value function $Q^*(s, a)$ is defined as the function whose value is maximum in each state-action pair. Learning the optimal strategy $\mathbf{x}^*$ is equivalent to learning the optimal action value function $Q^*(s, a)$. In order to learn $Q^*(s, a)$, the agent needs to explore all possible actions in all the states of the environment. On the other hand, as the learning progresses, in order to assess the performance of her strategy, the agent should exploit the estimation of the action value function by taking in each state the greedy action, i.e., the action whose action value is highest. A common exploration policy is the Boltzmann strategy:

$$x_i(s) = \frac{e^{\tau Q(s,a_i)}}{\sum_{j=1}^{m} e^{\tau Q(s,a_j)}} \tag{4}$$

where τ is the exploitation factor (the lower [higher] τ, the higher [lower] the exploration).

While the agent explores the environment according to Eq. 4, the estimation of the action value function should be updated on the basis of the rewards received by the agent. In Q-learning [13], one of the most used RL update rules, when the agent takes an action a and receives a reward r, the action value function is updated as:

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha\left(r_i + \delta \max_{a'} Q(s', a')\right) \tag{5}$$

where $\alpha \in [0, 1]$ is the learning rate and s' is the state after the execution of a.

In a multiagent context, the environment is populated by n agents, at each time instant k the state evolves according to the joint action $\mathbf{a}(k)$, and the reward for each agent depends on the joint action as well. In the simple case in which the interaction between the agents is described by a normal form game, the environment is characterized by a single state, and the reward r_i is defined by the payoff function $R_i(\mathbf{a})$ (Sec. 2.1). Although Q-learning is guaranteed to find

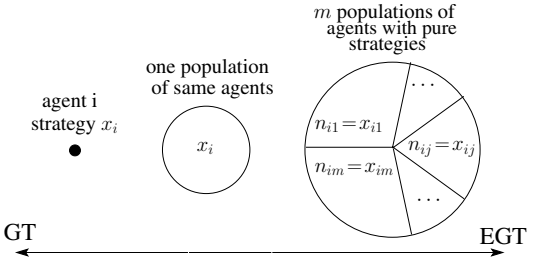


Fig. 2. In the replicator dynamics equations, a learning agent is represented by a set of m populations of identical agents that play pure strategies with proportions such that their densities correspond to the probability to play that strategy

the optimal strategy (that in normal form games corresponds to a Nash equilibrium) in stationary environments under very loose hypotheses [13], in multiagent problems payoffs depend on the joint action and, since agents do not know the other strategies, each agent perceives a non-stationary environment. Thus, the learning process is not guaranteed to converge and may exhibit complex dynamics that are often difficult to study by stochastically simulating single runs of execution.

Evolutionary Game Theory (EGT) is the application of population genetics-inspired models to game theory. With respect to classical game theory, it is more focused on the dynamics of proportions (i.e., the relative abundance or density, also called frequency) of homogeneous populations of agents all playing the same action. As depicted in Fig. 2, agent i can be imagined as a large population of $\mathbf{x}_i$ identical strategists, or, equivalently, as m homogeneous sub-populations of pure strategists, one for each action $a_{ij} \in A_i$ with proportions $n_{ij} = x_{ij}, j = 1, \ldots, m$. Thus, a pure strategist playing action a_{ij} is randomly extracted from the population with the same probability x_{ij} according to which agent i plays that action. Then, assuming that the game is repeatedly played many times in any small time interval dt and that from time to time (but still many times in dt) a pure strategist is randomly extracted from the population and offered the option of switching to the pure strategy of another randomly selected strategist, the continuous-time dynamics of the sub-population proportions or, equivalently, the strategy dynamics, are ruled by the replicator dynamics:

$$\dot{x}_{ij} = x_{ij} \left[(P_i \mathbf{x}_{-i})_i - \mathbf{x}_i P_i \mathbf{x}_{-i} \right] \tag{6}$$

where P_i is the payoff matrix of agent i. In [12], the Q-learning dynamics in normal form games is proved to converge to the following replicator-like dynamics:

$$\dot{x}_{ij} = x_{ij} \alpha \tau \left[(P_i \mathbf{x}_{-i})_i - \mathbf{x}_i \cdot P_i \mathbf{x}_{-i} \right] + x_{ij} \alpha \tau \sum_{k=1}^{m} x_{ik} \ln \left(\frac{x_{ik}}{x_{ij}} \right) \tag{7}$$

where the number of ordinary differential equations (ODEs) for agent i can be reduced to $|A_i| - 1$, since the probabilities $x_{ij}, j = 1 \ldots, n$ sum to 1.

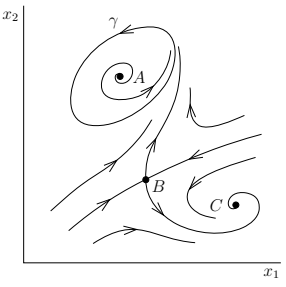


Fig. 3. A skeleton of 13 trajectories for a second-order dynamical system

4 Bifurcation Analysis of Dynamical Systems

In this section we recall the main notions on non-linear dynamical systems, with particular emphasis on structural stability and bifurcations. The following sections are adapted from [2] (Appendix A). We refer the reader to [7] for a more complete treatment of bifurcation theory.

4.1 Dynamical Systems Background

A continuous-time finite-dimensional dynamical system is defined by a system of ordinary differential equations (ODEs):

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)) \tag{8}$$

where the state vector $\mathbf{x}$ is n-dimensional and $\dot{\mathbf{x}}$ is its time derivative. Given the initial state $\mathbf{x}(0)$, the ODEs uniquely define a *trajectory* of the system, i.e., the state vector $\mathbf{x}(t)$ for all $t \geq 0$. Trajectories can be easily obtained through simulation (i.e., numerical integration). The set of all trajectories is the *state portrait* of the system. In Fig. 3 a representative example of 13 trajectories from a two-dimensional (i.e., $n = 2$) system is reported. Three trajectories (A, B, C) are points (corresponding to constant solutions of the system) called *equilibria*, while one (γ) is a closed trajectory (corresponding to a periodic solution of the system) called *limit cycle*. The behavior of the trajectories allow one to conclude that A is a *repellor* (no trajectory starting closing to A tends or remains close to A), B is a *saddle* (almost all trajectories starting close to B go away from B but two trajectories tend to B and compose the so-called *stable manifold*; the two trajectories emanating from B compose the *unstable manifold*) while C and γ are *attractors* (all trajectories starting close to C [γ] tend to C [γ]). Attracting equilibria and cycles are said to be (*asymptotically*) *stable* (*globally stable* if they attract all initial conditions, technically with the exclusion of sets with no measure in state space), while saddles and repellors are *unstable*. The trajectories in Figure 3 also identify the *basin of attraction* of each attractor: in fact all trajectories starting above [below] the stable manifold of the saddle tend toward the limit cycle γ [the equilibrium C].

The study of the stability of equilibria can be done through linearization of the dynamical system at equilibrium points, that is, by approximating the behavior of the system in the vicinity of an equilibrium $\bar{x}$ through the linear system $d/dt(\mathbf{x} - \bar{\mathbf{x}}) = \partial f/\partial \mathbf{x}|_{\mathbf{x}=\bar{\mathbf{x}}}(\mathbf{x} - \bar{\mathbf{x}})$. This way, it is possible to study the stability of $\bar{\mathbf{x}}$ by looking at the eigenvalues $\lambda_i, i = 1, \ldots, n$ of the Jacobian matrix $\partial f/\partial \mathbf{x}|_{\mathbf{x}=\bar{\mathbf{x}}}$. If all the eigenvalues have negative real part then the equilibrium is stable, while if at least one eigenvalue has positive real part the equilibrium is unstable. Similarly, the stability of limit cycles can be analyzed through linearization of the $(n-1)$-dimensional discrete-time dynamical system whose state is defined by the intersections of the system trajectories close to the limit cycle with a given transversal manifold (the so-called Poincaré section). Whenever these intersections converge to the equilibrium at which the cycle intersects the manifolds the limit cycle is stable, otherwise is unstable.

4.2 Structural Stability and Bifurcation Analysis

The goal of the structural stability analysis is the study of the asymptotic behavior of parametrized families of dynamical systems of the form:

$$\dot{\mathbf{x}} = f(\mathbf{x}(t), \mathbf{p}) \tag{9}$$

where $\mathbf{p}$ is a vector of *parameters*.

If a parameter is slightly perturbed, by continuity the position and form of the asymptotic behaviors of trajectories, namely attractors, saddles, and repellors, are smoothly affected (e.g., an equilibrium might slightly move or a limit cycle might become slightly bigger or faster), but all trajectories remain topologically the same (e.g., stable equilibria and cycles remain attractive). In regions of the domain of $\mathbf{p}$ in which this continuity holds, the system is *structurally stable*. The above continuity argument fails at particular parameter values called *bifurcation points* [7], which correspond in state space to collisions of attractors, saddles, and repellors. Thus, the robustness of the dynamical characteristics of the system, as summarized by the state portrait, depends on how far the parameters are from bifurcation points. A thorough robustness investigation therefore requires to produce the catalog of all possible modes of behavior of the system family, i.e., its complete bifurcation analysis. An exhaustive review of bifurcation theory is certainly beyond the scope of this paper. In the following, we focus on three types of bifurcations involving a single parameter p that are relevant for the case study in Sec. 4: *saddle-node*, *Hopf*, and *homoclinic* bifurcations.

The saddle-node bifurcation corresponds to the collision, at a critical value p^*, of two equilibria: a stable node N (i.e., a stable equilibrium characterized by real eigenvalues of the linearized system) and, in its simplest two-dimensional formulation, a saddle S (Fig- 4-top). For $p < p^*$, N has two negative eigenvalues, while the eigenvalues of S are one positive and one negative. For $p > p^*$ no equilibrium is present, so that at the bifurcation point $p = p^*$ the largest eigenvalue of N and the smallest eigenvalue of S both vanish. In short, a saddle-node bifurcation can be identified by the change of sign of one of the eigenvalues of an

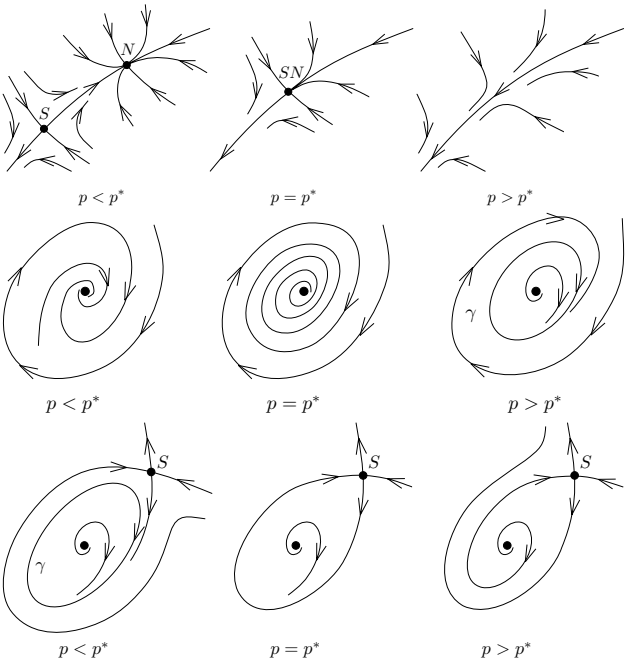


Fig. 4. Example of saddle-node (top), Hopf (mid) and homoclinic (bottom) bifurcations in a two-dimensional dynamical system

equilibrium when a parameter p is varied. Geometrically, two equilibria, which are non necessarily a node and a saddle in higher-dimensional systems, collide and disappear as p crosses the bifurcation.

The second type of bifurcation, the Hopf bifurcation, involves the appearance of limit cycles. With reference to two-dimensional systems, if a *focus*, that is an equilibrium with two complex conjugate eigenvalues, is stable for $p < p^*$ and becomes unstable at $p = p^*$, then a stable limit cycle γ may appear for $p > p^*$ (the so-called supercritical case, see Figure 4-mid). But the cycle may also be unstable and surround the stable focus for $p < p^*$ (subcritical case) and the distinction between the two cases depends upon the nonlinear terms in the expansion of the function f in Eq. 8 and is typically implemented in software packages for numerical bifurcation analysis [3,7]. In both cases, however, the cycle is small for p close to p^*, so that the Hopf bifurcation can be seen geometrically as the collision of a vanishing cycle with a focus. Moreover, the pair of complex conjugate eigenvalues cross the imaginary axis of the complex plane at the bifurcation, thus changing the stability of the equilibrium.

Finally, the homoclinic bifurcation is characterized by the collision of a limit cycle and a saddle (Figure 4-bottom). When p approaches p^*, the cycle γ gets closer to saddle S, so that the period of the cycle diverges, since the state of the system moves very slowly when close to S. At the bifurcation ($p = p^*$) the cycle

touches the saddle and collides with its stable and unstable manifolds which coincide at the bifurcation. The identification of a homoclinic bifurcation cannot rely on the analysis of eigenvalues but involve the global behavior of the system. For this reason, the homoclinic bifurcation is classified as a *global* bifurcation, in contrast with *local* bifurcations that can be detected through eigenvalue analysis.

Whenever a perturbation of the parameter from p to $p + \Delta$ ($p < p^* < p + \Delta$) triggers a transient toward a macroscopically different asymptotic regime (i.e., a different attractor), the bifurcation at p^* is called *catastrophic*. By contrast, if the catastrophic transition is not possible, the bifurcation is called *noncatastrophic*.

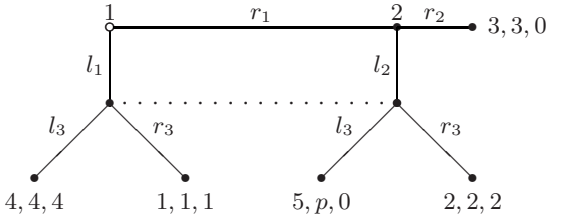
Although one might hope to detect bifurcations by simulating the system for various parameter settings and initial conditions, saddles, which have a fundamental role in bifurcation analysis, cannot be studied just through simulation. In fact, any small approximation introduced by the numerical scheme of integration would lead to trajectories that miss the saddle and go away from it along its unstable manifold. Moreover, the "brute force" simulation approach is never effective and accurate in practice, since bifurcations are often related to a loss of stability of equilibria and cycles, so that the length of simulations need to be dramatically increased while approaching the bifurcation. This is why the proper tools for numerical bifurcation analysis are based on continuation (see [7], Chap.10 and [3]), a simulation-free numerical method which locates bifurcations by continuing equilibria and cycles in parameter space, that is by studying their position in the state space when the parameter is changed.

5 Bifurcation Analysis on the Selten's Horse Game

In the following, we illustrate the results of the bifurcation analysis on the Selten's horse game [4] (named from its inventor and from the shape of its tree). This game, commonly adopted in GT for the study of sequential equilibria, is particularly suitable for our analysis because *(i)* it involves more than two agents, *(ii)* it is an extensive form game and *(iii)* one agent has imperfect information about the state of the game. All these factors of complexity lead to the definition of a complex learning system exhibiting interesting dynamics. At the same time, the game is simple enough to allow an intuitive analysis of its dynamics and a detailed bifurcation analysis on one of the payoffs.

5.1 Learning Dynamics in the Selten's Horse Game

The Selten's horse game [4] (Fig. 5-left) is an extensive form game with imperfect information involving three agents with two actions each ($A_i = \{l_i, r_i\}, i = 1, 2, 3$). While both agents 1 and 2 have perfect information about the state of the game, agent 3 cannot distinguish the state in which it plays (dotted line in the figure), that is, she is characterized by a single information set containing the two decision nodes where she plays. According to Definition 4, the game has a unique sequential equilibrium strategy (r_1, r_2, r_3) (we omit the derivation for lack of space). On the other hand, as one can easily verify, in the equivalent normal



l_1	l_2	r_2
l_3	4,4,4	**4,4,4**
r_3	1,1,1	1,1,1

r_1	l_2	r_2
l_3	5,p,0	3,3,0
r_3	2,2,2	**3,3,0**

Fig. 5. Selten's Horse game and payoff tables of the equivalent normal form game. Parameter p is set to 5 in the original configuration of the game.

form game (Fig. 5) there are two Nash equilibria (r_1, r_2, r_3) and (l_1, r_2, l_3) (in bold in payoff tables), while there are no mixed equilibria. The joint strategy for this game is the vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, where $\mathbf{x}_i = (x_{i1}, x_{i2})$ (being $x_{i2} = 1 - x_{i1}$) is the strategy of agent i. Strategy $\mathbf{x}_i = (1, 0)$ corresponds to action l_i, while $\mathbf{x}_i = (0, 1)$ is action r_i.

In the following, we analyze the dynamics of the strategies when all the agents learn through Q-learning. As discussed in Section 3, for the study of the learning process, we derive the dynamical system defined by the replicator dynamics for x_{11}, x_{21} and x_{31} as defined in Eq. 7, where

$$P_1 = \begin{bmatrix} 4\,1\,4\,1 \\ 5\,2\,3\,3 \end{bmatrix} \quad P_2 = \begin{bmatrix} 4\,1\,p\,2 \\ 4\,1\,3\,3 \end{bmatrix} \quad P_3 = \begin{bmatrix} 4\,4\,0\,0 \\ 1\,1\,2\,0 \end{bmatrix}$$

are the payoff matrices, where p is equal to 5 in the original setting.

The replicator dynamics is bounded in the open 3-dimensional cube of the state space (x_{11}, x_{21}, x_{31}) (the other three variables can be eliminated as $x_{i2} = 1 - x_{i1}$). The faces of the cube cannot be reached due to the exploration logarithmic terms in Eq. 7. The sequential equilibrium (r_1, r_2, r_3) corresponds to point $(0, 0, 0)$ in the state space, while the other Nash equilibrium is $(1, 0, 1)$.

Let us first consider the dynamics of the system in its original settings (panel 5 in Fig. 7). Although the equivalent normal form game has two Nash equilibria, the learning process converges to a point close to the sequential equilibrium. Starting from any initial joint strategy, the trajectories of the system reach a globally stable equilibrium close to the joint strategy (r_1, r_2, r_3). This would be the expected solution from a game theoretical perspective, since (r_1, r_2, r_3) is the unique sequential equilibrium in the game. At the opposite, the learning dynamics continues and, because of the residual exploration of the agents converges to a different equilibrium point. By exploring action l_2, agent 2 allows agent 3 to play r_3 and obtain a payoff greater than 0 (the payoff of agent 3 at the sequential equilibrium). Then, agent 3 takes advantage by mixing her strategy toward l_3, since this tempts agent 2 to play l_2 more frequently. In fact, the payoff of agent 2 for action l_2 is a weighted average between 5 and 2 depending on the agent 3 mixed strategy (Fig. 5), and the result can be greater than 3 (the equilibrium payoff). This is possible because of the payoff for agent 2 in (r_1, l_2, l_3) is sufficiently high, while this scenario is likely to change for lower values of the

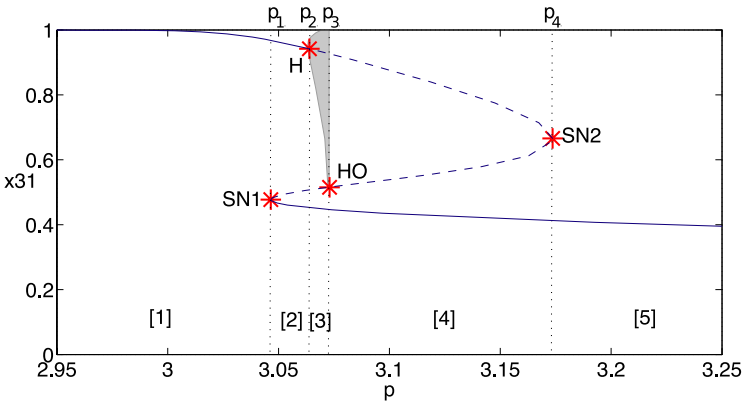


Fig. 6. Bifurcation diagram of the learning system. The curve identifies the number of equilibria and the value of their x_{31} component for each value of p in the considered range. The shaded area represents the x_{31} excursion along a family of stable limit cycles present in interval 3. Vertical dotted lines indicate bifurcation points (SN1,SN2: saddle-node, HO: homoclinic; H: Hopf). Dashed and solid parts of equilibrium curve distinguish unstable and stable equilibria (e.g., one stable and two unstable equilibria characterized by increasing value of x_{31} are present in interval 4). Parameter values: $\alpha = 0.3, \tau = 0.6$. Diagram obtained through the software package Matcont [3].

payoff. This means that, from this preliminary analysis, the system is expected to preserve structural stability only for a limited range of values of the payoff (in the following parameter p) and to possibly show complex dynamics otherwise. This is the reason for the following bifurcation analysis with respect to p.

5.2 Bifurcation Analysis

The bifurcation analysis of the learning system is reported in Fig. 6 (for variable x_{31}) and Tab. 1 and identifies five qualitatively different learning dynamics corresponding to five intervals of p (the learning dynamics are shown in Fig. 7). As discussed in Section 4, the identification of the bifurcations cannot be done by simulating the system for different values of p. A complete bifurcation analysis needs the *continuation* of equilibria and cycles in parameter space and the identification of the parameter values in which the system looses its structural stability.

The analysis starts with $p = 5$, the value in the original game setting, at which, as already discussed in Sec. 5.1, the system is characterized by a globally stable equilibrium close to $(0, 0, 0)$. The numerical continuation of this equilibrium with respect to p allows to track its position for different parameter values and produces the curve in Fig. 6. By decreasing p from its original value, the equilibrium moves away from $(0, 0, 0)$. In fact, in order to tempt agent 2 to play l_2, agent 3 is forced to mix her strategy more and more toward l_3, but so doing she vanishes her own return. Similar considerations can be made for x_{11} and x_{21}. Further

Table 1. Bifurcation analysis of the replicator dynamics with respect to parameter p

Parameter	Bifurcation
$p_1 = 3.04651$	SN1
$p_2 = 3.06392$	H
$p_3 = 3.07235$	HO
$p_4 = 3.17361$	SN2

Interval	Equilibria	Limit Cycles
[1] $p < p_1$	1 globally stable	-
[2] $p_1 \leq p < p_2$	2 stable, 1 saddle	-
[3] $p_2 \leq p < p_3$	1 stable, 2 saddles	1 stable
[4] $p_3 \leq p < p_4$	1 globally stable, 2 saddles	-
[5] $p \geq p_4$	1 globally stable	-

reductions of p are less easy to interpret on an intuitive ground, also because a different mix of Nash/sequential pure/mixed equilibria might arise, and this is indicative of impending dynamical complexity. In fact, the first bifurcation is encountered at $p{=}p_1$ (SN1), a saddle-node bifurcation at which the equilibrium collides with a saddle and they both disappear for $p < p_1$. Notice, however, that the equilibrium is not globally stable in intervals 2 and 3, since three more bifurcations occur for $p_1 < p < 5$, but involve other equilibria of the system and are therefore initially unnoticed by the local continuation of the equilibrium. The continuation direction reverts at a saddle-node bifurcation, so that we now continue the saddle for increasing values of p. The first encountered bifurcation is another saddle-node (SN2) at $p{=}p_4$, approaching which one of the two stable eigenvalues of the saddle vanishes, as well as one of the two unstable eigenvalues of another saddle, characterized by only one stable eigenvalue. The two saddles collide at the bifurcation and do not exist for $p > p_4$, while the continuation proceeds by tracking the new saddle for decreasing values of p. The two unstable eigenvalues are real close to p_4, but become complex (saddle-focus) somewhere before the Hopf bifurcation (H) detected at $p{=}p_2$. The Hopf is supercritical, so that a family of stable limit cycles can be continued for increasing values of p starting from $p{=}p_2$ (shaded area in Fig. 6). The saddle-focus becomes stable by crossing the bifurcation and its continuation to lower values of p does not point out new losses of structural stability. Moreover, the equilibrium becomes globally stable for $p < p_1$ and x_{31} approaches 1 as p is further reduced. At the same time, x_{11} and x_{21} approach 1 and 0, respectively, so that the equilibrium approaches $(1, 0, 1)$, i.e., the other Nash equilibrium of the original game. In particular, it is easy to verify that $(1, 0, 1)$ is Nash for all p and can be shown to be the only sequential equilibrium for p sufficiently small. Finally, increasing p from $p{=}p_2$, the family of limit cycles is interrupted by an homoclinic bifurcation (HO) at $p{=}p_3$, where the cycle gets in contact with the saddle originated at (SN1).

All together, the bifurcation analysis shows that the learning dynamics are dominated by two sequential equilibria, $(0, 0, 0)$ for large values of p and $(1, 0, 1)$ for small values of p, in the sense that close to them there is an equilibrium of the learning dynamics (the lower [upper] equilibrium in Fig. 6 for large [small] p) which attracts all initial conditions (see intervals 1 and 5 and in particular interval 4 where, though the presence of two saddles, all trajectories, except those composing the saddle stable manifolds, converge to the stable equilibrium). The switch from one equilibria to the other as p is varied involves two *catastrophes*: the homoclinic bifurcation for increasing values of p (HO) and the saddle-node

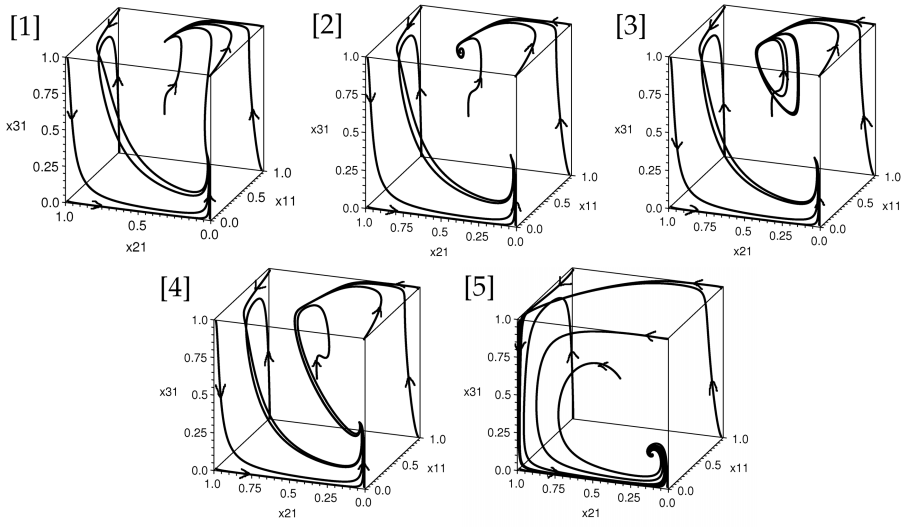


Fig. 7. Learning dynamics for five different values of p $(3.0, 3.05, 3.07, 3.1, 5)$

(SN1) for decreasing values of p. In particular, in the first case, the learning dynamics follow the family of limit cycles. The period of the cycle diverges as the bifurcation is approached and the joint strategy remains for most of the time very close to the saddle, at the point that finite-time simulations can erroneously reveal convergence to a stable equilibrium. Crossing the bifurcation, the cycle suddenly disappears and the dynamics converge to the lower equilibrium.

Finally notice that in intervals 2 and 3 the system has two alternative attractors, two equilibria in 2 and an equilibrium and a cycle in 3. The attractor which is reached by the learning process depends on the initial joint strategy and the saddle (actually its stable manifold) delimits the two attraction basins.

6 Conclusions and Future Works

In this paper we applied bifurcation analysis to the study of Q-learning multiagent dynamics in the continuous-time limit provided by the replicator dynamics of evolutionary game theory. A preliminary one-parameter analysis of the Selten's Horse game is presented as a case study. The first result of the analysis is that in extensive form games with imperfect information Q-learning may exhibit complex learning dynamics, including multiple stable equilibria and periodic non-convergent attractors. Furthermore, the analysis pointed out that Q-learning is not robust to payoff perturbations and that the corresponding dynamical system looses stability in four different bifurcation points. In particular, at the two catastrophic bifurcations, small variations of the payoff correspond to radically different asymptotic regimes, thus leading the three agents to significantly change their strategies. In general, we showed that bifurcation analysis

can be an effective way to study the structural stability of learning systems and that it could also be used to compare the robustness of different learning algorithms.

Although the bifurcation analysis presented in the paper focused on a structural parameter (i.e., a payoff), the same analysis can be carried out when learning parameters are varied, and this could lead to useful suggestions about parameter settings. A preliminary joint analysis with respect to the payoff p and the exploitation factor τ showed that for low values of τ (high exploration) the bifurcation points disappear and the system is globally structurally stable, while for high values of τ (low exploration) the system becomes more robust to payoff perturbations as the regions of structural stability become larger.

In general, we believe that this novel, though preliminary, analysis opens interesting scenarios for a more complete investigation of the dynamics of multiagent learning systems. Future efforts will be devoted to: *(i)* the development of a replicator dynamics model more compliant to learning algorithms (e.g., decreasing learning rates and exploration factors), *(ii)* two parameters bifurcation analysis (e.g., a joint analysis with respect to learning and structural parameters), *(iii)* study of more complex games (e.g., signaling game, bilateral negotiations, auctions).

References

1. Börgers, T., Sarin, R.: Learning through reinforcement and replicator dynamics. Journal of Economic Theory 77(1), 1–14 (1997)
2. Dercole, F., Rinaldi, S.: Analysis of Evolutionary Processes: The Adaptive Dynamics Approach and its Applications. Princeton University Press, Princeton, NJ, (forthcoming)
3. Dhooge, A., Govaerts, W., Kuznetsov, Y.A.: MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs. ACM Trans. Math. Software 29, 141–164 (2002)
4. Gintis, H.: Game Theory Evolving. Princeton University Press, Princeton, NJ (2000)
5. Kreps, D.M., Wilson, R.: Sequential equilibria. Econometrica 50(4), 863–894 (1982)
6. Kunigami, M., Terano, T.: Connected replicator dynamics and their control in a learning multi-agent system. In: IDEAL, pp. 18–26 (2003)
7. Kuznetsov, Y.A.: Elements of Applied Bifurcation Theory. 3rd edition (2004)
8. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: ICML, pp. 157–163. New Brunswick, NJ, Morgan Kaufmann, San Francisco (1994)
9. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press, Cambridge (1991)
10. Sato, Y., Crutchfield, J.P.: Coupled replicator equations for the dynamics of learning in multiagent systems. Phys. Rev. E 67(1), 15206 (2003)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
12. Tuyls, K., Hoen, P.J., Vanschoenwinkel, B.: An evolutionary dynamical analysis of multi-agent learning in iterated games. JAAMAS 12(1), 115–153 (2006)
13. Watkins, C.J., Dayan, P.: Q-learning. Machine Learning 8, 279–292 (1992)