UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Département d'écologie et évolution

EVOLUTION OF DECISION-MAKING AND LEARNING UNDER FLUCTUATING SOCIAL ENVIRONMENTS

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine

de l'Université de Lausanne

par

Slimane DRIDI

Master en Sciences Cognitives de l'Université de Grenoble

Jury

Prof. Ted Farmer, Président

Prof. Laurent Lehmann, Directeur de thèse

Prof. Alexandre Roulin, Co-directeur

Dr. Jean-Baptiste André, Expert

Prof. Tadeusz Kawecki, Expert

Lausanne 2013

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur	Prof. Edward E. **Farmer**
Directeur de thèse	Monsieur	Prof. Laurent **Lehmann**
Co-directeur de thèse	Monsieur	Prof. Alexandre **Roulin**
Experts	Monsieur	Prof. Tadeusz **Kawecki**
	Monsieur	Dr Jean-Baptise **Andre**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Slimane Dridi

Master en Sciences cognitives de l'Université de Grenoble, France

intitulée

EVOLUTION OF DECISION-MAKING AND LEARNING UNDER FLUCTUATING SOCIAL ENVIRONMENTS

Lausanne, le 7 février 2014

pour Le Doyen
de la Faculté de Biologie et de Médecine

Edward E. Farmer

Prof. Edward E. Farmer

Year : 2014

Evolution of decision-making and learning under fluctuating social environments

Slimane Dridi

Abstract

The capacity to learn to associate sensory perceptions with appropriate motor actions underlies the success of many animal species, from insects to humans. The evolutionary significance of learning has long been a subject of interest for evolutionary biologists who emphasize the benefit yielded by learning under changing environmental conditions, where it is required to flexibly switch from one behavior to another. However, two unsolved questions are particularly important for improving our knowledge of the evolutionary advantages provided by learning, and are addressed in the present work. First, because it is possible to learn the wrong behavior when a task is too complex, the learning rules and their underlying psychological characteristics that generate truly adaptive behavior must be identified with greater precision, and must be linked to the specific ecological problems faced by each species. A framework for predicting behavior from the definition of a learning rule is developed here. Learning rules capture cognitive features such as the tendency to explore, or the ability to infer rewards associated to unchosen actions. It is shown that these features interact in a non-intuitive way to generate adaptive behavior in social interactions where individuals affect each other's fitness. Such behavioral predictions are used in an evolutionary model to demonstrate that, surprisingly, simple trial-and-error learning is not always outcompeted by more computationally demanding inference-based learning, when population members interact in pairwise social interactions. A second question in the evolution of learning is its link with and relative advantage compared to other simpler forms of phenotypic plasticity. After providing a conceptual clarification on the distinction between genetically determined vs. learned responses to environmental stimuli, a new factor in the evolution of learning is proposed: environmental complexity. A simple mathematical model shows that a measure of environmental complexity, the number of possible stimuli in one's environment, is critical for the evolution of learning. In conclusion, this work opens roads for modeling interactions between evolving species and their environment in order to predict how natural selection shapes animals' cognitive abilities.

Résumé

La capacité d'apprendre à associer des sensations perceptives à des actions motrices appropriées est sous-jacente au succès évolutif de nombreuses espèces, depuis les insectes jusqu'aux êtres humains. L'importance évolutive de l'apprentissage est depuis longtemps un sujet d'intérêt pour les biologistes de l'évolution, et ces derniers mettent l'accent sur le bénéfice de l'apprentissage lorsque les conditions environnementales sont changeantes, car dans ce cas il est nécessaire de passer de manière flexible d'un comportement à l'autre. Cependant, deux questions non résolues sont importantes afin d'améliorer notre savoir quant aux avantages évolutifs procurés par l'apprentissage. Premièrement, puisqu'il est possible d'apprendre un comportement incorrect quand une tâche est trop complexe, les règles d'apprentissage qui permettent d'atteindre un comportement réellement adaptatif doivent être identifiées avec une plus grande précision, et doivent être mises en relation avec les problèmes écologiques spécifiques rencontrés par chaque espèce. Un cadre théorique ayant pour but de prédire le comportement à partir de la définition d'une règle d'apprentissage est développé ici. Il est démontré que les caractéristiques cognitives, telles que la tendance à explorer ou la capacité d'inférer les récompenses liées à des actions non expérimentées, interagissent de manière non-intuitive dans les interactions sociales pour produire des comportements adaptatifs. Ces prédictions comportementales sont utilisées dans un modèle évolutif afin de démontrer que, de manière surprenante, l'apprentissage simple par essai-et-erreur n'est pas toujours battu par l'apprentissage basé sur l'inférence qui est pourtant plus exigeant en puissance de calcul, lorsque les membres d'une population interagissent socialement par pair. Une deuxième question quant à l'évolution de l'apprentissage concerne son lien et son avantage relatif vis-à-vis d'autres formes plus simples de plasticité phénotypique. Après avoir clarifié la distinction entre réponses aux stimuli génétiquement déterminées ou apprises, un nouveau facteur favorisant l'évolution de l'apprentissage est proposé : la complexité environnementale. Un modèle mathématique permet de montrer qu'une mesure de la complexité environnementale – le nombre de stimuli rencontrés dans l'environnement – a un rôle fondamental pour l'évolution de l'apprentissage. En conclusion, ce travail ouvre de nombreuses perspectives quant à la modélisation des interactions entre les espèces en évolution et leur environnement, dans le but de comprendre comment la sélection naturelle façonne les capacités cognitives des animaux.

Acknowledgements

My first thanks go to my parents, Hedia and Mekki Dridi. They gave me all the love and trust, and I cannot thank them enough for this. My brother, Chems, who always encouraged me in following an ambitious path, and my sisters Chiraz and Myriam, whose visits I really enjoyed, were invaluable supports. I would like to thank my supervisor Laurent Lehmann for accepting my application for this thesis despite my non-biological background. I would also like to acknowledge his availability, in all circumstances, to answer all the mathematical and biological questions I had. I learned a lot from him.

I am grateful to Christian Graff and Gwenaël Kaminski for introducing me to the biology of behavior, and giving me the taste for evolutionary thinking, in an environment (psychology studies) where people are often reticent to natural selection and Darwinian arguments.

Thanks to my office mates (in order of appearance): Erica, Sam, Charles, Simon, Jorge, Matthias, and Mauricio. I especially thank Simon Powers for reviewing some parts of this thesis. Thanks also to Miguel, Manuel, Alberto, Pierre. I enjoyed learning from, and arguing with, you all on scientific and less scientific subjects. The combination of your different backgrounds helped me broaden my vision of biology.

Finally, thanks to Jalal, Hussein, Mohamed, Ilies, Bubaccar, Abdussalam, Mohamed Anas, Yacine (and many others from the GMU) with whom I shared enriching times, or helped me in various ways during my adventures in Switzerland. Thanks to my friend Alaa also, who never stopped asking about the status of this thesis. Our hikes were refreshing and our philosophical arguments a great training for the mind.

Publication arising from this thesis

The following article constitutes the content of Chapter 2:

Dridi, S., and L. Lehmann. 2013. On learning dynamics underlying the evolution of learning rules. Theoretical Population Biology (in press). doi:10.1016/j.tpb.2013.09.003.

Contents

Chapter 1

Introduction

1.1 Learning as a form of phenotypic plasticity

This thesis addresses the problem of the adaptive value of learning as it pertains to animal behavior. Learning, defined as the change in behavior in reaction to environmental information, is widespread across biological species. Animals learn to search for food, learn to find mating partners, or learn to socially interact with conspecifics (Dickinson, 1980; Shettleworth, 2009; Dugatkin, 2010). Seen as a process of adaptation during lifespan to external environmental conditions, learning is just a particular form of phenotypic plasticity (Pigliucci, 2001). Phenotypic plasticity concerns not only animals but all biological organisms. Mameli and Bateson (2006) convincingly claim that it is hard to find a biological trait whose expression is not somehow affected by the environment. Taken as a form of plasticity, learning does not need any particular evolutionary justification, because the selective advantage of plasticity over purely genetically determined phenotypes is already a well-established fact in evolutionary biology (West-Eberhard, 1989; Gavrilets and Scheiner, 1993; Pigliucci, 2005) .

Why then would this thesis focus on the adaptive value of learning? The reason is that there are actually different *levels of plasticity*, and each level of plasticity is adapted to different environmental conditions. Learning is one level of plasticity, and as such it is adapted to specific environmental conditions. In order to explain the concept of levels of plasticity, consider the reaction norm of an organism. The reaction norm of an organism of a given genotype is the function that describes the mapping from environmental conditions (which can be depicted on the x-axis of a two-dimensional plot) to phenotype (the y-axis). When describing, for example,

the plastic phenotype of a plant in relation to light, drawing a reaction norm function seems relatively easy: the x-axis would represent the quantity of light under which growth takes place, and the y-axis would represent the size of the plant. It is more difficult to draw the reaction norm corresponding to learning. What should one put on the x-axis? Learning refers to the change of response to stimuli as a function of experience, so one would actually need to draw several reaction norm functions (with the "quantity" of stimulus on the x-axis and the behavioral response on the y-axis), one for each time point during the learning experience. In other words, the "environment" in the description of the reaction norm of learning comprises the time dimension as a fundamental one, while it is not the case for other simpler norms of reaction. In animal behavior, an example of a norm of reaction that is simpler (or at a lower level) than learning is what are usually called innate behaviors, which correspond to the fact that the response to a particular stimulus is genetically determined and does not change as a function of experience. Since purely innate responses to stimuli are always somehow modified by experience, it might be preferable to call these behaviors "predispositions" or innate tendencies. In this introduction the phrase "innate behavior" will be kept, but it should be remembered that it does not mean that the interaction between innateness and learning in shaping ontogeny of behavior (Mery and Kawecki, 2004) is denied here (see also Shettleworth, 2009 on the use and misuse of the expression "innate behavior"). Under what environmental conditions does natural selection favor learning over innate behaviors?

1.2 Environmental conditions favoring learning

Without acknowledging explicitly the above distinction between different forms or levels of plasticity, students of behavioral ecology have tried to find the ecological circumstances favoring a learning ability over innate behaviors (Boyd and Richerson, 1988; Feldman et al., 1996; Wakano et al., 2004; Borenstein et al., 2008). Their conclusion is that learning is selected for under temporally variable environments. More precisely, there is a Goldilocks principle (Stephens, 1991; Kerr and Feldman, 2003; Dunlap and Stephens, 2009): for learning to provide a selective advantage, the environment should change between generations (so that it is pointless to express the behaviors genetically inherited from parents), but should not change too fast within a generation, so that individuals have time to learn before the consequences of actions change again. While they provide much insight, these models contain some conceptual and theoretical problems. The most important theoretical problem has to do with the assumption of optimality of

the learned actions. Indeed, in most models of the evolution of learning, one generally assumes that learners reach the optimal behavior in every type of new environment (with some cost due to, e.g. errors during exploration or augmented risk of predation, Boyd and Richerson, 1988; Feldman et al., 1996; Wakano et al., 2004; Borenstein et al., 2008).

However, empirical findings show that animals may learn the wrong behavior in complex tasks (Shettleworth, 2009). In fact, different tasks have different solutions and may require different uses of information. Animals that are learning vary in their tendencies to use different quantities and types of information, have different memory capacities, and employ different amounts of exploration. These different features lead to different behavior and have been captured under the idea that learning and, more generally, decision-making mechanisms have been shaped by natural selection to solve the specific set of ecological problems that each species must face (Hammerstein and Stevens, 2012). Hence, more work is needed in the direction of understanding the specificities of learning systems in relation to environmental patterns. The conditions under which learning generates truly adaptive behavior must be identified in a more precise way.

1.3 Psychological mechanisms underlying learning and ecological optimality

Because species vary in their ecological conditions, their psychological and cognitive needs vary as well. It is reasonable to think that natural selection should favor cognitive abilities that allow individuals to perform adapted actions under their species' own ecology. Accordingly, there is a long and very fruitful tradition in behavioral ecology of designing separate optimality models for each type of decision problem animals may face (Charnov, 1976; Mcnamara and Houston, 1985; Shettleworth et al., 1988; Krebs et al., 1993; Houston and McNamara, 1999) . However, taking into account cognitive economy, this is probably not what happens in reality when natural selection acts to favor individuals who express the fittest phenotypes. Rather than having a set of separate "modules", each one dedicated to solve a particular problem (e.g. one for foraging, another one for fighting, another one for social cooperation, etc.), it is more likely that general principles of decision-making that work well on average are implemented inside animals' brains to allow finding approximate solutions under a variety of circumstances (McNamara and Houston, 2009; Dijker, 2011; Fawcett et al., 2013; Lotem, 2013). Learning is one such general principle and, as explained above, the specific details of how an animal learns about its surrounding environment are critical for the animal to perform appropriate behavior.

1.4 Learning rules

The way in which an animal gathers and uses environmental information in the course of learning has been coined under the expression of "learning rule". Learning rules are conceptual ways of describing the probability distribution over actions of an animal at a given decision step depending on the combination of previous experience and newly gathered information. An example of a learning rule that captures the law of effect (Thorndike, 1911; Herrnstein, 1970) is the so-called "linear operator" rule (Bush and Mostelller, 1951; McNamara and Houston, 1987; Bernstein et al., 1988; Stephens and Clements, 1998) that stipulates that the probability to take an action at a given time step is a linear combination of the previous probability and the reward received at the current time step.

The behavioral dynamics of the linear operator rule and of various other learning rules have been studied by game theorists in the context of human social decision-making (Jordan, 1991; Erev and Roth, 1998; Fudenberg and Levine, 1998; Camerer and Ho, 1999; Hopkins, 2002; Hofbauer and Sandholm, 2002; Young, 2004; Sandholm, 2011). As a consequence of the focus on humans, some of these rules probably rely on excessively high cognitive demands (Hart and Mas-Colell, 2000; Foster and Young, 2003) to be applied in the field of animal behavior. But many other rules are in fact perfectly applicable to animal species because they are mainly based on the simple and universal principle of the law of effect (or reinforcement learning): rewarded actions (i.e. actions associated with positive payoffs, increasing biological fitness) tend to be repeated, and punished actions (negative payoffs decreasing fitness) tend to be avoided.

Biologists and psychologists also proposed possible learning rules that may describe animals' behavior (Rescorla and Wagner, 1972; Niv, 2009) but only few studies have considered to what extent the behavior produced by these rules provides a fitness advantage in natural environments. Some recent work takes this approach (Groß et al., 2008; Josephson, 2008; Hamblin and Giraldeau, 2009; Arbilly et al., 2010, 2011*a*,*b*; Katsnelson et al., 2011) and there is a need to pursue this avenue for a better integration of realistic accounts of psychological learning mechanisms and evolutionary arguments.

1.5 Evolution of learning rules for social interactions

Because learning rules in economics and game theory were primarily developed to understand the behavior of humans in social interactions, it is tempting to apply these results to

the field of evolutionary game theory, i.e. the study of natural selection on social behaviors (Maynard-Smith, 1982; Hofbauer and Sigmund, 1998). Indeed, it was claimed above that general decision-making mechanisms and learning rules must work in a wide range of situations, but when considering social situations, many complications arise. In biology, when population members interact and affect each other's fitness (which is the case for a large fraction of the world's biodiversity), there is frequency dependence, i.e. the performance of a strategy depends on the composition of the population.

A famous example is given by the evolution of conflict (Maynard-Smith and Price, 1973; McElreath and Boyd, 2007), captured by the "Hawk-Dove" game in which it is predicted that natural selection leads to the coexistence in the population of aggressive ("Hawks") and non-aggressive individuals ("Doves"). In the classical Hawk-Dove game, frequency dependence just leads to polymorphism in strategies of conflict, but the frequency dependence due to individuals having different learning rules interacting in a population may generate very complex stochastic processes (Lahkar and Seymour, 2013) that are difficult to analyze. Thus, while it is natural to ask what would be the evolutionary outcome of the Hawk-Dove game under the more realistic assumption that individuals learn to play this game rather than having innate strategies, the answer is far from being easy.

More generally in evolutionary game theory, the traditional assumption of behavioral ecology is almost always used. Namely, one assumes that natural selection shapes organisms to have innate strategies for each social game individuals are facing. Another important example is the paradox of cooperation that is also studied under this simplifying assumption, even if a larger set of strategies has been considered by including, for instance, automata strategies like Tit-for-tat or Win-stay-lose-shift (Axelrod and Hamilton, 1981; Nowak, 2006). But these strategies are still specifically designed for cooperation problems, e.g. captured by the Prisoner's Dilemma. Social games may involve many interactants and complex reward dependence, and it is crucial to understand how general learning rules can be applied in these situations, because after all, individual decision problems (e.g. foraging, nest building) are only special cases of social decision problems (in terms of mathematical optimization, Luce and Raiffa, 1989). Understanding sociality might be more ambitious than understanding individual decision tasks, but the former imply the latter, and biology would gain by continuing to build on game theory and economics to achieve this long standing goal.

1.6 Thesis outline

The three chapters presented in this thesis are formatted as scientific articles. Chapter 2 is a published paper (Dridi and Lehmann, 2013), Chapter 3 is a manuscript ready for submission, and Chapter 4 is another manuscript.

An important part of this thesis (Chapter 2) is dedicated to predicting how different learning rules can generate different behaviors in social interactions. Since a learning rule refers to how neurons process information in an animal's brain, it has proven difficult to empirically test whether a given rule describes accurately the psychological characteristics of animals or humans (for humans, see Camerer, 2003 and references therein; for animals, an example is given by Herrnstein, 1970). A first goal of Chapter 2 is then to give an intuition about the relations between the psychological characteristics underlying a learning rule with produced behavior. Another goal of Chapter 2 is to provide a basis in evolutionary theory for building models of the evolution of learning rules: knowing what behavior is produced by the learning rules, one can construct models of natural selection in order to find the learning rules that generate truly adaptive behavior.

In Chapter 3, the behavioral predictions of Chapter 2 are used to address one major problem in the evolution of learning rules and cognition: can animals infer, by reasoning, the rewards of actions that they do not necessarily try? This ability may underlie advanced cognition such as the capacity to form beliefs about the world. Humans are known to form mental representations (i.e. beliefs) of their world and of others' behaviors (theory of mind) but it remains controversial whether other animal species can also do it (Premack and Woodruff, 1978; Emery and Clayton, 2009). By comparing a simple form of learning, which is trial-and-error learning, with a model of inference-based learning, a contribution to finding an answer to this question is developed. In particular, the idea that the ability to infer missed payoffs (i.e. payoffs yielded by actions an individual did not explicitly try) gives an advantage in social interactions is tested.

The main work during this thesis (Chapters 2 and 3) was concerned with trying to distinguish the behavior produced by various learning rules, so a natural question linked to the above discussion is: what do all these learning rules share in common? In other words, what is the distinguishing feature of learning? In particular, from an evolutionary viewpoint, does learning provide an advantage over other simpler forms of plasticity? For some, it is not even a question that deserves attention as they claim that learning is just an emerging property of neural systems (Hollis and Guillette, 2011). With this view, learning is just a specific form of phenotypic

plasticity. However, this argument is not satisfying because animals do have innate preferences: some predetermined responses to particular stimuli, and these responses cohabit and interact with learning abilities (e.g. Mery and Kawecki, 2004; Gong, 2012). Why then are not all ecologically relevant situations encoded in an animal's brain? It seems obvious that such a proposition is computationally intractable: the world contains so many different situations that it is impossible to encode them all in a single brain. In Chapter 4, it is argued that learning helps dealing with the world's complexity, in particular thanks to the ability to forget past information.

Chapter 2

On learning dynamics underlying the evolution of learning rules

Abstract

In order to understand the development of non-genetically encoded actions during an animal's lifespan, it is necessary to analyze the dynamics and evolution of learning rules producing behavior. Owing to the intrinsic stochastic and frequency-dependent nature of learning dynamics, these rules are often studied in evolutionary biology via computer simulations. We show here that stochastic approximation theory can help to qualitatively understand learning dynamics and formulate analytical models for the evolution of learning rules. Individuals repeatedly interact during their lifespan, where the stage game faced by the individuals fluctuates according to an environmental stochastic process. Individuals adjust their behavioral actions according to learning rules belonging to the class of experience-weighted attraction learning mechanisms, which includes standard reinforcement and Bayesian learning as special cases. We use stochastic approximation theory in order to derive differential equations governing action play probabilities, which turn out to have qualitative features of mutator-selection equations. We then perform agent-based simulations to find the conditions where the deterministic approximation is closest to the original stochastic learning process for standard 2-action 2-player fluctuating games, where interaction between learning rules and preference reversal may occur. Finally, we analyze a simplified model for the evolution of learning in a producer-scrounger game, which shows that the exploration rate can interact in a non-intuitive way with other features of co-evolving learning rules. Overall, our analyses illustrate the usefulness of applying stochastic approximation theory in the study of animal learning.

2.1 Introduction

The abundance of resources and the environments to which organisms are exposed vary in space and time. Organisms are thus facing complex fluctuating biotic and abiotic conditions to which they must constantly adjust (Shettleworth, 2009; Dugatkin, 2010).

Animals have a nervous system, which can encode behavioral rules allowing them to adjust their actions to changing environmental conditions (Shettleworth, 2009; Dugatkin, 2010). In particular, the presence of a reward system allows an individual to reinforce actions increasing satisfaction and material rewards and thereby adjust behavior by learning to produce goal-oriented action paths (Thorndike, 1911; Herrnstein, 1970; Sutton and Barto, 1998; Niv, 2009). It is probable that behaviors as different as foraging, mating, fighting, cooperating, nest building, or information gathering all involve adjustment of actions to novel environmental conditions by learning, as they have evolved to be performed under various ecological contexts and with different interaction partners (Hollis et al., 1995; Chalmeau, 1994; Villarreal and Domjan, 1998; Walsh et al., 2011; Plotnik et al., 2011).

In the fields of evolutionary biology and behavioral ecology there is a growing interest in understanding how natural selection shapes the learning levels and abilities of animals, but this is met with difficulties (McNamara and Houston, 2009; Hammerstein and Stevens, 2012; Fawcett et al., 2013; Lotem, 2013). Focusing on situation specific actions does not help to understand the effects of natural selection on behavioral rules because one focuses on produced behavior and not the rules producing the behavior (e.g., Dijker, 2011). In order to understand the dynamics and evolution of learning mechanisms and other behavioral rules, an evolutionary analysis thus has to consider explicitly the dynamics of state variables on two timescales. First, one has to consider the timescale of an individual's lifespan; that is, the behavioral timescale during which genetically encoded behavioral rules produce a dynamic sequence of actions taken by the animal. Second, there is the generational timescale, during which selection occurs on the behavioral rules themselves.

It is the behavioral timescale, where learning may occur, that seems to be the most reluctant to be analyzed (Lotem, 2013). This may stem from the fact that learning rules intrinsically encompass constraints about the use of information and the expression of actions (in the absence unlimited powers of computation), which curtails the direct application of standard optimality approaches for studying dynamic behavior such as optimal control theory and dynamic programming. Indeed, the dynamics of even the simplest learning rule, such as reinforcement learn-

ing by trial-and-error, is hardly amenable to mathematical analysis without simplifying assumptions and focusing only on asymptotics (Bush and Mostelller, 1951; Norman, 1968; Rescorla and Wagner, 1972; Börgers and Sarin, 1997; Stephens and Clements, 1998; but see Izquierdo et al., 2007 for predictions in finite time).

Further, the difficulty of analyzing learning dynamics is increased by two biological features that need to be taken into account. First, varying environments need to be considered because learning is favored by selection when the environment faced by the individuals in a population is not absolutely fixed across and/or within generations (Boyd and Richerson, 1988; Rogers, 1988; Stephens, 1991; Feldman et al., 1996; Wakano et al., 2004; Dunlap and Stephens, 2009). Second, frequency-dependence needs to be considered because learning is likely to occur in situations where there are social interactions between the individuals in the population (Chalmeau, 1994; Hollis et al., 1995; Villarreal and Domjan, 1998; Giraldeau and Caraco, 2000; Arbilly et al., 2010, 2011*b*; Plotnik et al., 2011).

All these features taken together make the analysis of the evolution of learning rules more challenging to analyze than standard evolutionary game theory models focusing on actions or strategies for constant environments (e.g., Axelrod and Hamilton, 1981; Maynard-Smith, 1982; Binmore and Samuelson, 1992; Leimar and Hammerstein, 2001; McElreath and Boyd, 2007; André, 2010). Although there has been some early studies on evolutionarily stable learning rules (Harley, 1981; Houston, 1983; Houston and Sumida, 1987; Tracy and Seaman, 1995), this research field has only recently been reignited by the use of agent-based simulations (Groß et al., 2008; Josephson, 2008; Hamblin and Giraldeau, 2009; Arbilly et al., 2010, 2011*a*,*b*; Katsnelson et al., 2011). It is noteworthy that during the gap in time in the study of learning in behavioral ecology, the fields of game theory and economics have witnessed an explosion of theoretical studies of learning dynamics (e.g., Jordan, 1991; Erev and Roth, 1998; Fudenberg and Levine, 1998; Camerer and Ho, 1999; Hopkins, 2002; Hofbauer and Sandholm, 2002; Foster and Young, 2003; Young, 2004; Sandholm, 2011). This stems from an attempt to understand how humans learn to play in games (e.g., Camerer, 2003) and to refine static equilibrium concepts by introducing dynamics. Even if such motivations can be different from the biologists' attempt to understand the evolution of animal behavior, the underlying principles of learning are similar since actions leading to high experienced payoffs (or imagined payoffs) are reinforced over time.

Interestingly, mathematicians and game theorists have also developed tools to analytically approximate intertwined behavioral dynamics, in particular stochastic approximation theory

(Ljung, 1977; Benveniste et al., 1991; Fudenberg and Levine, 1998; Benaim and Hirsch, 1999*a*; Kushner and Yin, 2003; Young, 2004; Sandholm, 2011). Stochastic approximation theory allows one to approximate by way of differential equations discrete time stochastic learning processes with decreasing (or very small) step-size, and thereby understand qualitatively their dynamics and potentially construct analytical models for the evolution of learning mechanisms. This approach does not seem so far to have been applied in evolutionary biology.

In this paper, we analyze by means of stochastic approximation theory an extension to fluctuating social environments of the experience-weighted attraction learning mechanism (EWA model, Camerer and Ho, 1999; Ho et al., 2007). This is a parametric model, where the parameters describe the psychological characteristics of the learner (memory, ability to imagine payoffs of unchosen actions, exploration/exploitation inclination), and which encompasses as a special case various learning rules used in evolutionary biology such as the linear operator (McNamara and Houston, 1987; Bernstein et al., 1988; Stephens and Clements, 1998), relative payoff sum (Harley, 1981; Hamblin and Giraldeau, 2009) and Bayesian learning (Rodriguez-Gironés and Vásquez, 1997; Geisler and Diehl, 2002). We apply the EWA model to a situation where individuals face multiple periods of interactions during their lifetime, and where each period consists of a game (like a prisoner's dilemma game, a hawk-dove game), whose type changes stochastically according to an environmental process.

The paper is organized in three parts. First, we define the model and derive by way of stochastic approximation theory a set of differential equation describing action play probabilities out of which useful qualitative features about learning dynamics can be read. Second, we use the model to compare analytical and simulation results under some specific learning rules. Finally, we derive an evolutionary model for patch foraging in a producer-scrounger context, where both evolutionary and behavioral time scales are considered.

2.2 Model

2.2.1 Population

We consider a haploid population of constant size N. Although we are mainly interested in investigating learning dynamics, we endow for biological concreteness the organisms with a simple life cycle. This is as follows. (1) Each individual interacts socially with others repeatedly and possibly for T time periods. (2) Each individual produces a large number of offspring

according to its gains and losses incurred during social interactions. (3) All individuals of the parental generation die and N individuals from the offspring generation are sampled to form the new adult generation.

2.2.2 Social decision problem in a fluctuating environment

The social interactions stage of the life cycle, stage (1), is the main focus of this paper and it consists of the repeated play of a game between the members of the population. At each time step $t = 1,2,\ldots,T$, individuals play a game, whose outcome depends on the state of the environment ω. We denote the set of environmental states by Ω, which could consist of good and bad weather, or any other environmental biotic or abiotic feature affecting the focal organism. The dynamics of environmental states $\{\omega_t\}_{t=1}^{T}$ is assumed to obey a homogeneous and aperiodic Markov Chain, and we write $\mu(\omega)$ for the probability of occurrence of state ω under the stationary distribution of this Markov Chain (e.g., Karlin and Taylor, 1975; Grimmett and Stirzaker, 2001).

For simplicity, we consider that the number of actions in the game stays constant across environmental states (only the payoffs vary), that is, at every time step t, all individuals have a fixed behavioral repertoire that consists of the set of actions $\mathscr{A} = \{1,\ldots,m\}$. The action taken by individual i at time t is a random variable denoted by $a_{i,t}$, and the action profile in the population at time t is $\mathbf{a}_t = (a_{1,t},\ldots,a_{N,t})$. This process generates a sequence of action profiles $\{\mathbf{a}_t\}_{t=1}^{T}$. The payoff to individual i at time t when it takes action $a_{i,t}$ and the game is in state ω_t is denoted $\pi_i(a_{i,t},\mathbf{a}_{-i,t},\omega_t)$, where $\mathbf{a}_{-i,t} = (a_{1,t},\ldots,a_{i-1,t},a_{i+1,t},\ldots,a_{N,t})$ is the action profile of the remaining individuals in the population (all individuals except i). Note that this setting covers the case of an individual decision problem (e.g., a multi-armed bandit), where the payoff $\pi_i(a_{i,t},\omega_t)$ of individual i is independent of the profile of actions $\mathbf{a}_{-i,t}$ of the other members of the population.

2.2.3 Learning process

We assume that individuals learn to choose their actions in the game but are unable to detect the current state ω_t of the environment. Each individual is characterized by a genetically determined learning rule, which prescribes how its current actions depend on its private history. The learning rules we consider belong to the class of rules defined by the so-called experience-weighted-attraction (EWA) learning model (Camerer and Ho, 1999; Camerer, 2003; Ho et al.,

2007). The reason why we use EWA is that it encapsulates many standard learning rules and translates well the natural assumption that animals have internal states, which are modified during the interactions with their environment, and that internal states have a direct (but possibly noisy) influence on action (Enquist and Ghirlanda, 2005). In EWA learning, the internal states are attractions or "motivations" for actions, and the mapping from internal states (motivations) to action choice is realized via a probabilistic choice rule.

Dynamics of motivations

We first describe the dynamics of motivations. To each available action a of its action set $\mathscr{A}$, individual i has an associated motivation $M_{i,t}(a)$ at time t that is updated according to

$$M_{i,t+1}(a) = \frac{n_{i,t}}{n_{i,t+1}}\phi_{i,t}M_{i,t}(a) + \frac{1}{n_{i,t+1}}\{\delta_i + (1-\delta_i)\mathbb{1}(a,a_{i,t})\}\pi_i(a,\mathbf{a}_{-i,t},\omega_t), \tag{2.1}$$

where

$$n_{i,t+1} = 1 + \rho_i n_{i,t}. \tag{2.2}$$

is individual i's count of the number of steps of play. The initial conditions of eq. 2.1 and eq. 2.2 are the values of the state variables at the first period of play ($t = 1$); that is, $M_{i,1}(a)$ and $n_{i,1}$.

The updating rule of motivations (eq. 2.1) is a weighted average between the previous motivation to action a, $M_{i,t}(a)$, and a reinforcement to that action, $\{\delta_i + (1-\delta_i)\mathbb{1}(a,a_{i,t})\}\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$, which itself depends on the payoff $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$ that would obtain if action a was played at t. Eq. 2.1 is equivalent to eq. 2.2 of Camerer and Ho (1999) with the only difference that the payoff depends here on the current state of the environment, ω_t, so that individuals face a stochastic game.

The first term in eq. 2.1 weights the previous motivation by two factors: $\phi_{i,t}$, a positive dynamic memory parameter that indicates how well individual i remembers the previous motivation; and the experience weight $n_{i,t}/n_{i,t+1}$, which is the ratio between the previous experience count to the new one. Eq. 2.2 shows that the experience count is updated according to another memory parameter, $\rho_i \in [0,1]$. If $\rho_i = 1$, the individual counts the number of interactions objectively, i.e., $n_{i,t} = t$ (if $n_{i,1} = 1$), otherwise subjectively.

The reinforcement term to action a in eq. 2.1 is weighted by $1/n_{i,t+1}$ and depends on δ_i, which varies between 0 and 1. This captures the ability of an individual to observe (or mentally simulate) non-realized payoffs, while $\mathbb{1}(a,a_{i,t})$ is the action indicator function of individual i,

given by

$$\mathbb{1}(a,a_{i,t}) = \begin{cases} 1, & \text{if } a_{i,t} = a, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

With these definitions, we can see that depending on the value of δ_i, an individual can reinforce an unchosen action according to the payoff that action would have yielded had it been taken. Indeed, when individual i does not take action a at time t [$\mathbb{1}(a,a_{i,t}) = 0$], the numerator of the second term is $\delta_i \pi_i(a,\mathbf{a}_{-i,t},\omega_t)$. If $\delta_i = 0$, this cancels out and the payoff associated to the unchosen action a has no effect on the update of motivational states. But if $\delta_i = 1$, the numerator of the second term is $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$, and the motivation is updated according to the payoff individual i would have obtained by taking action a. All values of δ_i between 0 and 1 allow to reinforce unchosen actions according to their potential payoff.

On the other hand, if action a is played at time t; namely, $\mathbb{1}(a,a_{i,t}) = 1$, the numerator of the second term reduces to the realized payoff $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$, irrespective of the value of δ_i. Hence, δ_i plays a role only for updating motivations of unchosen actions, which occurs when individuals are belief-based or Bayesian learners as will be detailed below, after we have explained how the actions themselves are taken by an individual.

Action play probabilities

The translation of internal states (motivations) into action choice can take many forms. But it is natural to assume that the probability $p_{i,t}(a) = \Pr\{a_{i,t} = a\}$ that individual i takes action a at time t is independent of other individuals and takes the ratio form

$$p_{i,t}(a) = \frac{f(M_{i,t}(a))}{\sum_{k\in\mathscr{A}} f(M_{i,t}(k))}, \tag{2.4}$$

where $f(\cdot)$ is a continuous and increasing function of its argument (this ratio form could also be justified by appealing to the choice axiom of Luce, 1959, p. 6).

The choice rule (eq. 2.4) entails that the action that has maximal motivation at time t is chosen with the greatest probability. This is different from choosing deterministically the action that has the highest motivation. Indeed, this type of choice function allows one to model errors or exploration in the decision process of the animal (an action with a low motivation has still a probability of being chosen).

Errors can be formally implemented by imposing that

$$p_{i,t}(a) = \mathbb{P}\{a = \operatorname{argmax}_{b\in\mathscr{A}}[M_{i,t}(b) + \varepsilon(b)]\} \tag{2.5}$$

where $(\varepsilon(b))_{b\in\mathscr{A}}$ are small perturbations that are independently and identically distributed (i.i.d.) among choices. The idea is here to first perturb motivations by adding a small random vector ε of errors and then choose the action that has the biggest motivation. The probability that action a has maximal perturbed motivation defines the probability $p_{i,t}(a)$ with which action a will be chosen.

The maximizing assumption in eq. 2.5 restricts the possibilities for the form taken by f. In fact, the only function satisfying at the same time both eqs. 2.4-2.5 is $f(M)=\exp(\lambda_i M)$ for $0<\lambda_i<\infty$ depending on the distribution of perturbations (Sandholm, 2011, Chap. 6). Replacing this expression for f in eq. 2.4, we obtain that an organism chooses its actions according to the so-called logit choice function

$$p_{i,t}(a)=\frac{\exp[\lambda_i M_{i,t}(a)]}{\sum_k \exp[\lambda_i M_{i,t}(k)]}, \tag{2.6}$$

which is in standard use across disciplines (Luce, 1959; Anderson et al., 1992; McKelvey and Palfrey, 1995; Fudenberg and Levine, 1998; Sutton and Barto, 1998; Camerer and Ho, 1999; Achbany et al., 2006; Ho et al., 2007; Arbilly et al., 2010, 2011*b*).

The parameter λ_i can be seen as individual i's sensitivity to motivations, errors in decision-making, or as a proneness to explore actions that have not been expressed so far. Depending on the value of λ_i, we can obtain almost deterministic action choice or a uniform distribution over actions. If λ_i goes to zero, action a is chosen with probability $p_{i,t}(a)\to 1/m$ (where m is the number of available actions). In this case, choice is random and individual i is a pure explorer. If, on the other hand, λ_i becomes very large ($\lambda_i\to\infty$), then the action $a^*=\text{argmax}_{b\in\mathscr{A}}[M_{i,t}(b)]$ with the highest motivation is chosen almost deterministically, $p_{i,t}(a^*)\to 1$. In this case, individual i does not explore, it only exploits actions that led to high payoff. For intermediate values of λ_i, individual i trades off exploration and exploitation.

Learning rules in the EWA genotype space

In the EWA model, individuals differ by the value of the four parameters $\phi_{i,t},\rho_i,\delta_i,\lambda_i$ and the initial values of the state variables, $M_{i,1}(a)$ and $n_{i,1}$. These can be thought of as the genotypic values of individual i, and particular choice of these parameters provide particular learning rules. In Table 2.1, we retrieve from the model (eq. 2.1) some standard learning rules, which are special cases of the genotype space. The Linear Operator rule (Bush and Mostelller, 1951; Rescorla and Wagner, 1972; McNamara and Houston, 1987; Bernstein et al., 1988; Stephens and

Clements, 1998; Hamblin and Giraldeau, 2009), Relative Payoff Sum (Harley, 1981; Houston, 1983; Houston and Sumida, 1987; Tracy and Seaman, 1995), Cournot Adjustment (Cournot, 1838), and Fictitious Play (Brown, 1951; Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002; Hopkins, 2002) all can be expressed as special cases of EWA.

One of the strengths of EWA is that it encompasses at the same time both reinforcement learning (like the linear operator or relative payoff sum) and belief-based learning (like fictitious play) despite the fact that these two types of learning rules are usually thought of as cognitively very different (Erev and Roth, 1998; Hopkins, 2002; van der Horst et al., 2010). Reinforcement learning is the simplest translation of the idea that actions associated to high rewards are more often repeated, while belief-based learning relies on updating beliefs (probability distributions) over the actions of other players and/or the state of the environment, which occurs in Bayesian learning. In the EWA model, belief-based learning is made possible thanks to the ability to imagine outcomes of unchosen actions (Emery and Clayton, 2004); this is captured by the parameter δ_i, which is the key to differentiate reinforcement from belief-based learning models.

Belief-based learning is captured in the EWA model since motivations can represent the expected payoff of action over the distribution of beliefs of the actions of other players (Camerer and Ho, 1999), and the logit choice function further allows an individual to best respond to the actions of others. It then turns out that the Smooth Fictitious Play (FP) rule (Table 2.1) is equivalent to Bayesian learning for initial priors over the actions of others (stage game $t = 1$) that follow a Dirichlet distribution (Fudenberg and Levine, 1998, p. 48–49).

In EWA, the learning dynamics of an individual (eqs. 2.1–2.4) is a complex discrete time stochastic process because action choice is probabilistic and it depends on the (random) actions played by other individuals in the population, and on the random variable ω_t. In Fig. 2.1, we show a simulation of a typical learning dynamics of two interacting individuals who learn according to the EWA model (eqs. 2.1–2.4) in a repeated Hawk-Dove game, and with actions play probabilities following the logit choice rule (eq. 2.6). Is it possible to approximate this dynamics in order to obtain a qualitative understanding of the change of play probabilities?

Table 2.1: Special cases of EWA learning (eqs. 2.1–2.4). The first column gives the usual name of the learning rule found in the literature. The other columns give the parameter values in the EWA model to obtain this rule and the explicit expression of motivation updating ($M_{i,t+1}(a)$). A cell with a dot (.) means that the parameter in the corresponding column can take any value. A value of $\phi_{i,t}$ with the subscript t removed means that ϕ_i is a constant. The first part of the table gives the rules already defined in the literature while the second part gives PRL, ERL, and IL, the three learning rules introduced in this paper. See Appendix A.4 for an explanation of how to obtain Tit-for-Tat from EWA, where $L_i(a)$ is the aspiration level by i for action a.

Learning rule	$\phi_{i,t}$	ρ_i	δ_i	λ_i	$n_{i,1}$	$M_{i,t+1}(a)$	$p_{i,t}(a)$
Linear Operator	$0<\phi_i<1$	$\rho_i=\phi_i$	0	.	$1/(1-\rho_i)$	$\phi_i M_{i,t}(a)+(1-\phi_i)\mathbb{1}(a,a_{i,t})\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	.
Relative Payoff Sum	$0<\phi_i\leq 1$	0	0	.	1	$\phi_i M_{i,t}(a)+\mathbb{1}(a,a_{i,t})\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto M_{i,t}(a)$
Cournot Adjustment	0	0	1	∞	1	$\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$
Stochastic Fictitious Play (FP) (equivalent to Bayesian learning with Dirichlet distributed priors)	1	1	1	$\lambda_i>0$	.	$\frac{t}{t+1}M_{i,t}(a)+\frac{1}{t+1}\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$
Tit-for-Tat	0	0	1	∞	1	$\pi_i(a,\mathbf{a}_{-i,t},\omega_t)-L_i(a)$	$\propto \exp[\lambda_i M_{i,t}(a)]$
Pure Reinforcement Learning (PRL)	$1+\frac{1}{t}$	1	0	$\lambda_i>0$	1	$M_{i,t}(a)+\frac{1}{t+1}\mathbb{1}(a,a_{i,t})\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$
Exploratory Reinforcement Learning (ERL)	1	1	0	$\lambda_i>0$	1	$\frac{t}{t+1}M_{i,t}(a)+\frac{1}{t+1}\mathbb{1}(a,a_{i,t})\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$
Payoff-Informed Learning (IL)	$1+\frac{1}{t}$	1	1	$\lambda_i>0$	1	$M_{i,t}(a)+\frac{1}{t+1}\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$

The expression $\propto \exp[\lambda_i M_{i,t}(a)]$ refers to the logit choice rule (eq. 2.6).

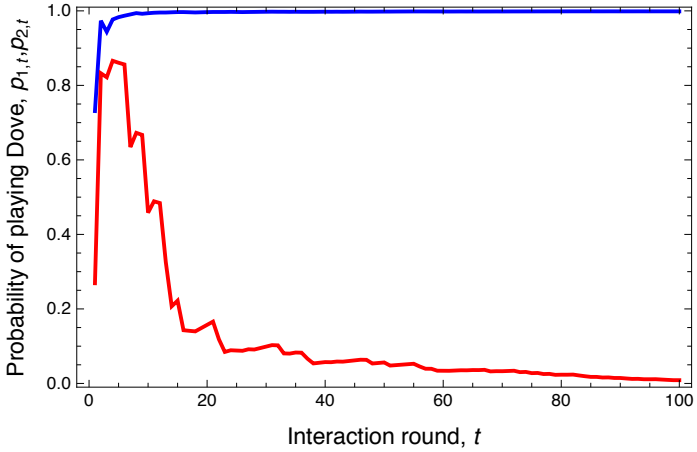

Figure 2.1: Example of learning dynamics for two interacting individuals (1 and 2) in a 2×2 Hawk-Dove game with $\pi(1,1)=B/2$, $\pi(1,2)=0$, $\pi(2,1)=B$, $\pi(2,2)=B/2-C$, where $B=5$ and $C=3$. The blue line represents the probability $p_{1,t}$ to play Dove for individual 1 and the red line the probability $p_{2,t}$ to play Dove for individual 2 when the learning rule is characterized by $\phi_{i,t}=1+1/t$, $\rho_i=1$, $\lambda_i=1$, $n_{i,1}=1$ for both players (rule called Pure Reinforcement Learning, PRL in Table 2.1). Parameters values for player 1 are $\delta_1=0$, $M_{1,1}(\text{Dove})=1$, and $M_{1,1}(\text{Hawk})=0$ (hence $p_{1,1}\approx 0.73$), while for player 2 they are $\delta_2=0$, $M_{2,1}(\text{Dove})=0$, and $M_{2,1}(\text{Hawk})=1$ (hence $p_{2,1}\approx 0.27$).

2.2.4 Stochastic approximation

Differential equations for motivations

We now use stochastic approximation theory (Ljung, 1977; Benveniste et al., 1991; Benaim, 1999; Kushner and Yin, 2003) in order to derive a system of differential equations (ODE) for the motivations and choice probabilities, which produces qualitative and quantitative results about learning dynamics.

The idea behind stochastic approximation is to write eq. 2.1 under the form of a difference equation with decreasing step-size, which then allows one to compute the expected change of the dynamics over one time step. These expected dynamics give rise to differential equations, which describes very closely the long-run stochastic dynamics of the motivations (see Benaim, 1999, for a standard reference, and Hopkins, 2002, for an application of this principle to learning). To that aim, we write eq. 2.1 as

$$M_{i,t+1}(a)-M_{i,t}(a)=\frac{1}{n_{i,t+1}}\Big[-\epsilon_{i,t}M_{i,t}(a)+R_i(a,a_{i,t},\mathbf{a}_{-i,t},\omega_t)\Big]. \tag{2.7}$$

where

$$\epsilon_{i,t} = 1 + n_{i,t}(\rho_i - \phi_{i,t}) \tag{2.8}$$

is a decay rate and

$$R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t) = \left[\delta_i + (1-\delta_i)\mathbb{1}(a, a_{i,t})\right] \pi_i(a, \mathbf{a}_{-i,t}, \omega_t) \tag{2.9}$$

can be interpreted as the net reinforcement of the motivation of action a.

In order to use stochastic approximation theory, we need that the step-size of the process satisfies $\sum_{t=1}^{\infty}(1/n_{i,t}) = \infty$ and $\lim_{t\to\infty}(1/n_{i,t}) = 0$ (Benaim, 1999, p. 11), where the first condition entails that the steps are large enough to eventually overcome initial conditions, while the second condition entails that the steps eventually become small enough so that the process converges. This is ensured here by setting $\rho = 1$ in eq. 2.2. We further assume a constant value of ϵ_i from now on (this is the case for all rules in Table 2.1, but a slowly varying $\epsilon_{i,t}$ is still amenable to an analysis via stochastic approximation). Note however that the assumption that $\rho = 1$ reduce to some extent the number of learning rules that one can analyze in the EWA model, but the approximation can still be useful for small constant step-sizes, for instance if one considers a Linear Operator Rule (Table 2.1) with ϕ_i close to 1 (see Benaim and Hirsch, 1999*b*; Izquierdo et al., 2007, for results on processes with constant step-sizes).

With these assumptions, we show in Appendix A.1 (eqs. A.1–A.12) that the differential equation arising from taking the expected motion of the stochastic dynamics in eq. 2.7 is

$$\dot{M}_i(a) = -\epsilon_i M_i(a) + \bar{R}_i(a), \tag{2.10}$$

where

$$\bar{R}_i(a) = [p_i(a) + \delta_i(1 - p_i(a))] \sum_{\mathbf{a}_{-i} \in \mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i}) \bar{\pi}_i(a, \mathbf{a}_{-i}) \tag{2.11}$$

and

$$\bar{\pi}_i(a, \mathbf{a}_{-i}) = \sum_{\omega \in \Omega} \mu(\omega) \pi_i(a, \mathbf{a}_{-i}, \omega). \tag{2.12}$$

Here, a dot accent is used to denote a derivative, i.e., $\mathrm{d}x/\mathrm{d}t = \dot{x}$, $\bar{R}_i(a)$ is the expected reinforcement to the motivation of action a of individual i over the distribution of action probabilities in the population (where $\mathscr{A}^{N-1}$ is the set of action profiles of individuals different than i), and $\bar{\pi}_i(a, \mathbf{a}_{-i})$ is the average payoff over the distribution of environmental states. Because the action play probabilities of the focal individual, $p_i(a)$, and the remaining individuals in the population $p_{-i}(\mathbf{a}_{-i}) = \prod_{i \neq j} p_j(a_j)$(eq. A.2), depend on the motivations, eq. 2.10 is a differential of the form

$\dot{M}_i(a) = F_i(\mathbf{M})$, for all actions a and individual i in the population, where $\mathbf{M}$ denotes the vector collecting the motivations of all actions and individuals in the population. Hence, eqs. 2.10–2.12 define a bona fide autonomous system of differential equations.

Eq. 2.11 shows that the deterministic approximation rests on the "average game" with payoffs given by $\bar{\pi}_i(a, \mathbf{a}_{-i})$, i.e., a game where each payoff matrix entry (eq. 2.12) is a weighted average of the corresponding entries of the stage games over the distribution of environmental states $\mu(\omega)$. Hence, if one wants to consider a situation where the stage game fluctuates, one does not need to specify a series of stage games, but only the average game resulting from taking the weighted average of the payoffs of the original stage games.

Differential equations for action play probabilities

Using the logit choice rule (eq. 2.6) and the dynamics of motivations (eq. 2.10), we can derive a differential equation for the choice probability for each action a of individual i

$$\dot{p}_i(a) = p_i(a)\left[\epsilon_i \sum_{k \in \mathscr{A}} \log\left(\frac{p_i(k)}{p_i(a)}\right) p_i(k) + \lambda_i \left(\bar{R}_i(a) - \sum_{k \in \mathscr{A}} \bar{R}_i(k) p_i(k)\right)\right], \tag{2.13}$$

(Appendix A.1, eqs. A.13–A.20). Because $\bar{R}_i(a)$ depends on the action play probabilities, eq. 2.13 also defines a bona fide autonomous system of differential equations, but this time directly for the dynamics of action. The first term in brackets in eq. 2.13 describes a perturbation to the choice probability. This represents the exploration of action by individual i (it is an analogue of mutation in evolutionary biology), and brings the dynamics back into the interior of the state space if it gets too close to the boundary. The second term in the brackets takes the same form as the replicator equation (Hofbauer and Sigmund, 1998; Tuyls et al., 2003); that is, if the expected reinforcement, $\bar{R}_i(a)$, to action a is higher than the average expected reinforcement, $\sum_k \bar{R}_i(k) p_i(k)$, then the probability of expressing action a increases.

Eq. 2.13 is the "final" point of the stochastic approximation applied to our model. We now have a system of differential equations [of dimension $N \times (m-1)$], which describes the ontogeny of behavior of the individuals in the population. Standard results from stochastic approximation theory guarantee that the original stochastic dynamics (eqs. 2.1–2.4) asymptotically follows very closely the deterministic path of the differential equation 2.13. For instance, if the limit set of eq. 2.13 consists of isolated equilibria, the stochastic process (eqs. 2.1–2.4) will converge to one of these equilibria almost surely (Benaim, 1999; Borkar, 2008).

More generally, the differential equations for the action probabilities are unlikely to depend

only on the probabilities as is the case in eq. 2.13. For instance, when the choice rule is the so-called power choice, i.e., $f(M) = M^{\lambda_i}$ in eq. 2.4, which gives rise to

$$\dot{p}_i(a) = \left[\lambda_i p_i(a)/M_i(a)\right]\left[\bar{R}_i(a) - \sum_{k\in\mathscr{A}} \bar{R}_i(k)p_i(k)\{p_i(a)/p_i(k)\}^{1/\lambda_i}\right],$$

(eq. A.22), the dynamics of actions will also depend on the dynamics of motivations. This is one of the reasons why the logit choice rule is appealing; namely, it yields simplifications allowing one to track only the dynamics of choice probabilities (eq. 2.13).

2.3 Applications

2.3.1 Pure reinforcement vs. Payoff-Informed Learning

We now apply our main result (eq. 2.13) to a situation where two individuals ($N = 2$) are interacting repeatedly and can express only two actions during the stage game, action 1 and 2. In this case, only three generic symmetric stage games are possible; namely a game with a dominant strategy (e.g., a Prisoner's Dilemma game, PD), a game with two pure asymmetric Nash Equilibria (e.g., a Hawk-Dove game, HD), and a game with two pure symmetric NE (e.g., a Coordination Game, CG) so that the set of games can be taken to be $\Omega = \{\text{PD}, \text{CG}, \text{HD}\}$ (Weibull, 1997, Chap. 1). We call $\mathscr{R}_\omega$ the payoff obtained when the game is ω and both players play action 1 (see Table 2.2 for the description of the payoffs for each game ω), so that the average payoff obtained when both players play action 1 is $\mathscr{R} = \mu(\text{PD})\mathscr{R}_{\text{PD}} + \mu(\text{CG})\mathscr{R}_{\text{CG}} + \mu(\text{HD})\mathscr{R}_{\text{HD}}$. Likewise, one can evaluate the payoffs $\mathscr{S}$, $\mathscr{T}$, and $\mathscr{P}$ of the average game, when, respectively, player 1 plays action 1 and player 2 plays action 2, player 1 plays action 2 and player 2 plays action 1, and both players play action 2 (Table 2.2).

We assume that individuals playing this stochastic game use the learning rules characterized by

$$\phi_{i,t} = 1 + \frac{1}{t} \text{ and } \rho_i = 1 \tag{2.14}$$

so that when $\delta_i = 0$ we obtain a form of reinforcement learning, which we call Pure Reinforcement Learning (PRL: see Table 2.1) because motivations are updated only according to realized payoffs and there is no discounting of the past. When $\delta_i = 1$ we obtain a rule we call Payoff-Informed Learning (IL: see Table 2.1) since in that case an individual updates motivations according not only to realized but also to imagined payoffs. The individual has here all information about possible payoffs at each decision step t, hence the name of the learning rule.

Substituting eqs. 2.14 into eq. 2.8 gives $\epsilon_{i,t} = 0$ (since $n_t = t$) and thus $\epsilon_i = 0$ in eq. 2.13. Letting $p_1 = p_1(1)$ be the probability that individual 1 plays action 1 and $p_2 = p_2(1)$ be the probability that individual 2 plays action 1, we then obtain from eq. 2.8, the above assumptions, and Table 2, that the action play probabilities satisfy the dynamics

$$\dot{p}_1 = p_1(1-p_1)\lambda_1[\{p_2\mathscr{R} + (1-p_2)\mathscr{S}\}\{p_1 + \delta_1(1-p_1)\} - \{p_2\mathscr{T} + (1-p_2)\mathscr{P}\}\{\delta_1 p_1 + (1-p_1)\}], \quad (2.15)$$

$$\dot{p}_2 = p_2(1-p_2)\lambda_2[\{p_1\mathscr{R} + (1-p_1)\mathscr{S}\}\{p_2 + \delta_2(1-p_2)\} - \{p_1\mathscr{T} + (1-p_1)\mathscr{P}\}\{\delta_2 p_2 + (1-p_2)\}]. \quad (2.16)$$

In order to compare the dynamics predicted by eqs. 2.15–2.16 to that obtained from iterating eq. 2.1 with logit choice function (eq. 2.6; agent-based simulations), we assume that the average game is a Hawk-Dove game (Maynard-Smith and Price, 1973; Maynard-Smith, 1982). Hence, action 1 can be thought as "Dove" and action 2 as "Hawk". We now focus on two specific interactions in this Hawk-Dove game: PRL vs. PRL, and PRL vs. IL, and in order to carry out the numerical analysis, we also assume that the probability $\mu(\omega)$ that game ω obtains in any period obeys an uniform distribution, which gives $\mu(\text{PD}) = \mu(\text{CG}) = \mu(\text{HD}) = 1/3$.

PRL vs. PRL

When two PRL play against each other (eqs. 2.15–2.16 with $\delta_i = 0$ for both players) in the average Hawk-Dove game, we find that the deterministic dynamic admits three locally stable equilibria (Fig. 2.2A): the two pure asymmetric Nash equilibria, (Dove, Hawk) and (Hawk, Dove), and the Pareto efficient outcome where both individuals play Dove (Appendix A.2). Which outcome is reached by the differential equations depends on the initial conditions, and we characterized a region of the state space of initial conditions that always lead to the (Dove, Dove) equilibrium (the gray region in Fig. 2.2A).

In Fig. 2.3, we compare the deterministic model to the original stochastic learning dynamics by graphing the distance between the probability of playing Dove obtained from the equilibrium of eqs. 2.15–2.16 to that obtained from eq. 2.6 under agent-based simulations for various values of the duration of the game, T. The correspondence between the two processes is affected by the sensitivity to payoff, λ, and the initial difference in motivations to a player between playing Dove and Hawk: $\Delta M_{i,1} = M_{i,1}(1) - M_{i,1}(2)$. If this difference is positive ($\Delta M_{i,1} > 0$), player i is

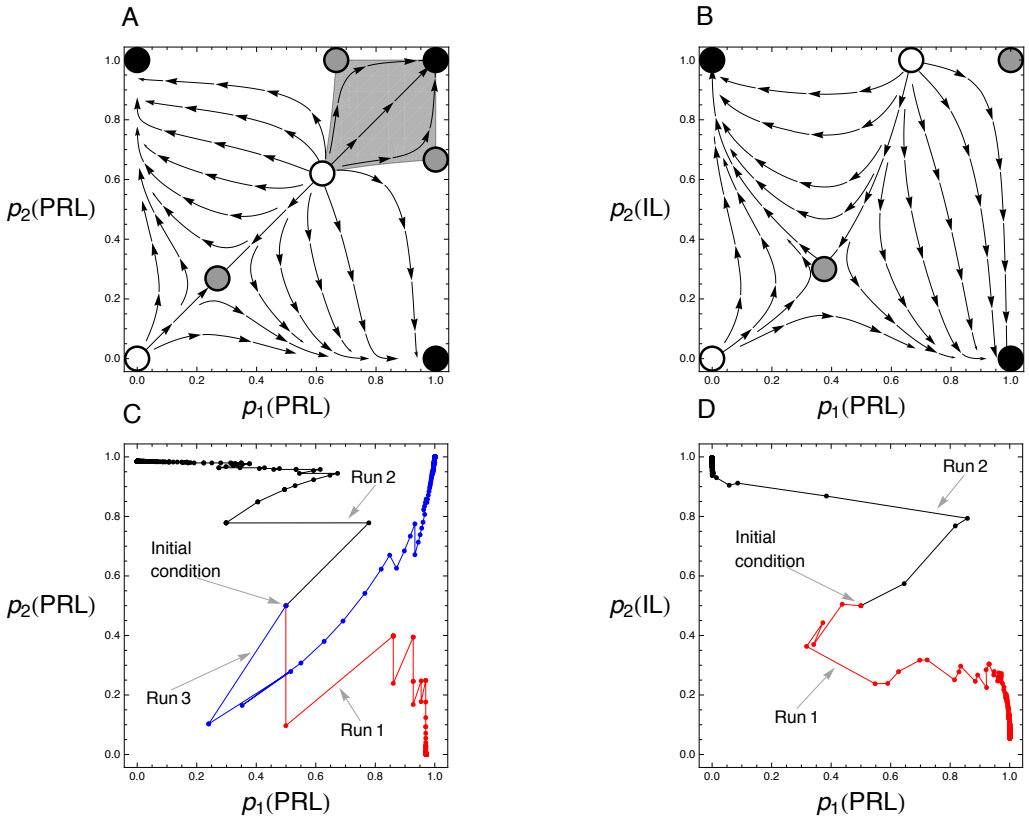

Figure 2.2: Solution orbits for the deterministic dynamics (panels A and B) and sample paths for the stochastic dynamics (panels C and D) for two learners in the average Hawk-Dove game ($B = 5$ and $C = 3$), where the x and y axis represent the probabilities of playing Dove by player 1 and 2, respectively. In panels A and C players 1 and 2 both use the PRL rule, while in panels B and D player 1 uses PRL and player 2 uses IL. The gray shaded area in A represents the initial conditions for which all trajectories go to the $(1, 1)$ equilibrium (Dove, Dove). In panels A and B a white-filled dot denotes an unstable node (both associated eigenvalues are positive), a gray-filled dot is a saddle, and a black dot is a locally stable equilibrium. For the stochastic trajectories (C, D), each color designates a given simulation run (with $\lambda_i = 0.5$ for both players and $T = 2000$) describing a sample path ending in a different equilibrium predicted by the deterministic dynamics. We started all simulations runs from the center of the state space (i.e., $p_1 = p_2 = 1/2$ and $\Delta M_{1,1} = \Delta M_{2,1} = 0$), and each point denotes an interactions round, t. We observe that points are far from each other at the beginning of a simulation run but accumulate near a stable equilibrium at the end of a simulation run. This is because the stochastic shocks are large at the beginning (when the step-size is big), but are smaller as the step-size decreases.

Table 2.2: Payoff matrices for the average Hawk-Dove game and the three associated sub-games. In each matrix, the rows correspond to the actions of player 1 (first row gives action 1, while second row gives action 2) and the columns correspond to the actions of player 2 (first column gives action 1, while second column gives action 2). Payoffs are to row player (player 1). The matrix at the top shows the payoffs in the average Hawk-Dove Game (denoted $\bar{G}$), and the three matrices below contain the payoffs of the sub-games ω (PD, HD, and CG). In the Prisoner's Dilemma (Left), we assume $\mathscr{T}_{\mathrm{PD}} > \mathscr{R}_{\mathrm{PD}} > \mathscr{P}_{\mathrm{PD}} > \mathscr{S}_{\mathrm{PD}}$ and $(\mathscr{T}_{\mathrm{PD}} + \mathscr{S}_{\mathrm{PD}})/2 < \mathscr{R}_{\mathrm{PD}}$. In the Hawk-Dove (Middle), we have $\mathscr{T}_{\mathrm{HD}} > \mathscr{R}_{\mathrm{HD}}, \mathscr{S}_{\mathrm{HD}} > \mathscr{P}_{\mathrm{HD}}, \mathscr{P}_{\mathrm{HD}} > \mathscr{R}_{\mathrm{HD}}$. In the Coordination Game (Right), $\mathscr{R}_{\mathrm{CG}} > \mathscr{S}_{\mathrm{CG}}, \mathscr{R}_{\mathrm{CG}} = \mathscr{P}_{\mathrm{CG}}, \mathscr{S}_{\mathrm{CG}} = \mathscr{T}_{\mathrm{CG}}$.

$\bar{G}$	**Dove**	**Hawk**
Dove	$\mathscr{R} = B/2$	$\mathscr{S} = 0$
Hawk	$\mathscr{T} = B$	$\mathscr{P} = B/2 - C$

PD	**Cooperate**	**Defect**
Cooperate	$\mathscr{R}_{\mathrm{PD}}$	$\mathscr{S}_{\mathrm{PD}}$
Defect	$\mathscr{T}_{\mathrm{PD}}$	$\mathscr{P}_{\mathrm{PD}}$

HD	**Dove**	**Hawk**
Dove	$\mathscr{R}_{\mathrm{HD}}$	$\mathscr{S}_{\mathrm{HD}}$
Hawk	$\mathscr{T}_{\mathrm{HD}}$	$\mathscr{P}_{\mathrm{HD}}$

CG	**Left**	**Right**
Left	$\mathscr{R}_{\mathrm{CG}}$	$\mathscr{S}_{\mathrm{CG}}$
Right	$\mathscr{T}_{\mathrm{CG}}$	$\mathscr{P}_{\mathrm{CG}}$

more likely to play Dove initially since $p_{i,1} > 0.5$, while the player is more likely to play Hawk if the difference is negative ($\Delta M_{i,1} < 0$, which entails $p_{i,1} < 0.5$).

We observe that when λ is very small, the probability of playing an action in the stochastic dynamics remains far from the equilibrium predicted by the deterministic dynamics even if T is large. But when λ_i becomes larger, the match between simulation and approximation becomes very good even for moderate T, unless the difference in initial motivations between actions is close to zero ($\Delta M_{i,1} = 0$). In this case, the initial probability of choosing actions is about $1/2$ for both players and one cannot predict which equilibrium is reached in the deterministic dynamics because the stochastic dynamics may go to any of the three locally stable equilibria (Fig. 2.2C). These features were generally observed when the initial motivations of the players in the stochastic simulations concord with the predicted equilibrium; namely, if the motivations entail initial play probabilities that are closer to the equilibrium than random choice of actions (e.g., if playing Dove is an equilibrium, then we say that $p_{i,1} > 0.5$ concords with the equilibrium).

When the initial motivations of an individual do not concord with an equilibrium, it has to revert his initial preferences. This may for instance be the case when the initial play probabilities of both player favor the equilibrium (Hawk, Hawk), so that $M_{i,1}(1) < M_{i,1}(2)$, which entails

$p_{i,1} < 0.5$. But (Hawk, Hawk) is an unstable equilibrium for the deterministic dynamics, and we know that reinforcement learners cannot learn a behavior that yields a strictly negative payoff as is the case when the equilibrium (Hawk, Hawk) is played. This means that at least one of the players will have to revert its initial preferences in order to reach one of the three stable outcomes, (Dove, Dove), (Hawk, Dove) or (Dove, Hawk).

If preferences need to be reversed and one further has a large λ value, the initial play probability of Dove is close to 0 $p_{i,1}(1) \approx 0$, which is very close to the lower-left corner of the state space in Fig. 2.2A. Preference reversal may then take a very long time, and Fig. 2.4 shows that the time t^* of such a reversal to occur is an increasing function of the magnitude of $\Delta M_{i,1}$. Consequently, while the value of T did not have an important influence on the correspondence between deterministic and stochastic dynamics when the initial preferences were concordant with the predicted equilibrium and λ is not too small (Fig. 2.3), T becomes very important when this is not the case. In effect, when preferences need to be reversed under small T and large $\Delta M_{i,1}$ and λ, one can predict that the difference in play probability observed under the deterministic and stochastic dynamics will be important (as in the lightly shaded regions in Fig. 2.3).

PRL vs. IL

When a PRL plays against an IL (eqs. 2.15–2.16 with $\delta_1 = 0$ for PRL and $\delta_2 = 1$ for IL), we find that asymptotically both players will learn one of the two pure asymmetric Nash equilibria, (Hawk, Dove) and (Dove, Hawk), of the average game, depending on the initial preferences of the players (Appendix A.2, eq. A.24).

As was the case for PRL vs. PRL, the match between deterministic model and stochastic simulation for finite time depends on λ_i and $\Delta M_{i,1}$ (Fig. 2.3B,D,F). However, in this case the region around $\Delta M_{i,1} = 0$, where the analysis gives poor predictions of the real behavior seems larger. Otherwise, the same caveat that we observed for PRL vs. PRL also apply to PRL vs. IL.

In summary, we observed that the deterministic dynamics (eqs. 2.15–2.16) generally approximates qualitatively well the quasi-equilibrium probabilities of playing action obtained under the stochastic learning processes (eqs. 2.1–2.4), but there are extreme cases which are not captured by the deterministic approximations. These are the cases where λ and $\Delta M_{i,1}$ are very small (actions are random) or λ and $\Delta M_{i,1}$ are too big (the dynamics get stuck in suboptimal equilibria). In particular, when $\Delta M_{i,1}$ is too big, the time t for which the stochastic learning process gets close to the deterministic approximation becomes very large (Fig. 2.4). Now that we have a feeling

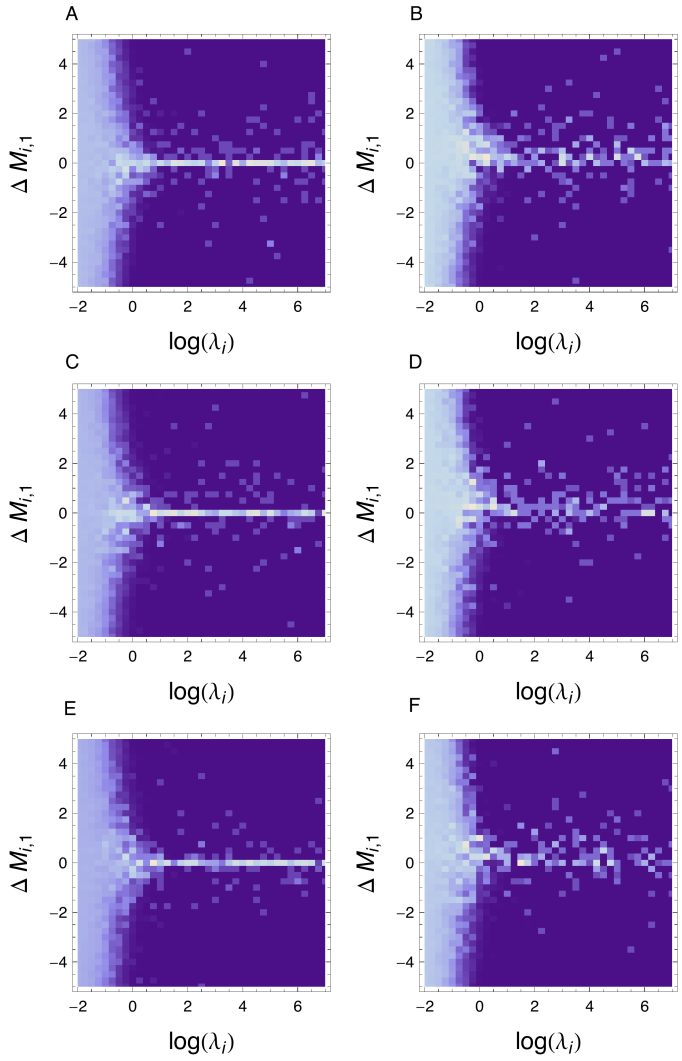

Figure 2.3: Density plot of the average distance between the probability of playing Dove obtained from the equilibrium of the deterministic model (eq. 2.15–2.16) and the stochastic learning dynamics, as a function of λ_i and $\Delta M_{i,1}$. Each data pixel is the average over 5 simulation runs. Lightly shaded regions indicates a big euclidean distance ($\sim \sqrt{2}$) between simulations and analytic prediction, while dark regions indicates a small distance (~ 0). We have $\lambda_1 = \lambda_2$, but motivations are set to opposed values in both individuals: $\Delta M_{1,1} = -\Delta M_{2,1}$. Parameters of the average Hawk-Dove game are $B = 5$ and $C = 3$. (A) PRL vs. PRL and $T = 100$. (B) PRL vs. IL and $T = 100$. (C) PRL vs. PRL and $T = 500$. (D) PRL vs. IL and $T = 500$. (E) PRL vs. PRL and $T = 1000$. (F) PRL vs. IL and $T = 1000$.

about the conditions under which the stochastic approximation can be applied, we turn to the analysis of an evolutionary model of the competition between two different learning rules.

2.3.2 Coevolution of learning and scrounging

Arbilly et al. (2010) explored using agent-based simulations an evolutionary model of foraging where individuals learn to find patches with high quality food. These producers can then be followed by scroungers with which they compete over resources found in the patches. In the same spirit as Arbilly et al. (2010), we analyze here a model for the coevolution between learning and scrounging. Our aim, however, is not to reproduce the results of this earlier model. Rather, it is to analyze a simplified model that is amenable to an illustrative application of the stochastic approximation method in a context where there are two time scales: behavioral and generational.

Biological setting

We consider a population of very large size (say $N \to \infty$), whose members are facing the problem of foraging in an environment consisting of two patch types, labeled 1 and 2. The resource value to an individual foraging in a patch of type a is written $V(a)$ ($a = 1,2$) and we assume that $V(1) > V(2) \geq 0$. In such an environment, learning is necessary if the location of the optimal patch type changes from one generation to the next.

At each decision step t ($1 \leq t < T$) during its lifetime, a learner has to make a choice on whether to forage on patch type 1 or 2 so that the two available actions to a learner are feeding on patch 1 or 2 and its action set can be written $\{1,2\}$. The payoff from feeding on a given patch depends on the number of other individuals on that patch. We assume that there can be no more than two individuals on a patch, so the payoff, $\pi_i(a_{i,t}, \mathbf{a}_{-i,t})$ to individual i taking action $a_{i,t} \in \{1,2\}$ at time t is either $V(a_{i,t})$ or $V(a_{i,t})/2$, where $\mathbf{a}_{-i,t}$ is the random indicator variable representing the presence of another individual on the patch of individual i at time t. If individual i is alone on the patch (which we write $\mathbf{a}_{-i,t} = 0$), it gets the whole resource: $\pi_i(a_{i,t}, 0) = V(a_{i,t})$. If there is another individual on the patch (which we write $\mathbf{a}_{-i,t} = 1$), individual i shares the value with him so its payoff is $\pi_i(a_{i,t}, 1) = V(a_{i,t})/2$.

We assume that there are three types of individuals in this population: Scroungers (S), Fictitious Players (FP), and Exploratory Reinforcement Learners (ERL, see Table 2.1), where both learners (ERL and FP) are exploratory ($\epsilon = 1$ in eq. 2.13). The life-cycle of these individuals is

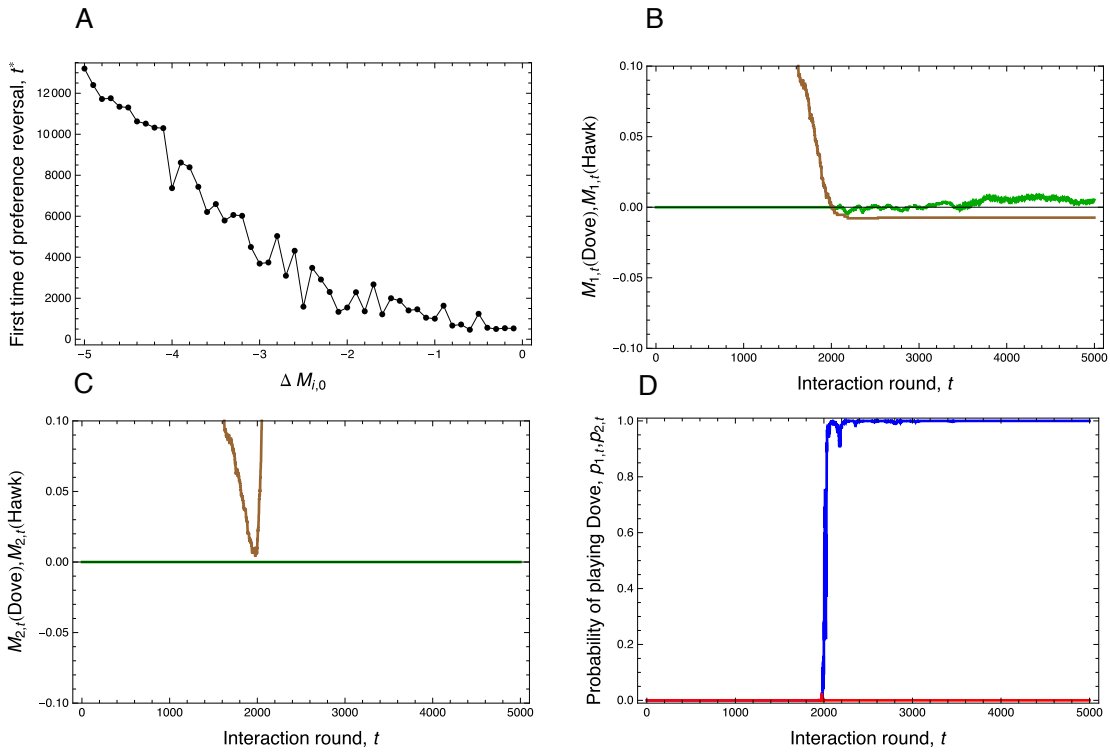

Figure 2.4: Preference reversal for two PRL playing the average Hawk-Dove game with same initial preferences for Hawk ($\Delta M_{i,1} < 0$) and a high sensitivity to motivations ($\lambda_i = 1000$). In panel A, we graph the first time t^* that a preference reversal occurs (i.e., first time that $\Delta M_{i,t^*} > 0$ for at least one of the players) as a function of the magnitude of $\Delta M_{i,1}$. Each point in A is the average over 100 simulation runs. In panel B, we graph the motivations for individual 1 to play Hawk (brown line) and Dove (green line) when $\Delta M_{1,1} = -3$ ($M_{1,1}(\text{Dove}) = 0$, $M_{1,1}(\text{Hawk}) = 3$). The individual has a larger but decreasing motivation for playing Hawk for approximately 2000 rounds, where the motivations reverse to favor Dove. In panel C, we have the corresponding motivations for individual 2 with same parameter values, and this shows that the motivation for Hawk first decreases (same trend as individual 1), but then increases again when its opponent has reversed his preferences. In panel D, we have the Dove action play probabilities for individual 1 (blue line) and 2 (red line).

as described above (section 2.2.1). Note that PRL and IL in the Hawk-Dove game model were characterized by $\epsilon = 0$ (no explicit exploration). Here, FP and ERL can be thought respectively as an extension of PRL and IL to the case of exploratory learning, because the only difference between the rules of the previous Hawk-Dove model (IL and PRL) and the ones in this foraging model (FP and ERL) is the value of ϵ, which determines the presence of explicit exploration (see Table 2.1 for a comparison of these rules). Scroungers do not learn but only follow learners (ERL and FP) and we now describe the learning dynamics of these two types. For simplicity, we do not consider innates (e.g., Feldman et al., 1996; Wakano et al., 2004) as these are likely to be replaced by learners if the latter visit patch type 1 with a probability larger than that obtained by encountering patches at random, and learning is not too costly.

Fictitious Play

Substituting $\epsilon = 1$ into eq. 2.13, and letting $p_{\mathbf{F}} = p_{\mathbf{F}}(1)$ be the probability that an individual of type FP visits patch 1, we obtain that the learning dynamics of FP obeys the differential equation

$$\dot{p}_{\mathbf{F}} = p_{\mathbf{F}}\left[\log\left(\frac{1-p_{\mathbf{F}}}{p_{\mathbf{F}}}\right)(1-p_{\mathbf{F}}) + \lambda_{\mathbf{F}}\left(\bar{R}_{\mathbf{F}}(1) - \left[\bar{R}_{\mathbf{F}}(1)p_{\mathbf{F}} + \bar{R}_{\mathbf{F}}(2)(1-p_{\mathbf{F}})\right]\right)\right]. \tag{2.17}$$

For this model of patch choice, the expected reinforcement to an FP (eq. 2.11) when foraging on patch type a is

$$\bar{R}_{\mathbf{F}}(a) = (1-s)V(a) + s\frac{V(a)}{2} = \left(1 - \frac{s}{2}\right)V(a), \qquad a = 1,2, \tag{2.18}$$

where s denotes the frequency of scroungers in the population. Setting $\dot{p}_{\mathbf{F}} = 0$ one obtains the probability of visiting patch 1 at steady state of learning as

$$\hat{p}_{\mathbf{F}} = \frac{1}{1 + \exp\left(\frac{\lambda_{\mathbf{F}}[2-s][V(2)-V(1)]}{2}\right)}. \tag{2.19}$$

Fig. 2.5A shows that the agreement between predicted equilibrium and that obtained in agent-based simulations is outstanding even if T is not too large, which stems from the fact that the dynamics has a single equilibrium.

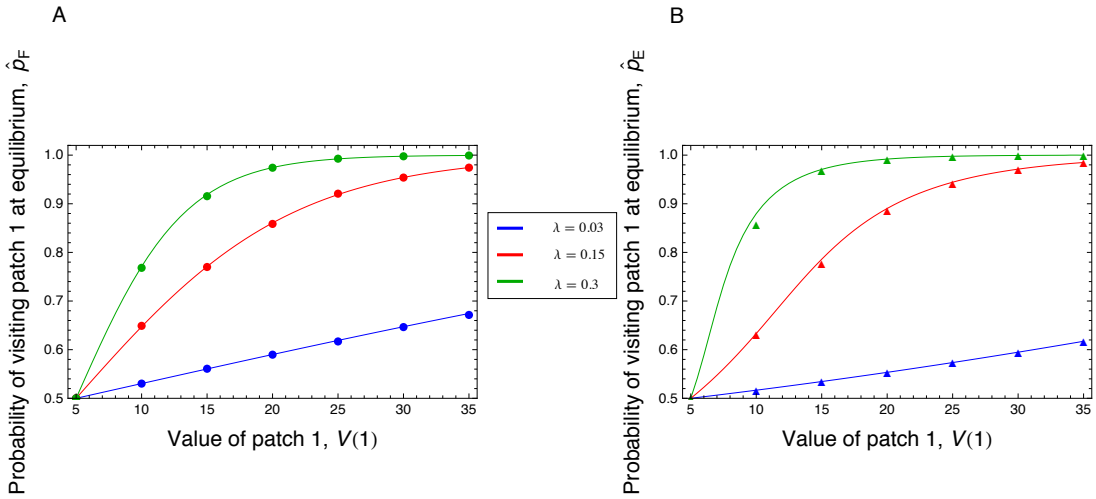

Figure 2.5: Panel A: equilibrium probability $\hat{p}_{\mathrm{F}}$ of visiting patch 1 for a FP (eq. 2.19) graphed as a function of the value of patch 1, $V(1)$, for different values of λ_i. We fixed the value of patch 2 at $V(2) = 5$, the frequency of scroungers to $s = \left(3 - \sqrt{5}\right)/2$ and $\lambda_1 = 0.03$ for the blue line, $\lambda_1 = 0.15$ for the red line, and $\lambda_1 = 0.3$ for the green line. Dots of corresponding colors were obtained from simulations of the original stochastic learning dynamics after $T = 1000$ decision steps in the environment. Panel B: equilibrium probability $\hat{p}_{\mathrm{E}}$ of visiting patch 1 for a ERL (obtained by solving for the equilibrium of eqs. 2.20–2.21) graphed for the same parameter values as in panel A. Triangles of corresponding colors give the average over 1000 runs of stochastic simulations with different initial conditions ($p_{\mathrm{E},1}$) ranging from 0 to 1.

Exploratory Reinforcement Learning

For exploratory reinforcement learning, substituting $\epsilon = 1$ into eq. 2.13, we obtain that the probability that an ERL visits patch 1 obeys

$$\dot{p}_{\mathrm{E}} = p_{\mathrm{E}}\left[\log\left(\frac{1-p_{\mathrm{E}}}{p_{\mathrm{E}}}\right)(1-p_{\mathrm{E}}) + \lambda_{\mathrm{E}}\left(\bar{R}_{\mathrm{E}}(1) - \left[\bar{R}_{\mathrm{E}}(1)p_{\mathrm{E}} + \bar{R}_{\mathrm{E}}(2)(1-p_{\mathrm{E}})\right]\right)\right], \tag{2.20}$$

where the expected reinforcements (eq. 2.11) of going on patch types 1 and 2 are, respectively, given by

$$\begin{aligned} \bar{R}_{\mathrm{E}}(1) &= p_{\mathrm{E}}\left(1-\frac{s}{2}\right)V(1), \\ \bar{R}_{\mathrm{E}}(2) &= (1-p_{\mathrm{E}})\left(1-\frac{s}{2}\right)V(2). \end{aligned} \tag{2.21}$$

The equilibria of eq. 2.20 cannot be solved analytically (this is a transcendental equation in p_{E}). We thus relied on a numerical analysis to obtain its fixed points ($\dot{p}_{\mathrm{E}} = 0$) and focused on the variation of λ_{E} values by performing a bifurcation analysis (with Newton's method using Mathematica, Wolfram Research, Inc., 2011). Fig. 2.6 shows that the phase line passes through three different regimes as λ_{E} increases. These regimes are separated by two critical values of λ_{E}, which we will call $\lambda_{\mathrm{E}}^{\mathrm{crit}_1}$ and $\lambda_{\mathrm{E}}^{\mathrm{crit}_2}$, and provide the following cases.

(I) When $0 < \lambda_{\mathrm{E}} < \lambda_{\mathrm{E}}^{\mathrm{crit}_1}$, the learning dynamics admit one stable interior equilibrium, which is close to 0.5 when λ_{E} is close to 0 and increases as λ_{E} increases, until λ_{E} reaches $\lambda_{\mathrm{E}}^{\mathrm{crit}_1}$.

(II) When $\lambda_{\mathrm{E}}^{\mathrm{crit}_1} < \lambda_{\mathrm{E}} < \lambda_{\mathrm{E}}^{\mathrm{crit}_2}$, there are three interior equilibria. The "completely interior" equilibrium is unstable and the two other interior equilibria are stable. As λ_{E} increases, the two stable equilibria get closer to 0 and 1 and finally collapse when λ_{E} approaches $\lambda_{\mathrm{E}}^{\mathrm{crit}_2}$.

(III) When $\lambda_{\mathrm{E}} > \lambda_{\mathrm{E}}^{\mathrm{crit}_2}$, there is only one interior equilibrium that is unstable. As λ_{E} gets bigger, this equilibrium approaches $V(2)/[V(1)+V(2)]$. In this case, the learner visits only patch 1 when the initial condition is above this value, and visits only patch 0 when the initial condition is below that value.

We also ran simulations of the original stochastic learning dynamics to test the robustness of this numerical analysis and observed that simulations agree very well on average with our numerical analysis based on the stochastic approximation results (Fig. 2.5B).

In the two cases where there are two stable equilibria (cases II and III), which equilibrium is reached depends on the initial conditions of the system (the initial motivations). We postulate

that the initial preference for the patch types is drawn at random from a uniform distribution (Appendix A.3, eq. A.27), which allow us to obtain an expected equilibrium probability that an ERL visits patch 1. When λ_E becomes large, this expectation is the average over visiting only patch 1 or 2, which gives

$$\hat{p}_E = \frac{V(1)}{V(1)+V(2)} \tag{2.22}$$

(Appendix A.3). This gives the matching law (Herrnstein, 1970), if one rescales $V(1)$ and $V(2)$ between 0 and 1 so that they describe the probability to find food at all in the respective patches rather than measuring the "value" of the patches.

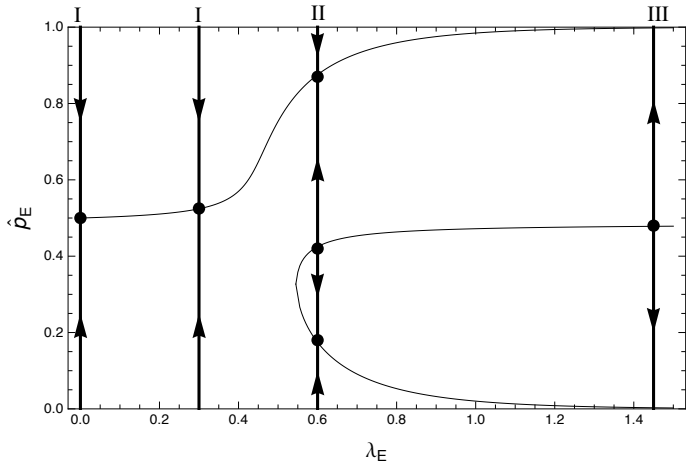

Figure 2.6: Bifurcation diagram for the differential equation 2.20 that describes the learning behavior of ERL in the producer-scrounger model as a function of λ_E. The thin curves describe the equilibrium values of $\hat{p}_E$ and the the thick vertical lines are phase lines at the corresponding values of λ_E. Dots on the phase lines denote interior equilibria. Our numerical exploration suggests that there are three possible phase lines depending on the value of λ_E (indicated by I, II, and III). Parameter values: $s = \left(3-\sqrt{5}\right)/2$, $V(1) = 5.3$, $V(2) = 5$.

Payoff functions

In order to derive the fecundity (or payoff) functions of the three types (ERL, FP, and S), we make the assumption that learners have reached the equilibrium behavior described in the previous section ($\hat{p}_F$ given eq. 2.19 for FP and $\hat{p}_E$ given by the expectation over the various equilibria like in eq. 2.22). We further denote by q the frequency of FP, so that $1-q-s$ gives

the frequency of ERL. With probability $\hat{p}_i$ ($i \in \{\mathbf{E},\mathbf{F}\}$), a learner goes to patch 1, while with probability $1-\hat{p}_i$, it goes to patch 2. The fecundity (or payoff) of the two learners is then given by

$$\begin{aligned} b_{\mathbf{F}} &= \alpha + \hat{p}_{\mathbf{F}}\left(\frac{2-s}{2}\right)V(1) + (1-\hat{p}_{\mathbf{F}})\left(\frac{2-s}{2}\right)V(2) - k, \\ b_{\mathbf{E}} &= \alpha + \hat{p}_{\mathbf{E}}\left(\frac{2-s}{2}\right)V(1) + (1-\hat{p}_{\mathbf{E}})\left(\frac{2-s}{2}\right)V(2) - k, \end{aligned} \tag{2.23}$$

where k is the cost of individual learning and we assumed that all individuals have a baseline reproductive output of α.

Because we assumed that only a single scrounger can follow a producer, the expected frequency of interactions of a scrounger with a producer is proportional to $(1-s)/s$, and the fecundity of a scrounger is assumed to be given by

$$\begin{aligned} b_{\mathbf{S}} = \alpha + \frac{1-s}{s}\Bigg[\frac{q}{1-s}\left(\hat{p}_{\mathbf{F}}\frac{V(1)}{2} + (1-\hat{p}_{\mathbf{F}})\frac{V(2)}{2}\right) \\ + \frac{1-q-s}{1-s}\left(\hat{p}_{\mathbf{E}}\frac{V(1)}{2} + (1-\hat{p}_{\mathbf{E}})\frac{V(2)}{2}\right)\Bigg]. \end{aligned} \tag{2.24}$$

This entails that scroungers have no preference for FP or ERL. They follow an FP with a probability $q/(1-s)$ and follow an ERL with the complementary probability $(1-q-s)/(1-s)$. When a scrounger follows a learner of type i on patch a, the scrounger gets half of the value of the patch, $V(a)/2$. This learner goes to patch 1 with a probability $\hat{p}_i$, or goes to patch 2 with probability $1-\hat{p}_i$, hence the expected payoff to a scrounger conditional on the event that it follows a learner of type i is $\hat{p}_i[V(1)/2] + [1-\hat{p}_i][V(2)/2]$.

ESS analysis

With the above assumptions, the change in frequencies of the types after one generation is given by

$$\begin{aligned} \Delta q &= q\left(b_{\mathbf{F}} - \bar{b}\right)/\bar{b} \\ \Delta s &= s\left(b_{\mathbf{S}} - \bar{b}\right)/\bar{b}, \end{aligned} \tag{2.25}$$

where $\bar{b} = q\,b_{\mathbf{F}} + s\,b_{\mathbf{S}} + (1-q-s)b_{\mathbf{E}}$ is the mean reproductive output in the population. The evolutionary dynamics (eq. 2.25 with eqs. 2.23–2.24) displays five stationary states, which we write under the form $(q^*, s^*, 1-q^*-s^*)$. There are the three trivial equilibria $[(1,0,0),(0,1,0),(0,0,1)]$, one equilibrium with a coexistence between FP and scroungers, $(\frac{1}{2}\left(\sqrt{5}-1\right), \frac{1}{2}\left(3-\sqrt{5}\right), 0)$,

and one with a coexistence between ERL and scroungers at the same frequency as in the previous case: $(0, \frac{1}{2}\left(3-\sqrt{5}\right), \frac{1}{2}\left(\sqrt{5}-1\right))$. Because the payoff to scroungers exceeds that of producers when they are in low frequency $s \to 0$ (for $V(1) > 0$ and/or $V(2) > 0$), the two equilibria where there is a mix between scroungers and producers are stable in a reduced 2-strategy dynamics (on the faces of the simplex). Hence, the three trivial equilibria are unstable. The question then is which one of the two other equilibria obtains. Because the fecundity of each type of producer does not depend on the other type and in the same way on the frequency of scroungers (eq. 2.23), the mix between scroungers and FP is invaded by ERL if they produce more resources. Namely, if the latter visit more often the optimal patch, which obtains if

$$\hat{p}_{\mathrm{E}} > \hat{p}_{\mathrm{F}}. \tag{2.26}$$

This invasion condition is not necessarily satisfied when $\lambda_{\mathrm{E}} > \lambda_{\mathrm{F}}$, and in Fig. 2.7 we display the regions of values of λ_{F} and λ_{E} where it is satisfied. These regions seem to alternate in a non-trivial way. Interestingly, the region where ERL outcompetes FP looks fairly large for our parameter values. When λ_{E} becomes very large, it is possible to have an exact invasion condition by substituting eq. 2.19 and eq. 2.22 into eq. 2.27, which implies that ERL invades the stable mix of FP and S if and only if

$$\lambda_{\mathrm{F}} < \frac{2\log\left(\frac{V(1)}{V(2)}\right)}{\left[V(1)-V(2)\right]\left(1+\sqrt{5}\right)/2}. \tag{2.27}$$

Summing up the above analysis, there is a globally stable state for the 3-strategy replicator dynamics in this producer-scrounger model that is the mix between scroungers and the most performant producer. In this unique evolutionarily stable state, producers are in frequency $\left(\sqrt{5}-1\right)/2$. Which of the producer type will be maintained in the population (FP or ERL) critically depends on the exploration rate (λ_{F} and λ_{E}). It is noteworthy that it is not the learner with the highest value of λ_i that will invade. The main reason for this is that increasing λ_{E} for ERL does not always leads to a higher probability of visiting patch 1. When λ_{E} is relatively small, this is actually true (in regime I of the learning dynamics of ERL, Fig. 2.6) but when λ_{E} grows (regimes II and III), ERL suddenly becomes prone to absorption in a state where it visits patch 2 with a probability greater than 0.5 ($\hat{p}_{\mathrm{E}} < 0.5$). This makes ERL less performant than FP for high values of λ_i (the upper-right region in Fig. 2.7). Further, when λ_i is very small (the lower-left region in Fig. 2.7), ERL seems to be less sensitive than FP to an increase in λ_i.

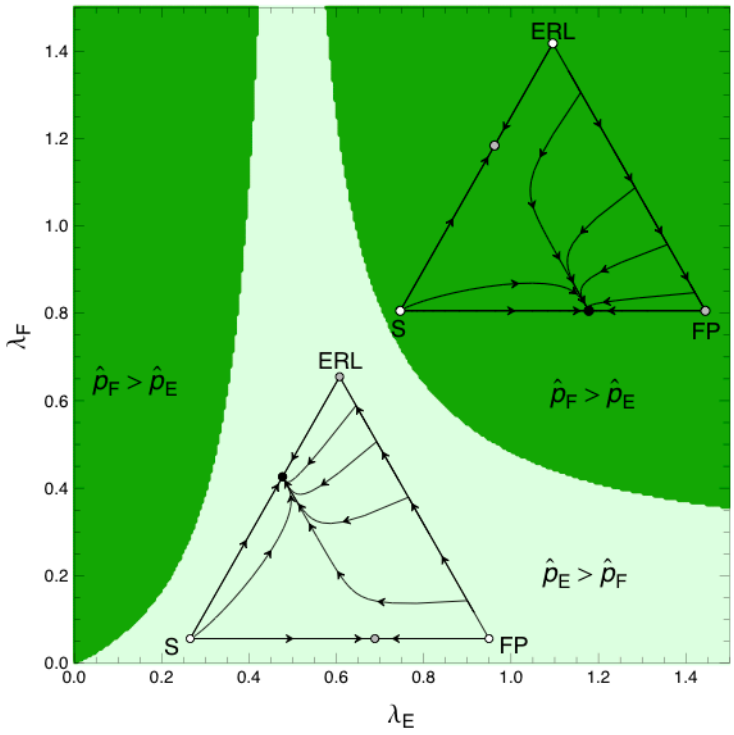

Figure 2.7: Solution orbits of the evolutionary dynamics (eq. 2.25) in the producer-scrounger model on the 3-strategy simplex as a function of λ_E and λ_F. In the light shaded region, ERL is the most performant ($\hat{p}_E > \hat{p}_F$), while in the dark shaded region, FP is the most performant ($\hat{p}_F > \hat{p}_E$). The simplex drawn in the light region is plotted for $\hat{p}_E > \hat{p}_F$ and hence verifies that the mix between ERL and scroungers is the unique ESS. The simplex in the dark region corresponds to the case where the unique ESS is the mix between FP and scroungers. At the corner labeled FP on the simplices, we have $(q = 1, s = 0)$, at the corner ERL we have $(q = 0, s = 0)$ and at the corner S we have $(q = 0, s = 1)$. A white-filled dot denotes an unstable node, a gray-filled dot is a saddle, and the black dot corresponds to the unique ESS. These simplices were produced using the Baryplot package (McElreath, 2010) for R (R Development Core Team, 2011). Parameter values for the shading: $s = \left(3 - \sqrt{5}\right)/2$, $V(1) = 5.3$, $V(2) = 5$.

2.4 Discussion

In this paper, we used stochastic approximation theory (Ljung, 1977; Benveniste et al., 1991; Fudenberg and Levine, 1998; Benaim, 1999; Kushner and Yin, 2003; Sandholm, 2011) in order to analyze the learning of actions over the course of an individual's lifespan in a situation of repeated social interactions with environmentally induced fluctuating game payoffs. This setting may represent different ecological scenarios and population structures, where interactions can be represented as an iterated N-person game or a multi-armed bandit. The learning dynamics was assumed to follow the experience-weighted attraction (EWA) learning mechanism (Camerer and Ho, 1999; Ho et al., 2007). This is a motivational-based learning process, which encompasses as special cases various learning rules used in biology such as the linear operator (McNamara and Houston, 1987; Bernstein et al., 1988; Stephens and Clements, 1998), relative payoff sum (Harley, 1981; Hamblin and Giraldeau, 2009) and Bayesian learning (Rodriguez-Gironés and Vásquez, 1997; Geisler and Diehl, 2002).

When a behavioral process has a decreasing step-size (or a very small constant step-size), stochastic approximation theory shows that the behavioral dynamics is asymptotically driven by the expected motion of the original stochastic recursions. Stochastic approximation is thus appealing because once the expected motion of the stochastic learning process is derived, one is dealing with deterministic differential equations that are easier to analyze. Further, the differential equations governing action play probabilities under the EWA model that we have obtained (eq. 2.13) have a useful interpretation. They show that learning is driven by a balance between two forces. First, the exploration of actions that tends to bring the dynamics out of pure states, which is analogous to mutation in evolutionary biology. Second, the exploitation of actions leading to higher expected reinforcement, which is analogous to selection in evolutionary biology. This second part actually takes the same qualitative form as the replicator equation (eq. 2.13), since actions leading to an expected reinforcement higher than the average expected reinforcement will have a tendency to be played with increased probability. Although it may be felt in retrospect that this result is intuitive, it is not directly apparent in the original stochastic recursions of the behavioral rule, which encompasses parameters tuning the levels of cognition of individuals (eq. 2.1).

Our model is not the first where analogues of replicator dynamics appear out of an explicit learning scheme (e.g., Börgers and Sarin, 1997; Hopkins, 2002; Tuyls et al., 2003; Hofbauer and Sigmund, 2003). But, apart from Hopkins (2002), we are not aware of results that link the repli-

cator dynamics to reinforcement learning and belief-based learning at the same time, which was extended here to take fluctuating social environments into account. Although we considered only individual learning without environmental detection in our formalization (i.e., individuals learn the average game), the reinforcement of motivations could take social learning into account (e.g., Cavalli-Sforza and Feldman, 1983; Schlag, 1998; Sandholm, 2011), and/or individuals may detect changes in the environment so that the motivations themselves may depend on environmental states (e.g., evaluate the dynamics of motivations $M_t(a,\omega)$ for action-state pairs). The consequences of incorporating these features for action ontogeny may be useful to analyze in future research.

We applied our results to analyze the dynamics of action play probabilities in a situation of repeated pairwise interactions in a 2×2 fluctuating game with average Hawk-Dove payoffs, where we investigated interactions between different learning rules, a situation that is very rarely addressed analytically (but see Leslie and Collins, 2005; Fudenberg and Takahashi, 2011). Comparison with stochastic simulations of the original learning dynamics indicate that the deterministic dynamics generally approximates qualitatively well the quasi-equilibrium of action play probabilities obtained under the original stochastic process. Even if the theory can only prove that the stochastic approximation of processes with decreasing step-sizes "works" when time becomes very large (the differential equation are guaranteed to track the solutions of the stochastic process only asymptotically, Benaim, 1999), our simulations suggest that stochastic approximation can, under good circumstances, give fair predictions for finite-time behavior (in our case, for $T = 100$, 500, and 1000), and also for the ontogeny of behavior (Fig. 2.8). This may be useful in the context of animal behavior, when lifespan is short.

We also observed one limitation associated with using stochastic approximation in our examples. Namely, there are situations that are not captured by the deterministic approximation. These involve the cases where the sensitivity to payoff (λ) and the difference between initial motivations ($\Delta M_{i,1}$) are very small so that actions are random, and the cases where λ and $\Delta M_{i,1}$ are very big so that the dynamics get stuck in suboptimal equilibria. In particular, when $\Delta M_{i,1}$ is very large, individuals may have to reverse their initial preferences and this makes very large the time for which the stochastic learning process gets close to its asymptotic approximation.

Finally, we applied our results to analyze the evolutionary competition between learners and scroungers in a producer-scrounger game, where we considered that learners are producers (who search and find good patches of food) and scroungers follow the producers. Three

types were present in the population: individuals who learn according to Exploratory Reinforcement Learning, individuals who learn according to Stochastic Fictitious Play (Table 2.1), and scroungers. This evolutionary model leads, at the ESS, to the co-existence of scrounger with the most performant of the two learning rules. In particular, we showed that the exploration rate (λ_i) influences which is the most performant producer, but the effect of λ_i is non-linear. This shows that different learning rules are very differently affected by varying the exploration rate. The exploration rate and the choice rule (eq. 2.4) thus makes part of the definition of a learning rule, and λ_i may interact in a non-intuitive way with the other parameter of the process that affect motivation updating.

While in this paper we analyzed certain learning rules with decreasing step-size, it remains an open empirical question to document how common this type of learning rules are in nature. It seems that previous work in animal psychology and behavioral ecology focused more on rules with constant step-sizes (e.g., the linear operator, Bush and Mostelller, 1951; Rescorla and Wagner, 1972; Hamblin and Giraldeau, 2009; Arbilly et al., 2010) because the step-size has here a clear interpretation in terms of a discount factor (or learning rate) and takes into account known phenomena such as habituation or forgetting. But it will be relevant to determine how well rules with decreasing step-size fit animal behavior. In particular, we suspect that such behavioral rules could describe accurately learning processes where early experience is critical to shape general behavior and where further information is used only to fine tune actions (e.g., developmental processes) and where preference reversal becomes unlikely.

In summary, although we illustrated some shortcoming of applying stochastic approximation, we showed that it can be a useful approach to learn about learning dynamics and to avoid "the behavioral gambit" (Fawcett et al., 2013; Lotem, 2013). But even if action play probabilities can be approximated by differential equations, there are many aspects of the concomitant dynamics that we did not analyze here, and that are likely to be relevant in the context of animal learning. This opens paths to future work, which could for instance analyze rules with constant step-size, produce finite-time predictions for play probabilities, evaluate the effect of learning speed on payoff under different patterns of environmental fluctuations, or investigate state-dependent motivations. Studying these aspects may be relevant to better understand learning dynamics and behavioral ontogeny.

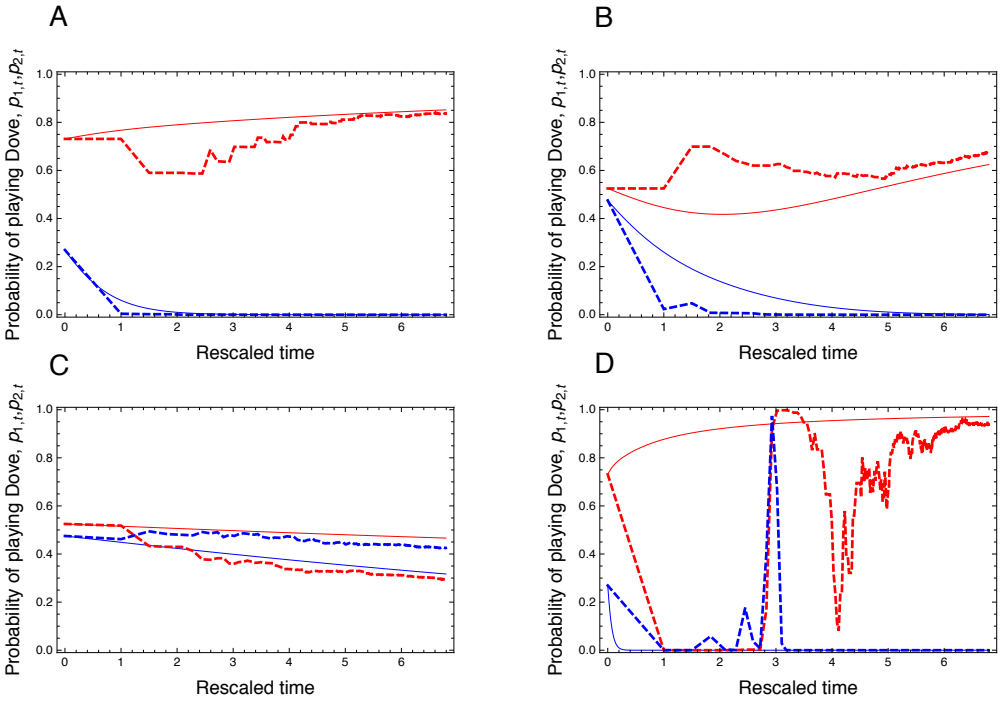

Figure 2.8: Comparison between the deterministic (thin, plain lines) and the stochastic (thick, dashed lines) time dynamics of playing Dove for a PRL meeting an IL, and for different values of λ_i and $\Delta M_{i,1}$. The blue line is for the PRL while the red line is for the IL. We always set opposed initial motivations to the players ($\Delta M_{1,1} = -\Delta M_{2,1}$) and $\lambda_1 = \lambda_2$, while the parameters of the game are $B = 5$ and $C = 3$. (A) $\Delta M_{1,1} = 1$ and $\lambda_i = 1$. (B) $\Delta M_{1,1} = 10^{-1}$ and $\lambda_i = 1$. (C) $\Delta M_{1,1} = 1$ and $\lambda_i = 10^{-1}$. (D) $\Delta M_{1,1} = 10^{-1}$ and $\lambda_i = 10$. We simulated the process for $T = 500$ interactions and the time on the x-axis is measured on the timescale of the interpolated stochastic process (the τ_n in Benaim, 1999), which is used to plot the numerical solution of the differential equations.

Acknowledgments

We are grateful to Mathieu Faure and Michel Benaïm for their generous answers to many questions about stochastic approximations. This work was supported by Swiss NSF grant PP00P3-123344.

Chapter 3

Does selection favor the use of inferred payoffs over reinforcement learning in social interactions?

Abstract

Different species are able to learn to associate behaviors with rewards as this gives fitness advantages in changing environments. However, social interactions might require more cognitive abilities than simple reinforcement learning, in particular the capacity to infer the material payoff consequences of different action combinations. It is unclear whether natural selection necessarily favors individuals that are able to use inferred payoffs as opposed to simple reinforcement of realized payoff (using trial-and-error) when social interactions occur between population members. Here we develop an evolutionary model where individuals are genetically determined to use either reinforcement learning or inference-based learning, and ask what is the evolutionarily stable rule in a situation of pairwise symmetric two actions stochastic games played during lifespan. We analyze through stochastic approximation theory and simulations the learning dynamics at the behavioral timescale, and derive conditions where reinforcement learning outcompetes inference-based learning at the evolutionary timescale when repeated interactions occur between pairs of individuals. By contrast, we find that inference-based learners tend to be favored under random interactions, but stable polymorphisms can also obtain where reinforcement learners are maintained at a low frequency. We conclude that specific game structures can select for trial-and-error even in the absence of cost of cognition, which hints at none gradual fitness benefits of "complex" cognition.

3.1 Introduction

Many species have a learning ability because this allows an individual to adapt, within its lifespan, to the currently fitness-relevant features of its environment (e.g., by tracking the location of food patches; Charnov, 1976; Mcnamara and Houston, 1985; Shettleworth et al., 1988). Hence, learning is likely to provide a selective advantage (Johnston, 1982; Stephens, 1991; Mery and Kawecki, 2002; Wakano et al., 2004; Dunlap and Stephens, 2009).

In terms of behavior, learning can be defined as the process of gathering information about the biotic and abiotic environment and using it to choose actions. It is commonly accepted that the simplest way of learning a behavior is through trial-and-error, which is also called instrumental or reinforcement learning (Thorndike, 1911; Bush and Mostelller, 1951). This consists in trying different actions, experiencing the rewards associated to each action, and repeating more often the actions yielding higher rewards (or, equivalently, avoiding actions that yield negative payoffs, or punishments). For example, rats in the Skinner box learn that pressing a lever is associated with obtaining food, and various instances of reinforcement learning in other mammals, birds, fish, and insects have been demonstrated (Shettleworth, 2009; Dugatkin, 2010).

Although reinforcement learning is the main paradigm for describing the learning of actions in animals (Dickinson, 1980; Shettleworth, 2009; Dugatkin, 2010), it cannot solve all decisions problems. In effect, with this behavioral rule, an individual has to physically try (or experience) an action to get the knowledge of the reward (or payoff) associated to it. Thus, an animal who uses only reinforcement learning risks to spend all its time trying and exploring without ever finding a good (rewarding) action. It also risks to face the dangers inherent to imprudent explorations (e.g., to meet a predator). This problem of reinforcement learning in an uncertain environment is often called the exploration-exploitation trade-off and has been investigated by students of learning for a long time (Arnold, 1978; Mcnamara and Houston, 1985; Shettleworth et al., 1988; Krebs et al., 1993; Sutton and Barto, 1998; Achbany et al., 2006).

Certain species circumvent the exploration-exploitation problem through cognitive abilities allowing them to mentally simulate and reason about potential solutions to the ecological problems they are facing (Emery and Clayton, 2004; Taylor et al., 2012). Mental simulation, or imagination, is the ability an organism has to represent in its own brain the outcome of a situation that is not physically present. With mental simulations, one can try solutions and infer material consequences without spending energy and time, or taking the risk of getting caught by a predator. Humans use imagination to plan courses of actions, or to better anticipate the behav-

ior of others (theory of mind); chimpanzees are able of role taking behaviors, so that they can find a solution to a problem in which another individual is involved (Premack and Woodruff, 1978); scrub jays display sophisticated food caching, where they hide food only if they are sure that no other conspecific is observing, which implies that they expect that others can steal their food (Emery and Clayton, 2001). According to Emery and Clayton (2004), the ability to mentally simulate outcomes of possible actions is one of the features that characterize species with "complex cognition", among which we find Corvids, Apes, and Humans.

The examples provided above all involve social interactions. While the usefulness of a learning ability in social interactions is well documented across species (Chalmeau, 1994; Hollis et al., 1995; Villarreal and Domjan, 1998; Plotnik et al., 2011), some authors go even further by arguing that the evolution of "complex" learning and cognition may have been a response to the emergence of "complex" social interactions (Humphrey, 1976; Alexander, 1979; Gavrilets and Vose, 2006; see also the references in Emery and Clayton, 2009). This hypothesis is to take with care (how should we define "complex" interaction and cognition? Shettleworth, 2009, Chap. 12), yet there is an intriguing related and simpler question: is reinforcement learning sufficient to learn in social interactions or will a learning rule based on payoff inference be necessarily favored by natural selection?

Since inference would allow an individual to have a mental representation of the payoff consequences of its own actions and that of its interacting partners, it may be felt that this ability should necessarily have a selective advantage over simple reinforcement learning, since it can lead to a better anticipation and response to opponents' actions. However, this intuition has never been given empirical or theoretical support as the effect of natural selection on cognition in a social context is rarely addressed (for exceptions see Heller, 2004; Arbilly et al., 2010; Mohlin, 2012). Our aim in this paper is then to determine under what conditions does natural selection favors inference-based learning versus trial-and-error in pairwise social interactions where the social environment is itself changing.

We will consider two learning rules, the first of which is standard reinforcement learning (Thorndike, 1911; Bush and Mostelller, 1951; Rescorla and Wagner, 1972; Stephens and Clements, 1998). The second rule we call inference-based learning and it allows an individual to infer forgone payoffs given its actions and that of its partners, and it is related to the class of belief-learning rules that have been extensively studied and is a special case of the experience-weighted attraction learning mechanisms (Fudenberg and Levine, 1998; Camerer and Ho, 1999;

Hopkins, 2002; Camerer, 2003). Inference-based learning allows an individual to infer from realized actions, the payoffs consequences of alternative courses of actions to maximize new rewards accordingly. For instance, an inference based learner interacting with a cooperator in the context of a Prisoner's Dilemma game will infer the forgone payoffs of playing defect instead of cooperate, and thus learn to play defect. Inference-based learning relies on imagination because an individual must be able to compute the payoff of an action it did not explicitly try given the belief. In other words, for a given action taken by an opponent, an inference-based learner must be able to mentally represent what is his payoff for each action it can take. This potential payoff can be termed an hypothetical or inferred reinforcement, and it has been shown that belief-based learning can be derived from a model where individuals are able to compute inferred reinforcements (Camerer and Ho, 1999; Ho et al., 2007).

The behavior of individuals learning according to reinforcement or inference-based learning have been studied in isolation (i.e., either interactions between reinforcement learners only or between inference-based learners only, Börgers and Sarin, 1997; Fudenberg and Levine, 1998; Hopkins, 2002; Young, 2004) but, few studies have pitted these rules agains each other. A notable exception is Josephson (2008) who let compete belief-based with reinforcement learning in an evolutionary model with different standard games like the Prisoner's Dilemma or the Coordination game, and found that belief-based learning almost always outcompete reinforcement learning. Interestingly in Josephson's work, reasonably often both rules did not lead to fundamentally different behavior but the main difference was in the speed of learning: reinforcement learners were slower than belief-based learners (see also Hopkins, 2002). Using a different approach, Dridi and Lehmann (2013) found that the evolutionarily stable learning rule in the context of a producer-scrounger game depended on the exploration tendency of the learners and, surprisingly, it was not the rule that explored less that was the evolutionarily stable one, even in a very simple task only requiring the identification of a best action among two options. These contrasting results indicate that there is a need to better understand the environmental and social conditions that favor inference-based over reinforcement learning, and where these cognitive mechanisms can lead to different behavior.

Models studying the evolutionary competition between individual learning rules in social interactions are rare (Harley, 1981; Josephson, 2008; Hamblin and Giraldeau, 2009; Arbilly et al., 2010), and often do not consider the evolution of the learning rules under the very conditions where they are selected for: fluctuating environments (Stephens, 1991; Wakano et al., 2004; Dunlap and Stephens, 2009). In this paper, we extend the approach of Josephson (2008) and study the

evolutionary stability of reinforcement learning or inference-based learning in a situation where the payoffs of these social interactions (games) fluctuate during lifespan. We also study variants of reinforcement learning and inference-based learning that rely on the same dichotomy between trial-and-error and ability to mentally simulate non-realized payoffs. For three different types of fluctuating games, we study the evolutionary stability of these two learning rules. We analyze two ways of forming pairs of opponents from the population. In the first one, individuals are randomly paired at the beginning of lifespan and each formed pair interacts for the rest of lifespan. In the second matching scheme, there is random matching at each time step during lifespan so the individuals learn to play against the whole population. Because the underlying model is stochastic, we use, when possible, a deterministic approximation to analyze the equations and then compare this with individual-based simulations.

3.2 Model

3.2.1 Setting the stage

Population

We consider a haploid population of constant size N. The main life cycle stages are the following. (1) Each individual interacts socially with others for T time periods. (2) Each individual reproduces according to its gains and losses incurred during social interactions. (3) All individuals of the parental generation die and N individuals from the offspring generation are sampled to form the new adult generation.

Repeated game affected by environmental fluctuations

During stage (1), individuals play a game at each time period $t = 0, 1, 2, \ldots, T$, where the game that is played is determined by some environmental state, ω, which itself belongs to a set of environmental states, Ω. The environment can be anything that alters the payoffs associated to actions taken by individuals. We assume that the environmental process follows an ergodic Markov chain (Karlin and Taylor, 1975), and denote by $\mu(\omega)$ the stationary probability that state ω obtains.

For example, one can consider an environment in which individuals play alternatively two games, e.g., a Prisoner's Dilemma and a Hawk-Dove game; the set of environmental states is then

$\Omega = \{\text{Prisoner's Dilemma, Hawk-Dove}\}$. In this example, one could set $\mu(\text{Prisoner's Dilemma}) = \mu(\text{Hawk-Dove}) = 1/2$, meaning that it is as if we toss a fair coin at each time t to determine the game to be played. If the current environmental state is $\omega = \text{Prisoner's Dilemma}$, then all individuals in the population will have to choose between cooperating and defecting.

More generally, we consider that all the games in Ω consist of the same number of actions, say m (in our previous example we had $m = 2$). Hence, in every period of time, each organism in the population chooses its action from a fixed finite set of actions $\mathscr{A} = \{1, \ldots, m\}$, and we denote by ω_t the game played at time t, which is a random variable. The action taken by individual i at time t is also a random variable denoted by $a_{i,t}$ (we will allow individuals to use probabilistic action choice) and the action profile in the population is $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$ (this is the collection of the actions of all individuals in the population at time t). The payoff to individual i at time t when it takes action $a_{i,t}$ and the game is ω_t is denoted $\pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \omega_t)$, where $\mathbf{a}_{-i,t}$ is the action profile of the remaining individuals in the population (all individuals excluding i).

With this, we define the fecundity, b_i, of individual i as its mean payoff obtained during the whole sequence of interactions:

$$b_i = \frac{1}{T} \sum_{t=1}^{T} \pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \omega_t), \tag{3.1}$$

which provides its number of offspring produced during stage (2) of the life cycle.

3.2.2 Learning actions

In order to evaluate the fecundity, b_i, of individual i, we need to know how actions are taken by learning. We now present a model of learning that takes both reinforcement learning and inference-based learning into account.

Action choice

Our way of modeling learning is shared by many previous studies and relies upon two components: (1) dynamic preferences for action, and (2) a rule for choosing action given preferences (Harley, 1981; Camerer and Ho, 1999; Leslie and Collins, 2005; Ho et al., 2007; Hamblin and Giraldeau, 2009; Arbilly et al., 2010, 2011*b*; Dridi and Lehmann, 2013). Specifically, we let individuals having preferences or motivations for actions that they update through the repeated play of the game according to payoffs. For each action a in its behavioral repertoire $\mathscr{A}$, individual i has an associated motivation $M_{i,t}(a)$ that represents how much action a is valued by individual

i. Thus, the motivations can be thought as the states of the organism (Enquist and Ghirlanda, 2005; Niv, 2009) and we assume that action a is chosen at time t by individual i with probability

$$p_{i,t}(a) = \frac{\exp[\lambda M_{i,t}(a)]}{\sum_{k\in\mathscr{A}} \exp[\lambda M_{i,t}(k)]}. \tag{3.2}$$

This is a standard choice rule (Anderson et al., 1992; McKelvey and Palfrey, 1995; Fudenberg and Levine, 1998; Camerer and Ho, 1999; Ho et al., 2007; Arbilly et al., 2010, 2011*b*), where the action that has the highest motivation is chosen with the greatest probability. The parameter $\lambda \in [0,\infty)$ represents the sensitivity of an animal to its motivations. If λ is near 0, the animal is not very reactive to its motivations and has a tendency to explore (because $p_{i,t}(a)$ is close to $1/m$). If λ is high, we have an animal being "greedy", taking almost surely the action that has the highest motivation (if action a^* has the highest motivation, $p_{i,t}(a^*)$ is close to 1, see Dridi and Lehmann (2013) and references therein for more justifications underlying the use of eq. 3.2).

Motivations

The motivations of an individual are updated after each interaction stage, t. In order to achieve this updating, we use a special case of the EWA model of Camerer and Ho (1999) and its application to stochastically varying environments (Dridi and Lehmann, 2013). Individual i starts off with some initial preferences over actions at time $t = 1$ given by the initial motivations $M_{i,1}(a)$ for all actions a. We assume that the motivation $M_{i,t+1}(a)$ for action a of individual i at time $t+1$ is given by

$$M_{i,t+1}(a) = \frac{t}{t+1}\phi_{i,t}M_{i,t}(a) + \frac{1}{t+1}\{\delta_i + (1-\delta_i)\mathbb{1}(a,a_{i,t})\}\pi_i(a,\mathbf{a}_{-i,t},\omega_t). \tag{3.3}$$

Eq. 3.3 can be seen as a weighted average of the previous motivation $M_{i,t}(a)$ and of the new payoff $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$. In the first term, the motivation $M_{i,t}(a)$ is weighted by $\phi_{i,t} \geq 0$, a memory parameter, or learning rate, that indicates the relative importance of the last motivation as opposed to the current payoff (note that $\phi_{i,t}$ can change as a function of time and also has an initial value $\phi_{i,1}$, possibly genetically determined). This first term is also weighted by $t/(t+1)$, which entails that the previous motivation is weighted according to the number of interactions that have occurred up to time t.

The second term can be termed the increment, or reinforcement to the motivation. It has weight $1/(t+1)$ and depends on the payoff $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$ of action a when all other individuals in the population play $\mathbf{a}_{-i,t}$ and the game is in state ω_t at time t. The expression $\mathbb{1}(a,a_{i,t})$ is an

indicator function, which is

$$\mathbb{1}(a,a_{i,t}) = \begin{cases} 1, & \text{if } a_{i,t} = a, \\ 0, & \text{otherwise,} \end{cases} \tag{3.4}$$

i.e., if individual i takes action a at time t, then $\mathbb{1}(a,a_{i,t}) = 1$. We see that if $\mathbb{1}(a,a_{i,t}) = 1$, the numerator of the second term of eq. 3.3 reduces to $\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$. If, on the other hand, individual i does not play a at time t, then the numerator of the second term reduces to $\delta_i\pi_i(a,\mathbf{a}_{-i,t},\omega_t)$.

The parameter δ_i weights the ability of an individual to infer the payoffs of unchosen actions (non-realized or foregone payoffs) and reinforce motivations accordingly. If $\delta_i = 0$, the individual is not able to infer non-realized payoffs and thus only reinforces actions according to realized payoff. We will call this Reinforcement Learning (RL). By contrast, if $\delta_i = 1$, the payoffs associated to unchosen actions are always perfectly inferred. This will be called Inference-based learning (IL), where individuals have the capacity to access information about (or compute) non-realized payoffs. When an Inference-based learner plays action a at time t, it reinforces not only action a but also all other actions according to the payoffs they would have yielded (see Appendix B.1). This information about missed payoffs may be available if the organism can infer its potential payoffs from environmental cues (i.e., it observes the column corresponding to the action of its opponent in the payoff matrix of the game, but *a posteriori*) or if it can observe the payoffs of individuals involved in other interactions. A well-studied special case of IL obtained when $\phi_i = 1$ and $\delta_i = 1$ is Belief-based learning, where the motivations represent expected payoffs given the history of play by individual's i opponents (Brown, 1951; Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002; Hopkins, 2002).

Note that the motivations of all actions are updated by individual i after each time t. Consequently, an Inference-based learner [with $\delta_i = 1$], who observe foregone payoffs, computes as many payoffs as there are available actions (m), at each period of time t. On the other hand, a Reinforcement Learner [with $\delta_i = 0$] has a much lighter computational task because it computes only the payoff of the action it actually took. Thus, we will postulate in the analysis below that the cost of additional computations will affect the fitness of IL by an amount k.

3.2.3 Evolutionary analysis

Assumptions

Equations 3.2–3.3 define the learning dynamics of action a for individual i. Our aim is to investigate the co-evolution of reinforcement and belief-based learning under these dynamics. To that aim, we consider that the parameter δ_i is the genotype of individual i, and that there can be only two types of individuals in the population, reinforcement learners with $\delta_i = 0$, and belief-based learners with $\delta_i = 1$. We make the following simplifying assumptions in order to investigate these dynamics.

(a) We consider two special values of $\phi_{i,t}$, the learning rate. First, we are interested in the case $\phi_{i,t} = (1/t)+1$, where analytical results can be derived about the learning behavior (see next section). For this special value of the learning rate, we obtain a special special version of RL, which we call Pure Reinforcement Learning (PRL), and a special version of IL, which we call Pure Inference-based Learning (PIL). These names follows from the fact that at a behavioral equilibrium individuals will express essentially only pure actions.

Second, we use a constant value $\phi_i = 1$ for the learning rate, which entails that individuals are likely to be more exploratory at a behavioral equilibrium and are likely to express mixed actions. For this value of the learning rate, we obtain a special special version of RL, which we call Exploratory Reinforcement Learning (ERL), and a special version of IL, which we call Exploratory Inference-based Learning (EIL), where the latter strategy corresponds to Belief-based learning of the game theory literature (it is also called Stochastic Fictitious Play, as defined in Chapter 2; Fudenberg and Levine, 1998).

In Appendix B.1, we provide more details on the dynamics of the motivations (eq. 3.3) of these learning rules, and in our analysis we always consider competition between RL and IL under the same learning rate.

(b) We assume pairwise interactions and we analyze two matching rules. The first we call one-shot matching, where individuals are randomly paired at the beginning of the game ($t = 1$) and each pair interacts together for the whole duration of the game (until time T and reproduction). Second, we consider repeated matching, where individuals are re-matched during each stage game, i.e., individuals meet different partners at each time t.

(c) All games consist of only two actions ($m = 2$), that is, individuals play 2×2 symmetric games. This means that, at each time t, individuals play one out of three types of games: a

Prisoner's Dilemma (PD), a Hawk-Dove game (HD), or a (pure) Coordination Game (CG); i.e., the set of games is $\Omega = \{\text{PD}, \text{HD}, \text{CG}\}$, and the stationary distribution of games thus satisfy $\mu(\text{PD}) + \mu(\text{CG}) + \mu(\text{HD}) = 1$. These three games are instances from the three possible categories of 2×2 games (the PD is a game with a dominant action, the HD is a game with two pure asymmetric Nash equilibria, and the CG is a game with two pure symmetric equilibria, Weibull, 1997, Chap. 1). The payoffs for game ω are written $\mathscr{R}_\omega$, $\mathscr{S}_\omega$, $\mathscr{T}_\omega$, and $\mathscr{P}_\omega$, where $\omega \in \{\text{PD}, \text{CG}, \text{HD}\}$ and where for example $\mathscr{T}_\omega$ is the payoff obtained by both players when they choose action 2 and the game is ω. Similarly, $\mathscr{R}_\omega$, $\mathscr{S}_\omega$, and $\mathscr{P}_\omega$ describe the payoffs for each other possible combination of actions by the players (Table 3.1).

Table 3.1: Payoff matrix for a typical stage game ω, where $\omega \in \{\text{PD}, \text{CG}, \text{HD}\}$. One player chooses a row and its opponent chooses a column. Payoffs are to row player. In order to numerically implement the three possible sub-games $(\text{PD}, \text{CG}, \text{HD})$ we used the following constraints on the payoffs. In the Prisoner's Dilemma game (PD): $\mathscr{T}_{\text{PD}} > \mathscr{R}_{\text{PD}} > \mathscr{P}_{\text{PD}} > \mathscr{S}_{\text{PD}}$ and $(\mathscr{T}_{\text{PD}} + \mathscr{S}_{\text{PD}})/2 < \mathscr{R}_{\text{PD}}$. In the Hawk-Dove game (HD): $\mathscr{T}_{\text{HD}} > \mathscr{R}_{\text{HD}}, \mathscr{S}_{\text{HD}} > \mathscr{P}_{\text{HD}}, \mathscr{P}_{\text{HD}} > \mathscr{R}_{\text{HD}}$. In the Coordination Game (CG): $\mathscr{R}_{\text{CG}} > \mathscr{S}_{\text{CG}}, \mathscr{R}_{\text{CG}} = \mathscr{P}_{\text{CG}}, \mathscr{S}_{\text{CG}} = \mathscr{T}_{\text{CG}}$.

	Action 1	**Action 2**
Action 1	$\mathscr{R}_\omega$	$\mathscr{S}_\omega$
Action 2	$\mathscr{T}_\omega$	$\mathscr{P}_\omega$

Stochastic approximation

Although eqs. 3.2–3.3 describe a bona fide learning process, this is a non-homogeneous multidimensional Markov process that is very difficult to analyze (see Fig. 2.1). It is thus necessary to approximate it in order to obtain analytical results, which is useful to form an intuition about behavioral dynamics. It turns out that this analysis is possible only under the one-shot matching scheme and when the learning rate $\phi_{i,t}$ takes the dynamic value $\phi_{i,t} = (1/t) + 1$.

Using the above assumptions (a-c) with $\phi_{i,t} = (1/t) + 1$ and stochastic approximation theory (Appendix B.2; Benveniste et al., 1991; Benaim, 1999), we can write a system of differential equations that describe the learning dynamics of a given pair of individuals (i, j) in the population

for the one-shot matching model as

$$\dot{p}_i = p_i(1-p_i)\lambda[\{p_j\mathscr{R}+(1-p_j)\mathscr{S}\}\{p_i+\delta_i(1-p_i)\} - \{p_j\mathscr{T}+(1-p_j)\mathscr{P}\}\{\delta_i p_i+(1-p_i)\}], \quad (3.5)$$

$$\dot{p}_j = p_j(1-p_j)\lambda[\{p_i\mathscr{R}+(1-p_i)\mathscr{S}\}\{p_j+\delta_j(1-p_j)\} - \{p_i\mathscr{T}+(1-p_i)\mathscr{P}\}\{\delta_j p_j+(1-p_j)\}], \quad (3.6)$$

where a dot accent is used to symbolize a time derivative, i.e., $\dot{p}_i = \mathrm{d}p_i/\mathrm{d}t$, p_i is the probability that individual i plays action 1 and p_j is the probability that individual j (the opponent of individual i) plays action 1 . Since individuals cannot detect the state of the game ω, one must note that the parameters $\mathscr{R}$, $\mathscr{S}$, $\mathscr{T}$, and $\mathscr{P}$ are the payoffs of the average game faced by the individuals in the fluctuating environment (Table 3.2); that is, the average over the distribution $\mu(\omega)$ of type of games. For instance when players i and j both choose action 2 they both obtain the average payoff $\mathscr{T} = \mu(\mathrm{PD})\mathscr{T}_{\mathrm{PD}} + \mu(\mathrm{HD})\mathscr{T}_{\mathrm{HD}} + \mu(\mathrm{CG})\mathscr{T}_{\mathrm{CG}}$, and the three other payoffs $\mathscr{R}$, $\mathscr{S}$, and $\mathscr{P}$, are similarly computed . We see that stochastic approximation shows that, asymptotically, the probability of playing actions is driven by the average payoff to actions, which can be thought to determine itself a game (eq. B.9), which we call the average game (see Dridi and Lehmann, 2013, for more details). Note also that the probability that an individual takes action a evolves according to the difference between the expected reward of action a and the average expected reward of all actions (the expected reward of a is the expected value of the second term of eq. 3.3 over the distribution of environmental states and choice probabilities).

To perform the analysis, we assume that the learning dynamic (eqs. 3.5–3.6) has reached an equilibrium during an individual's lifespan in order to evaluate fitness. This is equivalent to say that we let the time horizon, T, of the game become very large (ideally infinite) so that the equilibrium of eqs. 3.5–3.6 determines the fecundity of individuals i and j. Then, the fecundity of individual i is defined as the average (or expected) payoff attained at equilibrium of learning. That is

$$b_{ij} = \hat{p}_i(\hat{p}_j\mathscr{R}+(1-\hat{p}_j)\mathscr{S})+(1-\hat{p}_i)(\hat{p}_j\mathscr{T}+(1-\hat{p}_j)\mathscr{P}), \qquad i,j\in\{\mathbf{R},\mathbf{I}\}, \quad (3.7)$$

where we subscript b_{ij} by j to emphasize that the fecundity of individual i depends on its single opponent (j), $\hat{p}_i$ is the equilibrium probability that individual i plays action 1, and $\hat{p}_j$ is the equilibrium probability that individual j plays action 1 (i.e., $\hat{p}_i$ and $\hat{p}_j$ are the solutions of the equations $\dot{p}_i = 0, \dot{p}_j = 0$; Appendix B.3). We will use the subscript $\mathbf{R}$ to denote a PRL individual

and **I** to denote a PIL. For instance, $b_{\mathbf{RI}}$ is the fecundity of a PRL individual if it is paired with a PIL individual.

In order to use these payoffs to investigate evolutionary dynamics, we make the customary assumption that the population size is infinitely large so that we can use a deterministic evolutionary model. Calling q_i the frequency of type i in the population, the expected reproductive output of type i is then defined as

$$W_i = \alpha + q_i b_{ii} + q_j b_{ij}, \tag{3.8}$$

where α is a baseline fecundity. In the following, we call $q_{\mathbf{R}} \equiv q$ the frequency of RL in the population (so that $1-q$ is the frequency of IL). Hence, the change in frequency Δq of PRL in the one-shot matching model over one iteration of the life-cycle is given by the discrete-time replicator dynamics,

$$\Delta q = q(1-q)\left(\frac{W_{\mathbf{R}} - W_{\mathbf{I}}}{\bar{W}}\right), \tag{3.9}$$

where $\bar{W} = q W_{\mathbf{R}} + (1-q) W_{\mathbf{I}}$ is the mean reproductive output in the population. In the next section, we evaluate the replicator dynamics (eq. 3.9) for three different average games: the Prisoner's Dilemma (PD), the Hawk-Dove game (HD), and the Coordination game (CG) (Table 3.2). The details of the analysis are provided in Appendix B.4. Note that we use p_i and p_j to denote the probability to play action 1 when both individuals are of the same type (i.e., for the interactions PRL vs. PRL and PIL vs. PIL) but we use $p_{\mathbf{R}}$ and $p_{\mathbf{I}}$ for the interaction between a PRL and a PIL.

Table 3.2: Payoff matrices for the average games studied. The rows represent the actions of player 1 and the columns correspond to the actions of player 2. Payoffs are to row player. The matrix at the top shows the generic payoffs used in the paper. In the Prisoner's Dilemma (Left), we assume $B > C > 0$. In the Hawk-Dove (Middle), we have $B > C > B/2 > 0$. In the Coordination Game (Right), $B > 0$.

$\bar{G}$	**Action 1**	**Action 2**
Action 1	$\mathscr{R}$	$\mathscr{S}$
Action 2	$\mathscr{T}$	$\mathscr{P}$

$\overline{\text{PD}}$	**Cooperate**	**Defect**
Cooperate	$B-C$	$-C$
Defect	B	0

$\overline{\text{HD}}$	**Hawk**	**Dove**
Hawk	$B/2-C$	B
Dove	0	$B/2$

$\overline{\text{CG}}$	**Left**	**Right**
Left	B	0
Right	0	B

3.3 Results: one-shot matching

3.3.1 Prisoner's Dilemma

Equilibrium behavior

The learning dynamics for the Prisoner's Dilemma is obtained by replacing in eqs. 3.5–3.6 the payoffs $(\mathscr{R},\mathscr{S},\mathscr{T},\mathscr{P})$ by the values defined in Table 3.2 for that average game. Hence, action 1 can now be thought as "Cooperate" and action 2 as "Defect". In order to determine the fate of the PRL for this case (eq. 3.9), we need to evaluate the payoffs for each possible interaction between types of learners, which will depend on the equilibrium points of the learning dynamics. In the population, three types of pairwise interaction can occur: (i) PRL vs. PRL; (ii) PIL vs. PIL; (iii) PRL vs. PIL.

When a PRL is paired with another PRL we find that the learning dynamics can end in two possible states, depending on the initial preferences $(p_{i,1}, p_{j,1})$ individuals have for each action. If individuals have a high enough initial probability to play Cooperate, then both PRL will learn to cooperate $(\hat{p}_i = 1, \hat{p}_j = 1)$ at equilibrium of learning (Fig. 3.1A, Appendix B.4). For other initial conditions, both PRL learn to defect $(\hat{p}_i = 0, \hat{p}_j = 0)$. The interaction between two PIL gives a different result: irrespective of initial conditions, both PIL will learn to defect (Fig. 3.1B). Finally, when a PRL meets a PIL, both individuals learn to defect regardless of initial conditions, which means that PRL does not get exploited by PIL (Fig. 3.1C).

ESS analysis

Using the above results on equilibrium action play, we can now compute the fitnesses of both types (eq. 3.8). This gives

$$\begin{cases} W_{\mathbf{I}} = \alpha - k, W_{\mathbf{R}} = \alpha + q(B-C) & \text{if i.c. is in the basin of } (1,1), \\ W_{\mathbf{I}} = \alpha - k, W_{\mathbf{R}} = \alpha & \text{otherwise,} \end{cases} \tag{3.10}$$

where i.c. refers to the initial conditions of learning in the PRL vs. PRL interaction. Under the replicator dynamics (eq. 3.9), the frequency of the PRL type increases when $W_{\mathbf{R}} > W_{\mathbf{I}}$. In the first case of eq. 3.10 (i.e., when we start close enough to the equilibrium $(1,1)$), this means that PRL invades PIL if

$$q(B-C)+k>0. \tag{3.11}$$

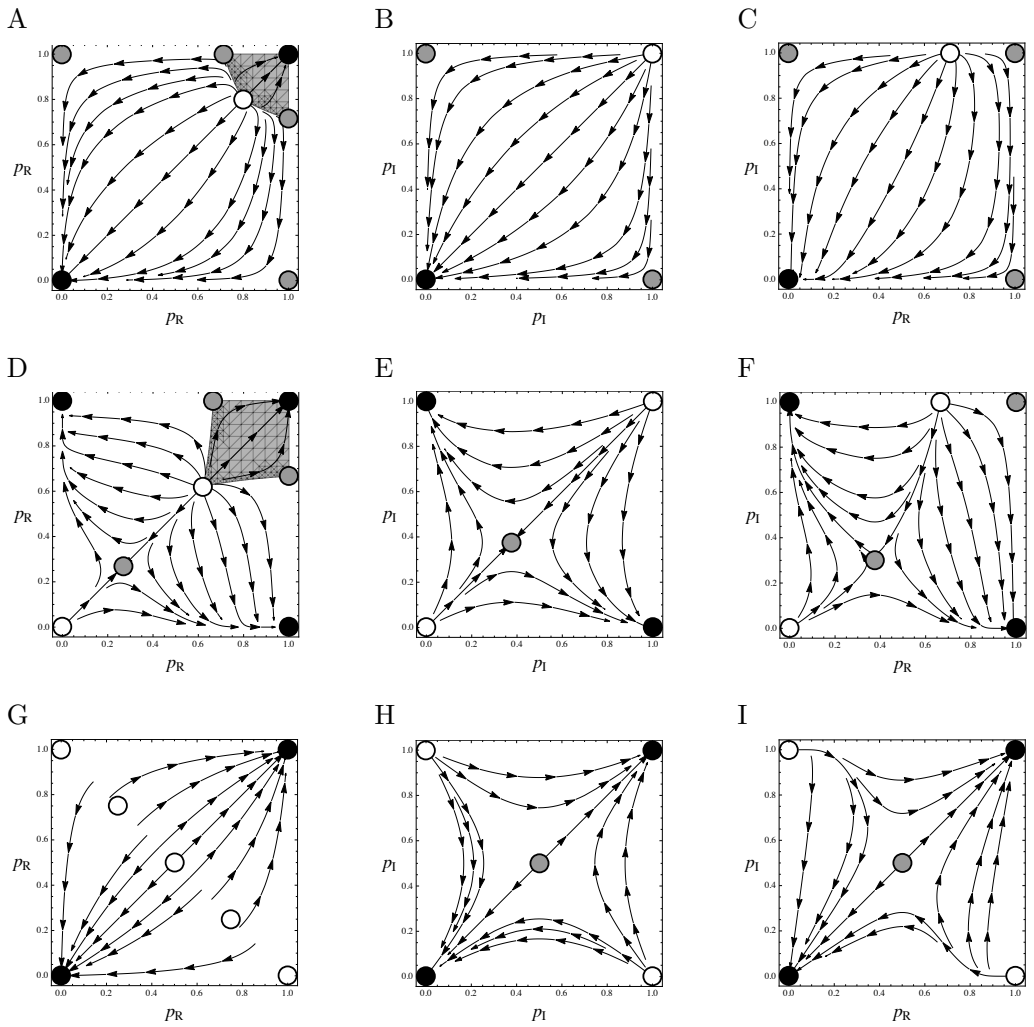

Figure 3.1: Solution orbits of the learning dynamical system in the three average games for the one-shot matching model. Top row (A, B, and C): Prisoner's Dilemma ($B = 5$ and $C = 3$). Middle row (D, E, and F): Hawk-Dove ($B = 5$ and $C = 4$). Bottom row (G, H, and I): Coordination Game ($B = 5$). In each row, the left panel represents the interaction between two PRLs, the center one between two PILs, and the right one between a PRL and a PIL. A white-filled circle denotes an unstable node (both associated eigenvalues are positive), a gray-filled circle is a saddle (one positive and one negative eigenvalue), and a black circle is a locally stable equilibrium. In panel A the gray shaded area represents the initial conditions for which all trajectories go to the (Cooperate, Cooperate) $(1, 1)$ equilibrium. In panel D the gray shaded area represents the initial conditions for which all trajectories go to the (Dove, Dove) $(1, 1)$ equilibrium.

Because the left-hand side is always positive, PRL is the evolutionarily stable learning rule (ESLR); that is, it cannot be invaded by PIL and can invade PIL. Note that, even when $k = 0$, PRL is still the ESLR.

In the second case of eq. 3.10, i.e., when the learning dynamics start in the basin of (Defect, Defect) of the PRL vs. PRL interaction, both PRL and PIL learn to defect. Here, $W_{\mathbf{R}} > W_{\mathbf{I}}$ always holds when $k > 0$, hence PRL is also the ESLR. For $k = 0$, we have neutrality ($W_{\mathbf{R}} = W_{\mathbf{I}}$ for all q).

Individual-based simulations

In order to see whether the above analytical approximation reflects accurately the underlying stochastic model of learning and evolution, we performed individual-based simulations. The main question we ask in these simulations is: can we find parameters such that the above approximation based on deterministic linear stability analysis is robust enough to predict the outcome of natural selection? In Appendix B.5, we provide a detailed description of these simulations.

Simulations: dynamic learning rate

Since the analytical prediction in the Prisoner's Dilemma depends on whether individuals have an initial preference for Cooperation or Defection, we run a set of simulations for each type of initial preference (see Appendix B.6 for a detailed description of results for both dynamic learning rate and constant learning rate for all games).

Both players initially prefer Cooperation. For this case, simulations results are similar to the analytical results. In particular, two PRL can learn to cooperate with each other but always learn to defect against the defector PIL, which favors PRL and leads to the fixation of PRL (Fig. 3.2A,B,C). However, it is noteworthy that the simulations slightly differ from the analysis: pairs of PRL can sometimes learn to defect (Fig. 3.2A), which is a consequence of the possibility of escaping the basin of attraction of cooperation due to stochastic fluctuations. But this does not affect our evolutionary prediction, because it suffices to have a small probability to cooperate with itself in order to outperform the defector PIL (Table 3.3).

Both players initially prefer Defection. For these initial preferences, the simulations results differ from the analytical results because PRL individuals sometimes learn to cooperate with each

other despite their initial tendency to defect. Hence, simulations for the present case are very similar to the simulations of the previous case where individuals initially prefer cooperation. Namely, some pairs of PRL individuals learn to cooperate and some other pairs learn to defect, while the interactions involving PIL always lead to Defection (Fig. 3.2D,E,F). As a consequence, we indeed observe that PRL fixes in the population in our evolutionary simulations (Table 3.3).

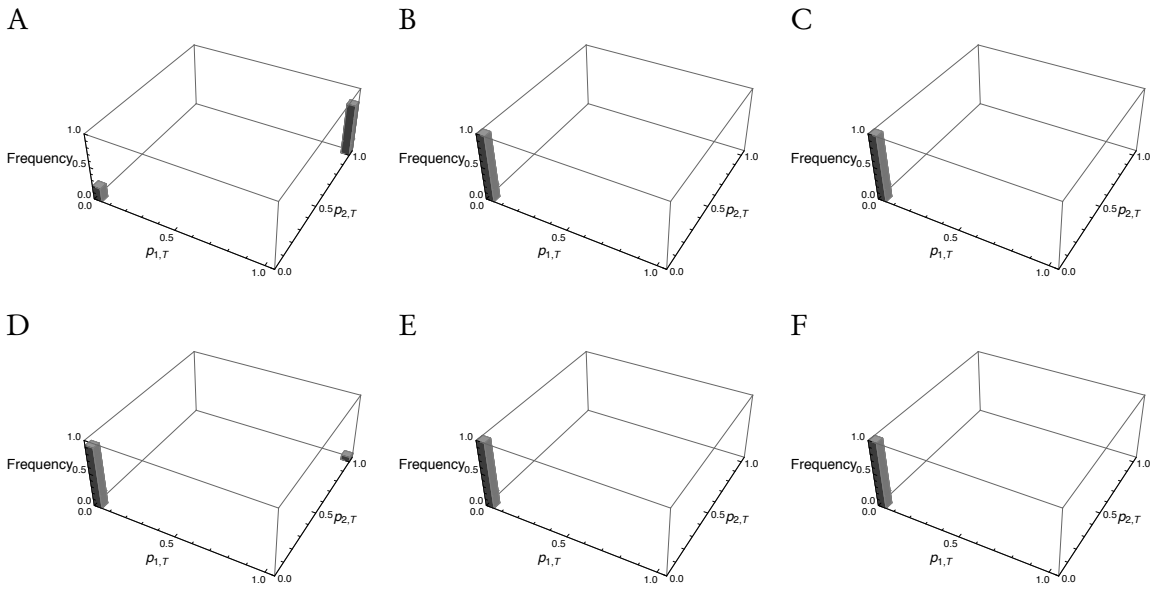

Figure 3.2: Behavioral equilibrium of learning in the average Prisoner's Dilemma for the one-shot matching model with dynamic learning rate for pairs of opponents. This represents the frequency of pairs having reached a given probability to play action 1 $(p_{1,T}, p_{2,T})$ at the end of lifespan, T. We used a total of 1000 individuals of each type in each simulation. First line: initial preference for Cooperation $(p_{i,1} = 0.85)$. Second line: initial preference for Defection $(p_{i,1} = 0.15)$. Left column: interaction between two PRLs. Middle column: interaction between two PILs. Right Column: interaction between PRL (player 1) and PIL (player 2).

Simulations: constant learning rate

Both players initially prefer Cooperation. There, the same qualitative results are obtained as in the dynamic learning rate situation, where ERL can learn to cooperate with itself (Fig. B.1A,B,C) and fixes in the population in the evolutionary long run (Table 3.3).

Both players initially prefer Defection. In this situation there is only one (but important) difference compared to dynamic learning rate. Namely, ERL individuals do not learn full defection and converge to a positive probability of cooperating against the defector BL, which gives an evolutionary advantage to the latter when present in high frequency in the population (Fig. B.1D,E,F). This implies that in our simulations of evolution we observe that ERL fixes in the population when initially frequent, but it is EIL that fixes when itself initially frequent in the population (Table 3.3).

3.3.2 Hawk-Dove Game

Equilibrium behavior

We proceed exactly as in the PD in order to analyze the dynamics of learning (eqs. 3.5–3.6), but we now use the payoffs of the HD game (Table 3.2). Here, action 1 corresponds to "Dove" and action 2 to "Hawk".

In the PRL vs. PRL interaction, the learning dynamics has three stable equilibria: (Hawk, Dove), (Dove, Hawk), or (Dove, Dove). In other words, depending on the initial conditions, individuals will either reach a Nash equilibrium or the "cooperative" outcome where both individuals choose Dove, a result similar to the one obtained in the average PD game above (Fig. 3.1D, Appendix B.4). When two PIL interact (Fig. 3.1E), they end up playing one of the two Nash Equilibria (Hawk, Dove) or (Dove, Hawk). Finally, when a PRL meets an IL, there are two possible endpoints: the equilibrium where PRL plays Hawk and PIL plays Dove; or the reverse situation where PRL learns to play Dove and PIL learns to play Hawk (Fig. 3.1F). Hence, in this heterogeneous interaction, depending on the initial conditions, either PRL or PIL will get exploited by its opponent.

ESS analysis

While the interactions between two individuals of the same type (PRL vs. PRL and PIL vs. PIL), always lead to a payoff of $\frac{B}{2}$ because both individuals have equal chances of learning to become a Hawk or a Dove, the payoffs in the PRL vs. PIL interaction depend on whether the initial condition is in the basin of attraction of $(0,1)$ or $(1,0)$. Thus, the reproductive output is

$$\begin{cases} W_{\mathbf{I}} = \alpha + qB + (1-q)\frac{B}{2} - k, W_{\mathbf{R}} = \alpha + q\frac{B}{2} & \text{if i.c. in basin of } (0,1) \\ W_{\mathbf{I}} = \alpha + (1-q)\frac{B}{2} - k, W_{\mathbf{R}} = \alpha + q\frac{B}{2} + (1-q)B & \text{if i.c. in basin of } (1,0), \end{cases} \tag{3.12}$$

where i.c. is the initial condition of the PRL vs. PIL interaction.

In the first case of eq. 3.12, PRL increases in frequency when rare ($W_R > W_I$) if

$$k - \frac{B}{2} > 0. \tag{3.13}$$

This implies that when $k > B/2$, PRL is the ESLR; when $k = B/2$ evolution is neutral; when $k < B/2$, PIL is the ESLR.

In the second case of eq. 3.12, PRL increases in frequency if

$$\frac{B}{2} + k > 0, \tag{3.14}$$

which is always true because $B > 0$ and $k \geq 0$. In this case, PRL is the only ESLR.

Simulations: dynamic learning rate

PRL initially prefers to play Hawk and PIL prefers Dove. In this case, the simulation results agree qualitatively well with the analytical results. The only difference is that in the simulations, we observe some pairs of PRL that learn the outcome (Dove, Dove) (Fig. 3.3A,B,C). This is due to stochasticity and (as occurred above for the PD) is not possible with initial preferences for Hawk in the analytical model. Consequently, PRL fixes in the population in the long run (Table 3.3).

PRL initially prefers to play Dove and PIL prefers Hawk. Because this situation is the exact opposite of the previous initial conditions, the analysis predicts that behavior should be the same for homogeneous interactions but should be inverted under the heterogeneous interaction (PRL vs. PIL), and this is what we observe. The PIL individuals learn Hawk against the PRL who learn Dove (Fig. 3.3D,E,F), and the former thus fixes in the population at an evolutionary equilibrium (Table 3.3).

Simulations: constant learning rate

ERL initially prefers to play Hawk and EIL prefers Dove. Here the results are the same than under dynamic learning rate and ERL fixes in the population in the long run (Fig. B.2A,B,C, Table 3.3).

ERL initially prefers to play Dove and EIL prefers Hawk. This case displays the same situation as dynamic learning rate, where EIL individuals outcompete ERL and fix to a frequency close to 1 at equilibrium of evolution (Fig. B.2D,E,F, Table 3.3).

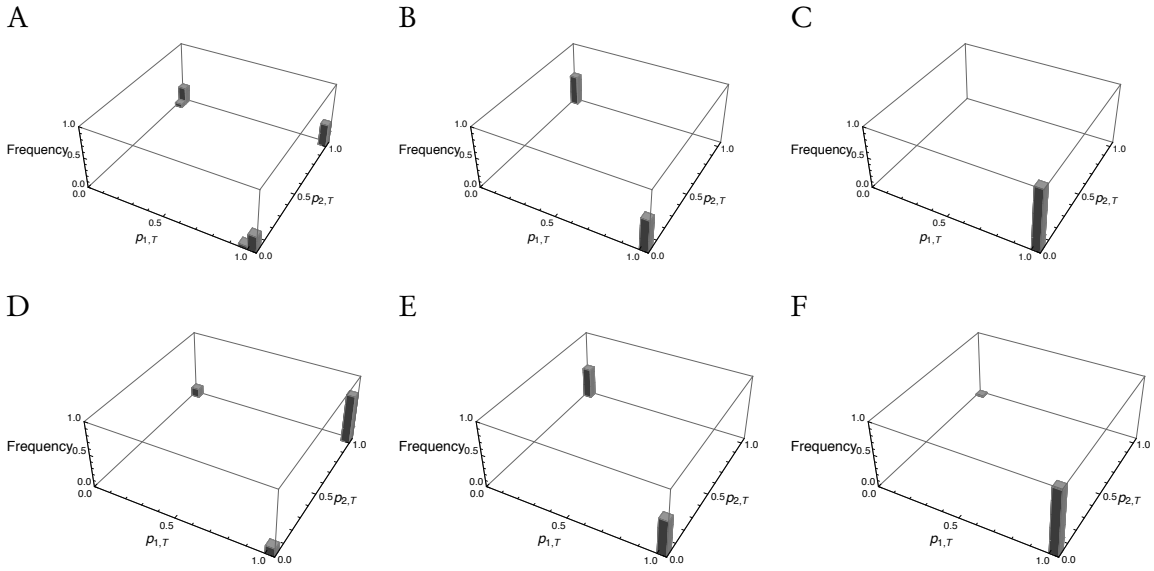

Figure 3.3: Same as in Fig. 3.2 but for the behavioral equilibrium under the average Hawk-Dove game. First line: RL initially prefer Hawk ($p_{\mathbf{R},1} = 0.15$) and IL prefers Dove ($p_{\mathbf{I},1} = 0.85$). Second line: RL initially prefer Dove ($p_{\mathbf{R},1} = 0.85$) and IL prefers Hawk ($p_{\mathbf{I},1} = 0.15$). Left column: interaction between two RLs. Middle column: interaction between two ILs. Right Column: interaction between RL (player 1) and IL (player 2).

3.3.3 Coordination Game

Equilibrium behavior

If we use the payoffs of the CG (Table 3.2) in eqs. 3.5–3.6, we obtain the learning dynamics of a pair of opponents in the coordination game. In the labeling of Table 3.2, action 1 corresponds to Left and action 2 corresponds to Right. In this game, all three types of pairs succeed in learning to coordinate in the long run (Fig. 3.1G,H,I), and depending on the initial preferences for Right or Left, the equilibrium reached will either be (Right, Right) or (Left, Left).

ESS analysis

In the three interactions, the players coordinate on a single action and get a payoff at equilibrium of B. The fitness of type PRL is then $W_{\mathbf{R}} = \alpha + B$ and the fitness of PIL is $W_{\mathbf{I}} = \alpha + B - k$. Trivially, PRL is the ESLR for all positive k; for $k = 0$, evolution is neutral.

Simulations: dynamic learning rate

In this game, our analytical results predict that evolution does not favor one type over the other for any initial condition because both types always coordinate on a single action, which should lead to a neutral evolutionary dynamics where both types co-exist in equal frequency ($q = 1/2$). In the simulations, we thus set initial preferences of individuals close to the outcome (Left, Left) and we observed that all types of pairs of individuals succeed in coordinating on a given action (but the particular action learned depends on the pair, Fig. 3.4). This learning behavior corresponds to our analytical prediction but simulations of evolution do not match as we observe a stable polymorphism dominated by PIL (Table 3.3). This result may be explained by the variance in convergence time of PRL individuals. Our criterion for convergence of learning behavior was indeed based on the average time needed for individuals to converge, but this disregards the possibility that different PRL individuals may converge in different times. Because certain PRL fail to converge to a pure action at time T, they do not coordinate on certain interaction rounds and hence get the payoff 0 (Table 3.2) more often than PIL.

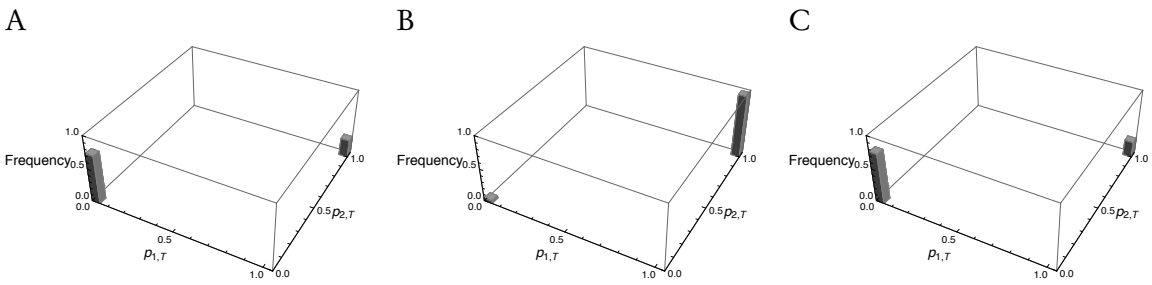

Figure 3.4: Same as in Fig. 3.2 but for the behavioral equilibrium under the average Coordination game. Left column: interaction between two RLs. Middle column: interaction between two ILs. Right Column: interaction between RL (player 1) and IL (player 2).

Simulations: constant learning rate

Learning in this case also leads to coordination of all pairs (Fig. B.3). However, here evolutionary simulations give a result closer to what is expected with a frequency of ERL of $q \approx 0.56$ at equilibrium of evolution (Table 3.3).

Table 3.3: Summary of results in the one-shot matching model. The column "Predicted q^*" shows the frequency of RL expected at evolutionary equilibrium under the deterministic approximation. The column "Simulated q^*" gives the approximate equilibrium frequency of RL obtained in the corresponding evolutionary simulation. The simulation results represent the average over simulation runs with different initial compositions of the population (see Appendix B.5 for details).

Average Game	**Initial condition of learning**	**Predicted q^***	**Learning rate**	**Simulated q^***
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of RLvsRL	1	Dynamic	0.99
			Constant	0.98
	Basin of (Defect, Defect) of RLvsRL	1/2	Dynamic	0.98
			Constant	0.5‡
Hawk-Dove Game	Basin of (Hawk, Dove) of RLvsIL	1	Dynamic	0.97
			Constant	0.98
	Basin of (Dove, Hawk) of RLvsIL	0	Dynamic	0.01
			Constant	0.01
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.27
			Constant	0.56

‡ In this case, the different simulation runs give disparate results with either type getting fixed depending on the initial conditions (see main text).

3.4 Results: repeated matching

We now assume that individuals in the population are randomly paired at every time t of lifespan, but otherwise keep all previous assumptions. An individual will now meet different partners during its lifespan. Its learning dynamics may then depend on the distribution of behaviors of all individuals in the population because anybody can be met for a one-shot interaction.

Unfortunately, we could not find an analytic approximation to the evolutionary dynamics for this matching model, so we used exclusively individual-based simulations to investigate the evolutionary stability of RL and IL. As for the one-shot matching model, we performed simulations of only the learning process on one hand and a full evolutionary analysis on the other hand. All parameters were set as for the one-shot matching model. In particular, to make possible the comparison between the two matching rules, we used the same type of initial conditions for the learning dynamics.

For the simulations of the learning dynamics, we cannot simply take pairs of individuals out of the population since the dynamics of learning of an individual depends on all other individuals in the population. In particular, the learning dynamics is sensitive to the types' frequencies. Thus, the learning simulations were performed by simulating one generation in a large population (10000 individuals) and by setting manually the frequencies of RL and IL in order to understand the effect of the types' frequencies on the learning behavior.

3.4.1 Prisoner's Dilemma

Dynamic learning rate

All individuals initially prefer Cooperation. For this case, we obtain that PRL individuals can learn to cooperate when very common in the population (precisely, when $q \geq 0.8$), while PIL individuals always learn to defect for all compositions of the population (Fig. 3.5A). This implies that, at the evolutionary timescale, the population will move neutrally through all the states such that $q < 0.8$ but is repelled from the states where $q \geq 0.8$. As a consequence we observe in our simulations that the equilibrium frequency of PRL is small but positive (Table 3.4).

All individuals prefer Defection. Here, both types always converge to Defection, so no one has an advantage (Fig. 3.5B). When we simulate evolution, the result is qualitatively in agreement

with this even if the frequency of PIL is slightly above 0.5 (Table 3.4). This can be explained by the variance in convergence time of PRL individuals. Some of them might converge more slowly to full Defection and will be exploited on some interaction rounds when they meet the defector PIL.

Constant learning rate

All individuals initially prefer Cooperation. Here ERL players learn to cooperate with a small probability for all frequencies, while EIL learn to defect as usual (Fig. 3.5C). As a consequence, the frequency of EIL in evolutionary simulations is close to 1 at equilibrium (Table 3.4).

All individuals initially prefer Defection. The result is here the same as in the previous section: ERL individuals learn to cooperate with a positive probability (Fig. 3.5D) which leads to the evolutionary superiority of EIL observed in our evolutionary simulations (Table 3.4).

3.4.2 Hawk-Dove Game

Dynamic learning rate

PRL initially prefers to play Hawk and PIL prefers Dove. Here, we obtain that individuals using the PRL rule learn to play Dove with a higher average probability than PIL does. As the frequency of PRL increases, their probability to play Dove also increases, while PIL plays Hawk more and more often (Fig. 3.5E). Consequently, when we perform simulations of evolution, we observe a stable polymorphism with a clear domination of PIL (Table 3.4). The reason is that playing Hawk with a higher probability than the opponent is beneficial in one-to-one interactions (which favors PIL in interactions against PRL) but playing Hawk too often renders susceptible to invasion by individuals who play the nicer Dove action (which favors PIL).

PRL initially prefers to play Dove and PIL prefers Hawk. The learning behavior of the types is similar to the previous case but now the average probability that PRL plays Dove is less affected by its frequency, and is always relatively high. Individuals using PIL still have a tendency to increase their probability to play Hawk as q increases (Fig. 3.5F), which gives them an advantage. This is confirmed by our evolutionary simulations where PIL dominates the population at a polymorphic evolutionary equilibrium (Table 3.4).

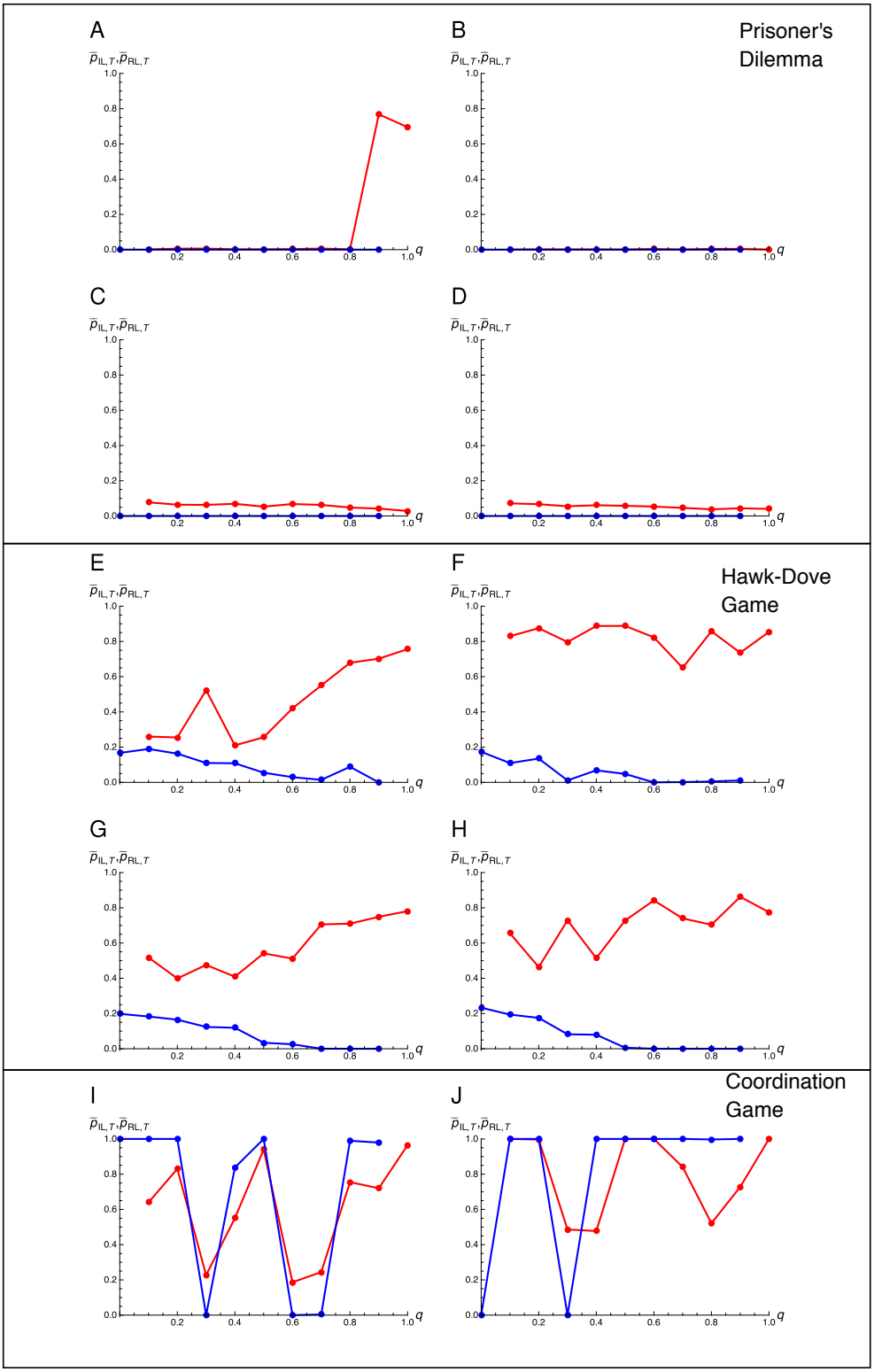

Figure 3.5: Average probability of choosing action 1 at the behavioral equilibrium for different frequencies q of RL in the population in the repeated matching model. Red line is for RL, and blue line is for IL. Panels A,B,C,D: Prisoner's Dilemma. Panels E,F,G,H: Hawk-Dove Game. Panels I,J: Coordination Game. In the Prisoner's Dilemma panel: first line: dynamic learning rate. Second line: constant learning rate. Left column: Individuals start with an initial preference for Cooperation ($p_{i,1} = 0.85$). Right column: initial preference for Defection ($p_{i,1} = 0.15$). In the Hawk-Dove panel: First line: dynamic learning rate. Second line: constant learning rate. Left column: RL initially prefer Hawk ($p_{\mathbf{R},1} = 0.15$) and IL prefers Dove ($p_{\mathbf{I},1} = 0.85$). Right column: RL initially prefer Dove ($p_{\mathbf{R},1} = 0.85$) and IL prefers Hawk ($p_{\mathbf{I},1} = 0.15$). In the Coordination panel: left: dynamic learning rate. Right: constant learning rate.

Constant learning rate

ERL initially prefers to play Hawk and EIL prefers Dove. Here the results are similar to the case with dynamic learning rate so ERL learns to play Dove more often than PIL (Fig. 3.5G) such that simulations of evolution lead to a polymorphic equilibrium where PIL constitutes almost all the population (Table 3.4).

ERL initially prefers to play Dove and EIL prefers Hawk. The results for the learning behavior are not different from the case with dynamic learning rate (Fig. 3.5H) so we also observe in our simulations of evolution that EIL constitutes almost all the population at an evolutionary endpoint (Table 3.4).

Table 3.4: Summary of results in the repeated matching model. This is the same table as for the one-shot matching case but without analytic prediction.

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.2
		Constant	0.05
	All individuals prefer Cooperation	Dynamic	0.48
		Constant	0.04
Hawk-Dove Game	RL prefers Hawk, IL prefers Dove	Dynamic	0.18
		Constant	0.11
	RL prefers Dove, IL prefers Hawk	Dynamic	0.07
		Constant	0.05
Coordination Game	All individuals prefer Left	Dynamic	0.01
		Constant	0.06

3.4.3 Coordination Game

Dynamic learning rate

The results for the learning behavior show here that PRL individuals fail to all coordinate on a given action. The PIL individuals are efficient in doing this (Fig. 3.5I) and thus fix in the population at the endpoint of evolution (Table 3.4).

Constant learning rate

The learning behavior of both types is qualitatively similar to the dynamic learning rate case (Fig. 3.5J) and this explains why EIL almost fix in the population at an evolutionary equilibrium in our simulations (Table 3.4).

3.5 Discussion

In order to assess whether selection favors cognitively more sophisticated individuals than simple reinforcement learners in social interactions, we analyzed an evolutionary model of the competition between reinforcement and inference-based learning. Reinforcement learners only use information about realized payoffs, while inference-based learners in our model also use information about the forgone payoffs of playing alternative actions. This learning mode can be thought of as one step up in the cognitive hierarchy and is further related to standard belief-based learning (see eq. B.6 in Appendix B.1).

We assumed that individuals in a large population are genetically programmed to be either reinforcement learners or inference-based learners and interact repeatedly in a two-player, two-action stochastically fluctuating games. We defined two matching schemes, that is, two ways in which individuals meet to play the games. In the one-shot matching model, individuals are paired at the beginning of the fluctuating game and each pair interacts for the rest of the game. In the second model, we used repeated matching: here, a random matching is realized at each time t of the game so that an individual has a negligible probability of playing twice against the same partner (since we also consider that the population is very large). Payoffs are evaluated at equilibrium of the learning process, and this defines the number of offspring produced (fecundity) by an individual.

We applied stochastic approximation theory to analyze learning during lifetime, and obtained that the equilibrium behavior of the learners could be characterized in terms of an average game of the fluctuating game (i.e., a game whose payoffs are averages of the sub-games payoffs of the original fluctuating game). We thus analyzed three standard cases of average game: the Prisoner's Dilemma, the Hawk-Dove game, and the Coordination game, and checked our analytic approximations with simulations of the exact process.

Overall, the presupposed domination of inference-based learning over reinforcement learning is not complete in our results. In other words, the ability to infer forgone payoff by sim-

ulating payoff outcomes of unchosen actions does not necessarily leads to a selective advantage in social interactions. Rather, we actually observed three main types of results, which hold regardless of the learning rate (constant or dynamic), and that we now describe.

(1) In simple social interactions, where the average game faced by individuals just require that two partners coordinate on the same action or "anti-coordinate" on two different actions, reinforcement and inference-based learning do not produce different behaviors at the behavioral equilibrium, in which case natural selection is neutral and does not favor one rule over the other.

(2) In social dilemmas like the Prisoner's Dilemma, the ability of two reinforcement learners to generate cooperative pairwise interactions by reinforcement of actual rewards, rather than to play a Nash equilibria, can give it an evolutionary advantage over inference-based learning. Importantly, we obtained this result even when inference-based learning payed no cost for cognitive complexity.

(3) Overall, we observed many examples where belief-based learning dominates the population at an evolutionary endpoint (i.e., it either gets to fixation or the population reaches a polymorphism where belief-based learning is at a high frequency), especially when two individuals cannot interact more than once (thereby eliminating the possibility for reciprocation). Since belief-based learning produces rational behavior at the level of the one-shot average game, this makes perfect sense, because it is the type of behavior that is selected when individuals cannot interact more than once.

Across all these results, the interaction between the learning rule and the initial preferences, which can be interpreted as an innate predisposition for a certain type of action, also plays an important role. We observed that like any predisposition in a learning context, this predisposition can be overcome or reversed in the long run, but this is much constrained by the dynamic properties of the interacting learning rules (e.g., size of the basin of attraction of the behavioral equilibria). This should be kept in mind in the following discussion about the more specific effect on evolutionary outcomes of the two different matching models.

In the one-shot matching model, we either observed results of type (1) or type (2) but we never observed results of type (3). The one-shot matching model allows to capture situations where the same animals interact many times together so that each animal has the possibility to tune its behavior to the actions of its partner. Examples where repeated interactions are likely to occur include many cooperatively breeding species that live in relatively small stable groups, or

if the home range of solitary animals overlap. In this situation, results of type (1) were obtained when the average game was either a Hawk-Dove game or a Coordination game. In these two games, the two learners did not differ from one another at a behavioral equilibrium, which leads to the absence of an advantage of one type of learner over the other. In this game, we found that both types of learners generally succeeded in coordinating on a single action, because it is only necessary to repeat the action that leads to positive payoffs, and both the reinforcement and inference-based learning rules are capable of this.

The Hawk-Dove game favors one learning rule or the other depending on the initial conditions, so no type always wins in this game, because both types of learners are able to reach the optimal (Nash equilibria) outcomes (Hawk, Dove) and (Dove, Hawk). This game actually illustrates well the interaction effect between genetic predisposition and learning rule, because we found that it was the learning rule that had the biggest predisposition for aggression (big initial probability of choosing "Hawk") that was evolutionarily stable. Besides, it is noteworthy that the behavior of reinforcement learners was slightly more cooperative in the sense that pairs of this type could also learn to play the socially peaceful outcome (Dove, Dove). This did not give an advantage nor a disadvantage compared to inference-based learners in this particular game, but this suggests that reinforcement learners tend to avoid aggression by staying away from the "Hawk" action.

In the average Prisoner's Dilemma, we also observed that reinforcement learners could reach the socially beneficial outcome where both partners cooperate (i.e., they "solve" the dilemma), but this time this ability made reinforcement evolutionarily stable. Indeed, pairs of reinforcement learners are able to learn to cooperate together. On the other hand, inference-based learning always leads to defection. Interestingly, reinforcement learners do not get exploited by inference-based learners as they succeed in learning to defect against them. We note that reinforcement learning behaves somehow similarly to the Tit-for-tat (TFT) strategy (Rapoport and Chammah, 1965; Axelrod, 1980; Axelrod and Hamilton, 1981): reinforcement learning cooperates with itself a high proportion of the time (but not always) and defect against the defector belief-based learning. The repeated Prisoner's Dilemma has been the topic of many studies aiming at understanding the evolution of cooperation, but most of the time no learning rules are used in evolutionary analysis, but qualitative strategies consisting of finite state automaton like TFT (or Win-stay-lose-shift, Grim, etc.). Learning rules may actually represent a more appropriate way of conceptualizing animal behavior because it describes several realistic features of animals, like incremental adaptation, forgetting, habituation. Moreover, learning rules provide

a quantitative approach (in our case through the motivations for actions) that can potentially be linked to neuronal decision making (Dayan and Abbott, 2005; Enquist and Ghirlanda, 2005; Niv, 2009). It is thus interesting to see that a fairly simple learning rule based on the widely accepted principle of trial-and-error produces qualitatively similar behavior as TFT, although reinforcement learning is much less domain specific than TFT.

In the repeated matching setting, individuals meet different partners at each interaction round. The learning task is here more complicated because an individual must adapt to an entire population composed of different types of learners who themselves adapt their behavior. Here, we mainly observed results of type (3), i.e., inference-based learning generally dominated the population at an evolutionary equilibrium. The most representative example is the average Hawk-Dove game, where we found that inference-based learning was able to exploit the tendency of reinforcement learners to play the "Dove" action too often, especially when reinforcement learners are in a high frequency in the population. Besides, we also observed in this repeated matching model an interaction between genetic predisposition for actions and learning rule in the average Prisoner's Dilemma. When individuals initially prefer "Defection", we obtained a neutral situation where both types learn full defection. However, when individuals had an initial preference for "Cooperation", reinforcement learning could lead to cooperation when in high frequency in the population, and get exploited by inference-based learning which always lead to defect.

The primary goal of this paper was to provide some intuition as to when a learning rule being located one step up in the cognitive hierarchy than reinforcement learning can be favored by selection in simple situations of social interactions. In comparative cognition (Shettleworth, 2009), trial-and-error learning is often the null hypothesis against which one tests hypotheses regarding more advanced cognition, and we adopted the same approach here by letting reinforcement compete with inference-based learning. Our findings illustrate that learning rules using more information do not necessarily outcompete reinforcement learning in social interactions, so the advantage of "complex cognition" is not automatic, unlikely to be gradual, and depends on the type of games played by the individuals in a population.

It might be difficult to assess the payoff structure of the games played by real animals, but our results show that when individuals cannot condition their actions on the type of games they are are playing, the results only depend on the average game over the entire lifespan of an individual, which can thus be used to produce prediction about the psychological capacities of that species.

More generally, being able to condition behavior on the type of game requires to detect features of the environment and combine them correctly in order to compute game payoffs accordingly, which is cognitively more demanding and requires more inference than we have assumed in this paper. Here we only assumed that individuals had access to limited information on missed opportunities. This is based on the observation that in nature, it is difficult to assess the value of actions before trying them (e.g., assessing the quantity of resource available on a food patch) but trying them can give cues not only about the actions themselves, but also about alternative related options (e.g., noticing that a food patch has been exploited by others is a cue that nearby patches have probably also been exploited).

Our paper also contributes to modeling in behavioral ecology, where researchers acknowledge the need for describing animal behavior in terms of general rules that are used to face several decision problems an individual may face (McNamara and Houston, 2009; Dijker, 2011; Hammerstein and Stevens, 2012; Fawcett et al., 2013). By modeling an environment where individuals face different social games and partners, but use the same behavioral rule (either reinforcement or belief-based learning), a learning rule can only works well on average. We concentrated here in this context on social behaviors only under the simplest games, but this approach can also be applied for studying rules of behavior that can serve under a variety of ecological contexts, or under more complex social structures. This may help to better delineate the possible evolutionary paths from simple to more sophisticated decision-making processes.

Chapter 4

Environmental complexity, reaction norms, and the evolution of learning

Abstract

Learning is a specific form of phenotypic plasticity and as such it may enhance fitness in changing environments. But learning is also different from simpler forms of plasticity, like innate responses to stimuli. However, classical models of the evolution of learning do not clearly distinguish between learning and other forms of plasticity. Thus, it remains unclear in what sense learning provides a fitness advantage. In this work, we consider explicitly the evolutionary transition from innate responses to stimuli to learning, by modeling an environment where an animal can encounter a large number of stimuli (or situations) in the course of its lifespan. In our model, animals are assumed to have a maximal brain size and the genotype of an individual codes for the proportion of memory "slots" dedicated to innate responses vs. dynamic slots used for learning. We argue that a reliance on a dynamic memory can help in dealing with a very large number of stimuli, thanks to the ability to forget past information. We study the validity of our argument in two special cases of the model. First, if an animal encounters environmental stimuli totally randomly, we find that learning does not provide a fitness advantage because the probability to encounter the same stimulus twice in a short period of time is too low. Second, in environments where a minimum memory size is required to remember interactions with stimuli, it is shown that reliance on a dynamic memory can evolve as long as the environment is complex enough, i.e. if the environment contains a sufficiently large number of stimuli. We conclude that environmental complexity could be an important but overlooked factor driving the evolution of learning and memory.

4.1 Introduction

Phenotypic plasticity, the ability of an organism with a given genotype to produce different phenotypes as a function of environmental conditions, is a universal biological feature that is favored by natural selection (Pigliucci, 2001). Its evolutionary advantage is due to the fact that it allows to express appropriate phenotypes under environmental variability (be it spatial or temporal), a fact supported theoretically and empirically (Gomulkiewicz and Kirkpatrick, 1992; Gavrilets and Scheiner, 1993; Pigliucci, 2005).

In the field of animal behavior, learning is probably the most studied form of phenotypic plasticity (Dukas, 2004; Shettleworth, 2009). Behaviorally speaking, learning may be defined as the change of actions as a function of experience, i.e., the use of environmental information to gradually change behavior. There is an important body of research within the field of evolutionary biology that focuses on the question of when natural selection would favor learning. Surprisingly, this work seems to be relatively independent from the work on the adaptive value of plasticity, despite the fact that the conditions under which learning is found to be selected for are qualitatively similar to the conditions favoring the evolution of general phenotypic plasticity. Indeed, there is a certain consensus in evolutionary biology on the idea that learning tends to be favored under (mainly between-generation) fluctuating environments (Boyd and Richerson, 1988; Stephens, 1991; Feldman et al., 1996; Kerr and Feldman, 2003; Wakano et al., 2004; Dunlap and Stephens, 2009). This result looks similar to the above idea that plasticity is favored under varying environments and, accordingly, the intuitive explanation of the adaptiveness of learning is the same as the one for the adaptiveness of plasticity. Namely, individuals who can adapt their phenotype (or behavior) to the present environmental conditions must have a fitness advantage over individuals who only express genetically determined actions inherited from their parents (who might have lived in different environmental conditions). In view of the two bodies of literature (general phenotypic plasticity on one hand, learning on the other hand) one might then ask: if learning is just a particular form of phenotypic plasticity and is selected under similar conditions than general phenotypic plasticity, why then formulate specific models of the evolution of learning (Hollis and Guillette, 2011)?

This question can be answered by realizing that learning has its mechanistic specificities, that distinguishes it from other forms of plasticity. For example, learning may be selected against in environments that change too fast within an individual's lifespan, because learning takes time: an individual must be able to reuse information gathered from past actions, and this is impos-

sible in environments that change too often. This is a difference with certain other (reversible) reaction norms, where an organism can readily change its phenotype as soon as the environment changes (e.g., temperature in ectotherms), because it has a genetically encoded program to tune the phenotype to environmental conditions.

This distinction between learning and other reaction norms leads to an important conceptual point: while learning is certainly a form of plasticity (and in this sense, it is not surprising that it is selected under somehow similar environmental conditions than general phenotypic plasticity), there are other, simpler, forms of plasticity in animal behavior that would also allow to deal with environmental fluctuations, but their evolutionary relevance have rarely been modeled and compared with learning in the specific context of animal behavior (Kerr, 2007). From an empirical standpoint, there are many examples in nature showing that individuals tend to have innate, hard-wired responses to certain stimuli (e.g., Mery and Kawecki, 2004; Riffell et al., 2008; Gong, 2012) and these responses are likely to be favored by natural selection (Dorosheva et al., 2011). As noted above, since time is a valuable resource for living organisms, why then take time and bother learning to adapt to a stimulus when an innate reaction could a be more direct way to cope with it?

An answer can be found in models for the evolution of learning (e.g., Wakano et al., 2004; Dunlap and Stephens, 2009), where it is argued that the fitness consequences of actions taken in response to a given stimulus might change because of environmental fluctuations. Under these circumstances, learning is favored over innate responses because it allows to change an animal's reaction to this particular stimulus. But, while being convincing, this argument does not settle the debate for two reasons. First, this argument gives one particular reason for why natural selection favors learning over innate reaction norms, but there may be other conditions that favor learning (in other words, one must make the distinction between a necessary and a sufficient condition). Second, models of the evolution of learning almost never take into account several stimuli or situations at the same time, but only focus on a given situation and ask what happens if the fitness consequences of actions applied to this situation change. In order to justify the evolution of learning on realistic grounds, one should rather take into account the full range of stimuli or situations an individual may face in the course of its lifespan, and the present paper proposes to do so.

When trying to account for the full range of situations an animal has to face, one quickly realizes that this number must be astronomical. The world is constituted of an enormous amount of

different situations, where each of these situations triggers a particular combination of sensory perceptions in an animal's brain. It seems then impossible to encode (even with the abstract encoding provided by neural networks, Dayan and Abbott, 2005; Enquist and Ghirlanda, 2005) in a finite brain the whole amount of environmental stimuli. This leads us to propose that learning and forgetting might be useful to deal with environmental complexity. The idea is that forgetting allows to remove from memory past situations, in order to encode and learn new situations as they are encountered by an animal, and thus makes possible to react to an arbitrarily large number of situations. The adaptive value of forgetting has already been discussed in previous research, but (again) on the grounds of environmental variability (Kraemer and Golding, 1997; Kerr and Feldman, 2003). According to these researchers, forgetting allows to face the same situation at distinct instants and if the optimal behavior for that situation has changed, an individual will be able to generate different behavior. We propose here a different role for forgetting: it allows to encode different situations, or different stimuli, because different stimuli may be encountered at distinct instants of time. Even in the absence of environmental fluctuations, an individual is likely to encounter in the course of its lifespan a large number of environmental stimuli and learning to interact with new stimuli should give a fitness advantage.

A model is presented below that formalizes this verbal argument. In this model, an animal is seen as having a maximum amount of memory at its disposal for brain functioning. The genotype of an individual prescribes to allocate different amounts of memory either to innate responses to stimuli or to a dynamic memory that can encode, learn about, and forget stimuli. The world is comprised of a finite (but possibly very large) number of stimuli (or situations), where each stimulus is characterized by its own set of optimal and suboptimal actions. Importantly, environmental fluctuations are not explicitly considered, because the goal here is precisely to assess the effect of environmental complexity *per se*. It is asked whether a large number of stimuli in the environment (a measure of environmental complexity) creates an evolutionary pressure in favor of a greater allocation of memory to learning.

4.2 Model

4.2.1 Environment

Consider animals acting during a finite decision period of length T. At each decision step t ($t = 1,2,\ldots,T$), an animal has to choose an action among a fixed set $A = \{1,\ldots,n\}$. The

action chosen is a response to the currently faced and environmentally determined stimulus, s_t. There is a set of stimuli $S = \{1, \ldots, k\}$, and at each time t a stimulus s_t is drawn in S according to a homogeneous and irreducible Markov chain $\{s_t\}_{t=1}^{T}$ with transition probabilities $\rho_{ss'} = \mathbb{P}\{s_{t+1} = s' | s_t = s\}$. Thus $\rho_{ss'}$ is the probability to encounter stimulus s' a time $t+1$ given that stimulus s was met at time t. Moreover, it will be assumed that this Markov chain in its stationary distribution, and that the stationary distribution is uniform, which can be written $\mu_s = 1/k$ for all $s \in S$, where μ_s is the stationary probability that stimulus s obtains. This has the consequence that the proportion of lifetime an individual will spend interacting with any one stimulus is equal to $1/k$.

As a result of taking action a in response to stimulus s, an individual receives payoff $\pi(a, s)$, which increments (or decrements) its fecundity by a proportional amount. For the sake of simplicity, action a applied in response to a given stimulus s is assumed to be either optimal and give payoff x_s, or be suboptimal and give payoff y_s. In other words, for each stimulus s, the action set can be divided into two sets: a set of optimal actions, and a set of suboptimal actions. Each stimulus s is characterized by its own set of optimal actions, so that θ_s is the number of optimal actions associated to stimulus s. The other $n - \theta_s$ are suboptimal for stimulus s. Even if this is an oversimplification of reality, this allows to associate to each stimulus two important characteristics: the difficulty of learning the optimal action for the stimulus, measured by θ_s, and the importance of a stimulus on fitness, given by $x_s - y_s$. The closer θ_s is to n, the easier it is to learn the optimal action for stimulus s. The bigger $x_s - y_s$ is, the bigger is the effect of learning a correct action for this stimulus on fitness.

4.2.2 Behavioral rule

The decision system of the animal is modeled as a finite number of associations between stimulus and action. It is assumed that the animal has a memory of size m, i.e., it can encode only m associations between stimulus and action. From these m memory "slots", the genotype prescribes to have j fixed associations. These j associations are present at birth and hold j templates of stimuli together with the innate response of the animal to these stimuli. It will be called the behavioral reaction norm of the animal. Note that if $j = 0$, the animal is born with a blank slate, with absolutely no innate tendency to respond to any stimulus (Kerr and Feldman, 2003). Let G be the set of stimuli to which the animal has an innate response ($|G| = j$, $G \subset S$). For each stimulus $s \in G$, the animal has a genetically determined mixed strategy

$\mathbf{z}(s) = (z(1,s),\ldots,z(n,s)) \in \Delta A$ (with $\sum_{a\in A} z(a,s) = 1$, where $z(a,s)$ is the probability to take action a when stimulus s is encountered). The collection $(\mathbf{z}(s))_{s\in G}$ constitutes what is called the behavioral reaction norm of the animal.

The rest of the memory ($m-j$ slots) is dynamic and can encode new stimuli encountered in the course of lifespan. Let denote by F the set of dynamic memory slots, which can be occupied by stimuli. In this "flexible" memory, new stimuli can replace old ones and the animal is assumed to respond to the stimuli in F. With this, a stimulus s_t encountered at time t can either be genetically encoded in the memory (the event $\{s_t \in G\}$ obtains), be located in the dynamic memory (event $\{s_t \in F\}$ obtains), or is unknown and this is denoted by $\{s_t \in U\}$. Actions associated to the dynamic part of the memory are updated according to a learning process, and $q_t(a,s)$ denotes the probability that action a is taken at time t if stimulus s in the dynamic memory is encountered in the environment.

In summary, the probability that an animal takes action a at any time t, when stimulus s_t is faced, can be written as

$$p_t(a) = \begin{cases} z(a,s_t), & \text{if } s_t \in G, \\ q_t(a,s_t), & \text{if } s_t \in F, \\ 1/n, & \text{if } s_t \in U, \end{cases} \tag{4.1}$$

where in the third case, when an unknown stimulus is met, a random behavior is applied. One can then introduce the following shorthand notation for the expected payoff obtained in each of these three cases,

$$\mathbb{E}[\pi(a_t,s_t)|s_t] = \begin{cases} \pi_G(s_t) = \sum_{a\in A} z(a,s_t)\pi(a,s_t), & \text{if } s_t \in G, \\ \pi_F(s_t) = \sum_{a\in A} q_t(a,s_t)\pi(a,s_t), & \text{if } s_t \in F, \\ \pi_U(s_t) = \sum_{a\in A} \frac{1}{n}\pi(a,s_t), & \text{if } s_t \in U. \end{cases} \tag{4.2}$$

4.2.3 Stimulus replacement mechanism

What does happen when a new, unknown stimulus, $s_t \in U$, is encountered? It was just said that the first action in response to a new stimulus is randomly chosen. But the next question is: should the animal store in memory the outcome of the interaction with that new stimulus? One possibility is to assume that, when an unknown stimulus is met, the animal always wants to encode it in its memory F. Consequently, the new stimulus s_t will either take a free slot, or will replace another stimulus. For the first $m-j$ encounters with non-innately encoded

stimuli, the new stimuli will take free slots in F and there is nothing to specify further. But for subsequent decision steps, new stimuli will have to replace other ones in F. This is done via a replacement rule. It is assumed the following replacement rule, which is taken from Kerr and Feldman (2003). A stimulus has a lifespan in memory of $m-j$ time steps (i.e., the size of the dynamic memory), starting from the last encounter with the stimulus. This means that if a stimulus is not met more than once in $m-j$ steps, it is forgotten. Otherwise, the stimulus stays in memory. This is a simplified view of forgetting, but the details of the forgetting mechanism do not influence the qualitative results of the model. However, note that the presence of forgetting is key for a learning system, because it allows to encode new stimuli as they are encountered by an individual. With this rule, the dynamic memory will never contain more than $m-j$ stimuli at the same time as required by the definition of the model. Importantly, it is assumed that when a stimulus is removed from memory, all the associated information is lost. If this stimulus is encountered later, then the individual will have to re-learn to interact with it.

The details of the learning rule are not critical for our main purpose but an example of how the learning rule could be implemented may help to give a better grasp of the model. For this, one could simply consider that the animal uses a linear operator updating rule for preferences (motivations) over actions, and take the action that maximizes these preferences with some perturbation due to exploration or error, as is described in eqs. 2.1–2.4 of Chapter 2, which leads to expressions for the vector $(q_t(a,s))_{(a,s)\in A\times F}$ of taking actions associated to the stimuli in the dynamic memory.

4.2.4 Model summary

Summarizing the description of this model, the genotype of an individual is

$$\mathbf{g}=(m,j,G,(\mathbf{z}(s))_{s\in G},\phi,\lambda), \tag{4.3}$$

which belongs to the space $\mathbf{G}$ of possible genotypes. In the analysis of the model, we will mainly be interested in the optimal value of j (how blank should the slate be at birth?), thus the other genotypic values will be considered as parameters of the model. One can also describe succinctly the state of an individual at time t of lifespan as its dynamics probabilities $(q_t(a,s))_{(a,s)\in A\times F}$ of taking learned actions.

The number of offspring produced by an individual (fecundity) will be assumed to depend on

the average payoff it acquires during lifespan, and we will be interested in the long term average

$$f = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \pi(a_t, s_t), \tag{4.4}$$

which can be thought of as the expected number of offspring produced by an individual. Hence, natural selection favors the individuals with genotype $\mathbf{g}$ that have maximum reproductive output, that is the genotype that solve the maximization problem

$$\max_{\mathbf{g}\in\mathbf{G}} f(\mathbf{g}). \tag{4.5}$$

4.3 Evolution of the optimal j value

Here, a number of additional assumptions are described that will simplify the model in order to focus on the main point of this model: environmental complexity, that is the number, k, of stimuli in the environment, can select for having an intermediate value of j (i.e., $0 < j < m$), which corresponds to individuals who have behavioral reaction norms but are also able to learn to react to several stimuli.

4.3.1 Homogeneous environment

It is assumed that the environment is completely homogeneous, whereby

— $\theta_s = \theta_{s'} = \theta$,

— $x_s = x_{s'} = x$ and $y_s = y_{s'} = y$,

for all $s, s' \in S$. This assumption of homogeneity greatly simplifies the model because the content of G is now irrelevant: we do not need to know which stimuli are innately recognized by the animal, since all stimuli are equivalent in terms of effect on fitness. Note, however, that the stimuli are still different, because the θ actions that are optimal for stimulus s can be different from the θ actions that are optimal for stimulus s'. Hence, a behavioral reaction norm is still the minimum requirement; playing always the same action (purely genetic behavior) is not adapted to this environment.

4.3.2 Optimal reaction norms

Further, it will be assumed that the behavioral reaction norms are optimal. This assumption might seem strong, but since our goal is to look for the conditions that favor the evolution of

learning as opposed to behavioral reaction norms, it is obvious that if learning can evolve in the presence of optimal reaction norms, it will *a fortiori* evolve in the presence of suboptimal reaction norms.

4.4 Analysis

It is now assumed that only j evolves, that is, an individual is characterized by the number of memory slots dedicated to innate vs. learned responses to stimuli.

4.4.1 Simplifications for payoffs

Letting an individual interact a very long time with its environment ($T \to \infty$), the behavioral dynamics will eventually enter into a stationary state and we will evaluate expected payoffs in this state. Because the optimal reaction norm is independent of the stimulus, we can write the expected payoffs when a response to a stimulus is innate as $\pi_G = \mathbb{E}[\pi(a_t, s_t)|s_t \in G] = x$. The expected payoff obtained by taking actions randomly (independently of time t) is $\pi_U = \mathbb{E}[\pi(a_t, s_t)|s_t \in U] = (\theta/n)x + (1 - [\theta/n])y$. Finally, let $\pi_F = \mathbb{E}[\pi(a_t, s_t)|s_t \in F]$ be the expected payoff obtained when a stimulus is in the dynamic memory. This expression is very complicated to evaluate explicitly, but under reasonable assumptions on the learning system, the payoff for learning will be higher than the one by taking actions at random $\pi_F > \pi_U$. Another reasonable assumption is that $\pi_F < x$, because the average learning payoff must be a convex combination of x and y. It is also noteworthy that π_F will generally depend on the genotype j, because the number of memory slots dedicated to learning affects the number of time steps that a stimulus stays in memory (and the more a stimulus stays in memory, the greater the opportunities to learn the optimal action for that particular stimulus). However, we will generally ignore this effect of j on π_F. This assumption might restrict the conditions under which learning can evolve, but the results presented below suggest that it unlikely totally impedes the evolution of a learning ability.

In the stationary state, we also have a constant probability $P_G = \mathbb{P}\{s_t \in G\}$ that a current stimulus, s_t, is innately encoded, a constant probability $P_F = \mathbb{P}\{s_t \in F|s_t \notin G\}$ that the current stimulus is remembered, given it is not innate (i.e., s_t is present in the dynamic memory), and $1 - P_F = \mathbb{P}\{s_t \notin F|s_t \notin G\}$ is the probability that s_t is not remembered, given it is not innate.

With the above, the expected payoff to a focal individual with memory j is

$$f(j) = P_G \pi_G + (1 - P_G)\left[P_F \pi_F + (1 - P_F)\pi_U\right]. \tag{4.6}$$

As will be seen below, the expression for P_F depends on the size of the dynamic memory $(m - j)$ as well as on the details of the environmental pattern of encounters with stimuli. However, P_G is always the same because of the assumption that the environment is Markovian and is in a stationary uniform distribution, i.e.,

$$P_G = \frac{j}{k}, \tag{4.7}$$

whereby $(1 - P_G) = (k - j)/k$. Thus, the expected payoff is

$$f(j) = \frac{j}{k}x + \frac{k - j}{k}\left[P_F \pi_F + (1 - P_F)\pi_U\right]. \tag{4.8}$$

In order to be able to maximize $f(j)$, one must find an explicit expression for P_F. Two special cases of environmental pattern of encounters with stimuli are studied now (which are special cases of the general Markovian environment postulated in this paper) and the probability P_F is derived in each case.

4.4.2 Case 1: i.i.d. stimulus draw

Let us first consider that the stimuli are generated according to a simple random process where, conditional on a given stimulus sent by the environment, each stimulus has the same probability $1/k$ to be met at the next time step. This will be called an "i.i.d. process" (where "i.i.d." stands for "independently and identically distributed"), and thus there are no temporal correlations in stimuli.

In order to obtain the expression for P_F, one must look at the history of interactions with stimuli. Since our condition for forgetting a stimulus is similar to Kerr and Feldman (2003), this probability is given by their eq. 10, i.e.,

$$1 - \left(\frac{k - 1}{k}\right)^{\min[m-j,\,t-1]}. \tag{4.9}$$

Because we consider only the asymptotic limit of payoffs (eq. 4.4), the first $m - j$ steps of interaction with stimuli can be ignored, and the above expression simplifies and gives

$$P_F = 1 - \left(\frac{k - 1}{k}\right)^{m-j}. \tag{4.10}$$

A justification for this equation is as follows. Considering that t is bigger than $m-j$, our implementation of memory described above implies that the stimulus s_t met at time t is in F if and only if it has been encountered in one of the last $m-j$ steps. On the other hand, s_t is not in F if it was not met in the last $m-j$ steps. It is easier to compute the probability of this last case, $s_t \notin F$. For s_t to not be in F, one needs that s_t is not met during the last $m-j$ steps. The probability that s_t is not met on one of these steps is $(k-1)/k$. Since we need that s_t is not met on $m-j$ such steps and since stimulus draws are independent across time steps (because of the assumption of i.i.d. stimulus draw), the probability that s_t is not in F is $1-P_F = \left(\frac{k-1}{k}\right)^{m-j}$. The probability to remember s_t is obtained by taking the complementary probability, hence eq. 4.10. We see that provided that $t > m-j$, the probability to remember a stimulus does not depend explicitly on time, as required by the assumption used to derive the equation for expected average payoff (eq. 4.6).

Using eq. 4.10, the average payoff can be written as

$$\begin{aligned} f(j) &= \frac{j}{k}x + \frac{k-j}{k}\left((1-P_F)\pi_U + P_F\pi_F\right), \\ &= \frac{j}{k}x + \frac{k-j}{k}\left[\left(\frac{k-1}{k}\right)^{m-j}\pi_U + \left(1-\left(\frac{k-1}{k}\right)^{m-j}\right)\pi_F\right]. \end{aligned} \tag{4.11}$$

The j value satisfying the first order condition $\mathrm{d}f(j)/\mathrm{d}j = 0$ for a maximum can be expressed in terms of a Lambert function, but there is no real value of j between 0 and m that satisfies this equation. Consequently, there are no local optima in $f(j)$, so it is either monotonically increasing or decreasing. A simple derivation shows that $f(m) > f(m-1)$, hence f is monotonically increasing, thus the optimal j is $j^* = m$ (Fig. 4.1).

This model thus yields a negative result: learning does not evolve when the encounters with environmental stimuli are totally random. In the next section, a simple, but abstract condition for natural selection to favor an intermediate value of j (i.e., $0 < j^* < m$) is provided.

4.4.3 Case 2: where a minimum memory size is required to remember stimuli

In the previous section, it was shown that when the stimuli follow an i.i.d. process, the probability P_F to remember a stimulus on a second encounter is a decreasing function of j. Here, leaving out the details of the pattern of encounters with environmental stimuli, this probability is assumed to be a threshold function of j.

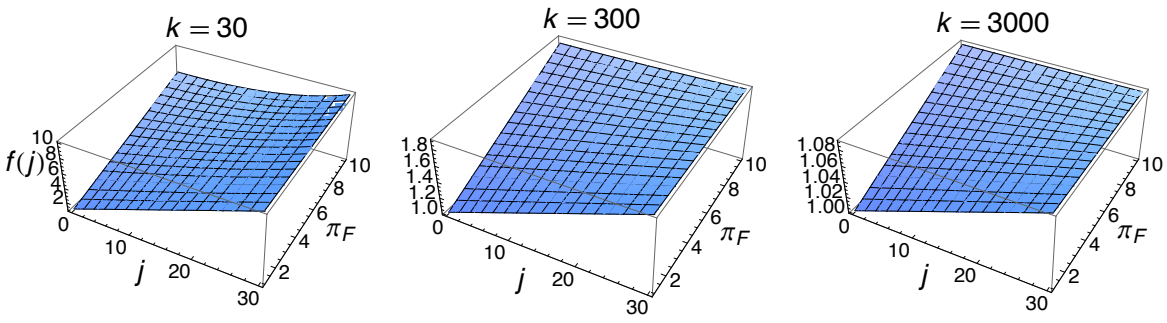

Figure 4.1: Average payoff, $f(j)$, in the i.i.d. model (eq. 4.11) as a function of the number of innate responses to stimuli, j, the average payoff for learned responses to stimuli, π_F, and the complexity of the environment, k. Parameter values: $m = 30, \pi_U = 1, x = 10$.

Formally, the assumption here is that there is a certain i $(0 < i < m)$ such that

$$P_F = \begin{cases} 0, & \text{if } j > i, \\ c, & \text{if } j \leq i, \end{cases} \tag{4.12}$$

where c is a constant $(0 < c < 1)$. In other words, an individual must have a dynamic memory (F) of size at least $m-i$ slots in order to have a chance to remember interactions with a stimulus; and there is no memory gain in increasing the size of the dynamic memory provided it is of size at least $m - i$ (a possible example where such a condition might be verified is a cyclic environment, where interactions with a given stimulus are separated by relatively long incompressible intervals). With this, it shall be shown that there exists a k sufficiently large such that $j^* = i$.

The first step is to notice that the total payoff for an individual with $j > i$ simplifies to

$$f(j) = \frac{j}{k}x + \frac{k-j}{k}\pi_U, \tag{4.13}$$

because of the assumption that $P_F = 0$ if $j > i$ (eq. 4.8). Among all individuals who have a $j > i$, $f(j)$ is strictly increasing, thus the best possible j is simply $j = m$, that is an individual should use all its memory for behavioral reaction norms.

For an individual with $j \leq i$, the total payoff is

$$f(j) = \frac{j}{k}x + \frac{k-j}{k}\left(c\pi_U + (1-c)\pi_F\right). \tag{4.14}$$

Among all individuals who have a $j \leq i$, again the relation $f(j+1) > f(j)$ holds, thus the individual with $f(i)$ has the optimal payoff among those. In order to show that $j^* = i$, it now

remains to prove that the inequality $f(i) > f(m)$ has a solution. Solving this inequality for k yields

$$k > \frac{i(c\pi_U + (1-c)\pi_F) + (m-i)x - m\pi_U}{(1-c)(\pi_F - \pi_U)}. \tag{4.15}$$

The right hand side of this expression will be called $k_{\min}$ so that this inequality becomes $k > k_{\min}$. This condition describes the minimum value of k for j^* to be equal to i, as a function of the parameters of the model. In particular, $k_{\min}$ is decreasing in i, which means that more complex environments select for a greater reliance on learning. In Fig. 4.2, the dependence of $k_{\min}$ on the model parameters is shown. In particular, it is a decreasing function of π_F, because this parameter can be thought as measuring the efficiency of the learning rule of the animal. The closer is π_F to the optimal payoff x, the more efficient the learning rule is and the easier it is for learning to evolve. The probability $P_F = c$ represents the difficulty of the environment in terms of pattern of encounters with stimuli (low values of c correspond to a small chance to remember stimuli and thus a "difficult" environment), and thus $k_{\min}$ is decreasing in c. The abrupt dependence of P_F on j seems to be important for the evolution of learning (Fig. 4.3).

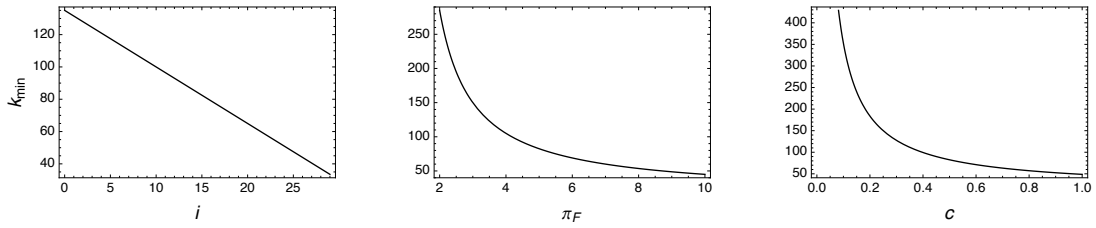

Figure 4.2: Dependence of $k_{\min}$ (eq. 4.15) on the model parameters in Case 2. When not varied, parameter values are: $m = 30, i = 15, d = 1, x = 10, \pi_F = 5, c = 0.5$.

4.5 Summary of results

In this work, the question of the evolution of learning as an alternative to innate responses to variable stimuli has been investigated. The distinguishing feature of the present research is that it tries to account for a wide range of situations/stimuli faced by an organism, in contrast to previous models for the evolution of learning that considered only one situation (or stimulus) under many environmental conditions. In the model, the decision system of an animal is viewed as a limited number of associations between stimulus and action. Our goal was to show that

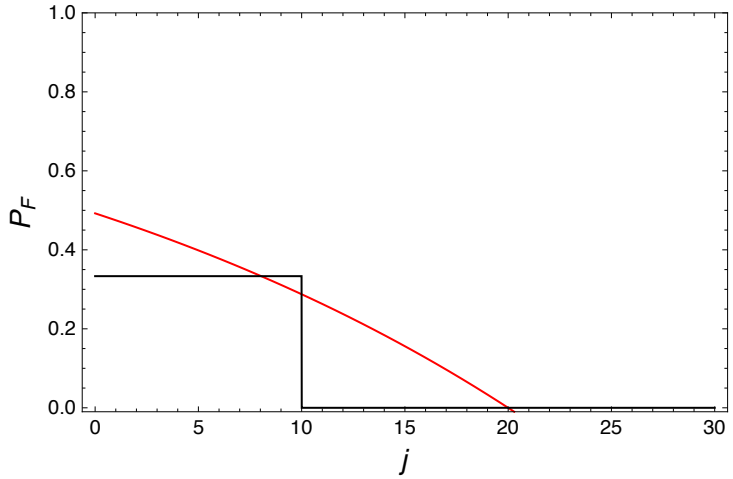

Figure 4.3: Probability, P_F, to recall a stimulus as a function of the number of memory slots, j, dedicated to innate responses to stimuli. Red: Case 1 (eq. 4.10); Black: Case 2 (eq. 4.12). Parameter values: $k = 30, m = 20, \pi_U = 1, x = 10, \pi_F = 5, c = 1/3, i = 20$.

given this limit on memory (the parameter m), a process of learning and forgetting could allow to deal with a great number of environmental stimuli (the parameter k), and would thus be selected for.

Several simplifying assumptions were made in order to obtain analytical tractability. Most importantly, it was assumed that all stimuli had an equal effect on fitness, and we considered that lifespan was long enough so that an individual has interacted with a great number of environmental stimuli and reached a behavioral equilibrium. Finally, it was postulated that individuals had optimal innate reactions norms, and this assumption was made on the ground that in the absence of learning, optimal reaction norms was the only possible evolutionary endpoint in the modeled environment.

Two particular cases of environmental pattern of encounters with stimuli were studied. In the first case, it was assumed that there was no particular structure to the environment so that an individual encounters stimuli at random. The result obtained in this case is negative: learning does not evolve, because the benefit of investing one slot of memory to learning was too small compared to the effect of obtaining the optimal payoff for one innately encoded stimulus, even if only a small number of stimuli-response situations can be genetically encoded. The low effect of investing a slot to the dynamic memory in complex environments is due to the fact that in our random setting of encounters with stimuli, the bigger is k the smaller is P_F (and $\lim_{k\to\infty} P_F = 0$),

whatever the value of j is.

In a second case, an abstract leap was taken, and no explicit environmental dynamics was considered, but rather a general assumption on the relation between the probability to recall a stimulus in memory and the size of the dynamic memory was used. Namely, in this case an individual must have a minimum size of the dynamic memory in order to be able to remember interactions with stimuli. This assumption allowed to derive a general result on the evolution of learning: as long as the environment is complex enough (i.e., if k is big enough), then any combination of reliance on learning vs. innate reaction norms can be achieved. The empirical significance of this result remains to be demonstrated however, because we could not find an example of environmental dynamics supporting it.

Looking ahead, many additional features of real environments would be interesting to study in relation to the evolution of learning in this model. First, it would be informative to see what happens if one considers more heterogeneous environments, where stimuli vary in their quality, and their frequency in an animal's environment. It seems indeed reasonable to think that frequent stimuli having an important effect on fitness should have an innate, immediate response, while for other less important stimuli, an animal can take the risk to lose time and energy in learning. However, the stimuli less frequently met are more difficult to learn because an individual usually needs to experiment many actions to find a good one. How these two effects interact to generate the evolution of learning and memory? Another important area of study would be to investigate how environmental complexity (the magnitude of k) interact with environmental fluctuations. It is well established that environmental fluctuations can select for the evolution of learning. In the present model, environmental fluctuations could be implemented by changing the fitness consequences of actions applied to stimuli. Does environmental complexity enhance or inhibit the evolution of learning under fluctuating environments? All these questions open the road to more general models of the evolution of learning that take into account relevant features of real environments.

Chapter 5

Conclusion

This thesis began with the aim to assess the evolutionary significance of specific learning mechanisms, based on observed behavior of animals, that can be used to cope with a wide range of ecologically relevant situations. A special focus was given to social interactions, because biological fitness very often depends on the actions of other population members. We also tried to find an extended set of conditions under which one can expect learning abilities to be favored by natural selection.

5.1 Summary of results and discussion

5.1.1 Specificities of learning rules

Because animals and humans use the same brain to cope with all the decisions they face, it is argued in this thesis (following a recent trend of research in animal behavior, McNamara and Houston, 2009; Fawcett et al., 2013; Lotem, 2013) that the learning abilities of a given organism must share common features to deal with the variety of social and non-social circumstances proposed by its ecological niche. In Chapter 2, mathematical results from game theory and stochastic approximation were used to follow this approach. The goal of this part of the thesis was to contribute to building a mapping from cognitive abilities to produced behavior. Empiricists interested in animal behavior generally face the inverse problem: given observed actions of a species, what can be said about its psychological, cognitive, and computational abilities? Years of field observations and experimental investigations have led behavioral ecologists to propose a variety of mechanisms, from simple ones like the law of effect, to more involved ones like

Bayesian learning, but it remains difficult to assess the plausibility of these mechanisms unless one studies what they really entail in terms of the shape of learning curves and endpoints of behavioral dynamics. Moreover, evolutionary biologists also need to know the phenotypes produced by these rules so that they can formulate statements about their evolutionary significance.

A number of learning rules from behavioral ecology and game theory were thus considered in the context of general social (and also non-social) decision problems. The learning rules studied were members of a general class of rules parametrized by several psychological characteristics or cognitive abilities. That is, we considered that animals could vary in their tendency to explore or be impulsive, in their abilities to infer payoffs of actions not tried, and also in their memory capacities. It was assumed that environments were fluctuating so that the specific situation encountered by an individual was possibly different at every decision step. The features of the environment that produced these variations were considered to be undetectable by the individuals. All information individuals could get was related to the payoffs of actions in their behavioral repertoire. To illustrate the predictive power of this approach, it was applied to two particular questions: (1) What is the behavioral outcome of the Hawk-Dove game played by learners of different types? (2) Does the sophisticated Bayesian Learning rule outcompete reinforcement learning in a foraging task, in the presence of scroungers (producer-scrounger game)? Question (1) was behavioral only, while question (2) was evolutionary, as we modeled natural selection on these two learning rules.

Equations approximating the long-run behavior of individuals interacting in social games were obtained and these equations provided a decomposition of behavior into the tendency of individuals to exploit actions leading to high payoff and the tendency to explore novel actions (eq. 2.13). Importantly, the "average game" played by individuals in the fluctuating environment determined long-run behavior. Comparison of approximation equations with simulations in the Hawk-Dove game showed that the approximation was quite good for individuals who have only a small tendency to explore, i.e., it works better for impulsive individuals. We also showed that the approximation concords to real stochastic behavior of the rules when time is large, especially if individuals display preference reversal in the course of learning. Preference reversal is defined by the fact that an action for which an animal has a genetic predisposition becomes less preferred after the animal acquires knowledge of current payoffs during learning. In the producer-scrounger model (question (2) above), the role of the exploration tendency of individuals was shown to be critical to determine the fate of co-evolving learning rules.

One of the psychological features modeled in the rules of Chapter 2 is the ability to infer payoffs (or rewards) of actions not tried by an individual. This entails that an animal is able to imagine (or mentally simulate) situations not physically perceived, in order to reason about alternative behaviors (Emery and Clayton, 2001, 2004; Taylor et al., 2012). The sophisticated strategies of food caching in corvids provide an example, where it is thought that animals can imagine that others may pilfer their caches (Emery and Clayton, 2001). More generally, it remains controversial to what extent animals can reason and deduce from environmental cues the rewards of actions that they do not explicitly try.

Chapter 3 thus focused on what may be called inference-based learning and reinforcement learning. In inference-based learning, an individual is able to compute the payoffs of actions it did not explicitly try and updates its motivations (or preferences over actions) accordingly. In reinforcement learning, an animal can only know payoffs associated to experienced actions. Since social interactions are thought to have driven the evolution of advanced cognition (Humphrey, 1976; Alexander, 1979; Gavrilets and Vose, 2006), we tested the idea that pairwise social interactions favor the ability to infer missed payoffs. These pairwise games varied in the same way as in Chapter 2, and we analyzed three cases of average game: Prisoner's Dilemma, Hawk-Dove game, and Coordination game. Natural selection on the two learning rules (inference-based and reinforcement learning) was modeled and it was asked what is the equilibrium frequency of these rules in a large population.

Results depended on the type of average game faced by individuals and also on the matching scheme (i.e., the presence or absence of repeated interactions between the same individuals). Of the different cases of environment and matching investigated, a majority led to the evolution of inference-based learning to a high equilibrium frequency. This suggests that the capacity to infer missed payoffs indeed gives a fitness advantage in most situations modeled in the Chapter 3, but several examples showed that reinforcement learning could also be the evolutionarily stable learning rule. For instance, when the average game was a Prisoner's Dilemma, pairs of repeatedly interacting reinforcement learners were able to learn to cooperate and extract the maximal mutual payoffs out of this social dilemma. Empirical investigation is needed to assess what are the ecological situations faced by animals among the variety of situations we tried to capture, but our model gives intuition about why trial-and-error learning is so common in animals, since it often resisted invasion by the sophisticated inference-based learning strategy.

5.1.2 Origins of learning

In the last part of this thesis (Chapter 4), a broader view of the evolution of learning was adopted, and the question of how learning abilities can emerge from other simpler forms of phenotypic plasticity was tackled. Because learning is a special type of reaction norm, it might be justified by appealing to the general conditions that favor the evolution of reaction norms (West-Eberhard, 1989; Gavrilets and Scheiner, 1993; Pigliucci, 2001). However, because of its specificities that distinguish it from lower levels of plasticity, learning requires special environmental patterns in order to be selected for. Previous research largely ignored this evolutionary transition, because in theoretical models, learning is generally opposed to purely genetically determined behavior. The intermediate step that consists of innate responses to stimuli is not accounted for. It is difficult to know why this step is not modeled: it might be because researchers just focus on a particular stimulus or situation, and implicitly consider that learning evolves from these innate responses; or because there is a conceptual misunderstanding on the levels of plasticity.

It was argued in Chapter 4 that because of this conceptual fuzziness, an important factor in the evolution of learning and forgetting has been overlooked, namely environmental complexity. What is meant by "environmental complexity" is the fact that biological organisms may interact with a very large number of stimuli (or situations) in the course of their lifespan. Importantly, the number of potential situations in one's environment may be larger than what can be encoded in a single brain, so individuals must resort to a higher level of plasticity than innate responses to stimuli. This higher level is learning, which, combined with forgetting (Kraemer and Golding, 1997; Kerr and Feldman, 2003), allow a dynamic process where stimuli enter and leave memory sequentially. In this way an individual can respond to an arbitrarily high number of stimuli as long as it has the time to learn the optimal solution for each single stimulus encountered. A mathematical model was developed to test the validity of this verbal argument where the genotype of an individual was conceptualized as the proportion of total memory size dedicated to learning vs. innate responses to stimuli.

It was found that an ability to learn to associate stimuli with optimal actions evolves only under special assumptions on the environment and the learning system. Importantly, an environment where stimuli are encountered completely randomly, without any particular structure, cannot select for learning. The reason is that when the number of stimuli is larger than the size of the animal's memory, the probability to interact with a stimulus twice in a short period of

time is low and learning hardly happens. On the other hand, an abstract condition was derived in a second environmental pattern. When a minimum memory size is required to remember interactions with stimuli, the corresponding amount of memory size dedicated to learning can be made evolutionarily stable as long as the environment is complex enough.

5.2 Essential contributions and outlook

The contributions described above were informative, but many questions and problems remain open for future research. A glimpse of the possible follow-up for each part of the present work is now presented.

The main contribution of Chapter 2 was to show that it is possible to analyze the behavioral dynamics produced by many learning rules relevant to evolutionary biology and animal behavior. In previous work in evolutionary theory, these rules were mainly studied via agent-based computer simulations, so the generality of these results is hard to assess (Groß et al., 2008; Josephson, 2008; Hamblin and Giraldeau, 2009; Arbilly et al., 2010, 2011*a*,*b*; Katsnelson et al., 2011). With analytical approximations, more can be said but one must be cautious regarding a number of elements. First, the method of stochastic approximation (Ljung, 1977; Fudenberg and Levine, 1998) used to derive equations of motion for learning dynamics only guarantee asymptotic convergence, but do not say much about finite time behavior. This is problematic since learning in a short time is likely to provide a selective advantage, but this effect cannot be tested with the method used in this work (Hopkins, 2002; Leslie and Collins, 2005; Izquierdo et al., 2007). Second, social learning, whereby an individual copy the actions of conspecifics, is an important related mechanism that we did not include, but may be relevant to compare to individual learning rules since it can allow to acquire information at a lower cost (Schlag, 1998; Laland, 2004; Rendell et al., 2010), for example when scroungers follow producers on food patches (Hamblin and Giraldeau, 2009).

Chapter 3 showed that reinforcement learning (or trial-and-error) is not necessarily outcompeted by sophisticated learning rules, and can generate cooperation in social dilemmas like the Hawk-Dove (or Snowdrift) game, and the Prisoner's Dilemma. However many cases displayed a superiority of inference-based learning over reinforcement learning and this suggests that there might be an evolutionary arms race in social cognition, because more complex rules than inference-based learning can be imagined, for instance rules that try to also infer the type of game under the fluctuating environment (Mengel, 2012). Where this arms race will stop is a

question for future research even if recent developments in evolutionary game theory indicate that individuals with low cognitive abilities may co-exist with more sophisticated types (Mohlin, 2012), even when sophisticated types pay no cost for cognitive complexity.

The evolutionary transition from behavioral reaction norms to learning was addressed in Chapter 4 and it was found that a large number of stimuli (or situations) in an animal's environment can drive the evolution of learning. But, it remains an open challenge to fully understand the interaction between environmental fluctuations and complexity, since these factors are likely to co-occur in nature. Moreover, we made several simplifying assumptions, such as equal effects on fitness of all environmental stimuli, that should be relaxed in future research to test the robustness, applicability, and generality of the results.

Overall, this thesis proposes a detailed view of the evolution of learning, integrating at the same time mechanisms and natural selection. In particular, we accounted for both behavioral and generational time scales because this is necessary in order to have a more complete picture of biological dynamics occurring in nature. We saw that the way in which organisms gather and use information cannot be overlooked in evolutionary arguments on the usefulness of a learning ability, and we laid the basis for bridging the gap that separates research on general phenotypic plasticity from studies of learning, memory, and advanced cognitive abilities.

Bibliography

Achbany, Y., F. Fouss, L. Yen, A. Pirotte, and M. Saerens. 2006. Optimal tuning of continual online exploration in reinforcement learning. In S. Kollias, A. Stafylopatis, W. Duch, and E. Oja, eds., Artificial Neural Networks – ICANN 2006, volume 4131 of *Lecture Notes in Computer Science*, pages 790–800. Springer Berlin / Heidelberg.

Alexander, R. D. 1979. Darwinism and Human Affairs. University of Washington Press, Seattle, WA.

Anderson, S. P., A. d. d. Palma, and J.-F. Thisse. 1992. Discrete Choice Theory of Product Differentiation. 1 edition. The MIT Press.

André, J.-B. 2010. The evolution of reciprocity: social types or social incentives? The American naturalist 175:197–210.

Arbilly, M., U. Motro, M. W. Feldman, and A. Lotem. 2010. Co-evolution of learning complexity and social foraging strategies. Journal of Theoretical Biology 267:573–581.

———. 2011*a*. Evolution of social learning when high expected payoffs are associated with high risk of failure. Journal of The Royal Society Interface 8:1604–1615.

———. 2011*b*. Recombination and the evolution of coordinated phenotypic expression in a frequency-dependent game. Theoretical Population Biology 80:244–255.

Arnold, S. J. 1978. The evolution of a special class of modifiable behaviors in relation to environmental pattern. The American Naturalist 112:415–427.

Axelrod, R. 1980. Effective choice in the prisoner's dilemma. The Journal of Conflict Resolution 24:3–25.

Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. Science 211:1390–1396.

Benaim, M. 1999. Dynamics of stochastic approximation algorithms. In Séminaire de Probabilités XXXIII, volume 1709, J. Azéma et al. edition, pages 1–68. Springer, Berlin.

Benaim, M., and N. El Karoui. 2005. Promenade aléatoire : Chaînes de Markov et simulations ; martingales et stratégies. Ecole Polytechnique.

Benaim, M., and M. W. Hirsch. 1999*a*. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. Games and Economic Behavior 29:36–72.

———. 1999*b*. Stochastic approximation algorithms with constant step size whose average is cooperative. The Annals of Applied Probability 9:216–241.

Benveniste, A., M. Metivier, and P. Priouret. 1991. Adaptive Algorithms and Stochastic Approximations. Springer-Verlag.

Bernstein, C., A. Kacelnik, and J. R. Krebs. 1988. Individual decisions and the distribution of predators in a patchy environment. Journal of Animal Ecology 57:1007–1026.

Binmore, K. G., and L. Samuelson. 1992. Evolutionary stability in repeated games played by finite automata. Journal of Economic Theory 57:278–305.

Borenstein, E., M. W. Feldman, and K. Aoki. 2008. Evolution of learning in fluctuating environments: when selection favors both social and exploratory individual learning. Evolution 62:586–602.

Börgers, T., and R. Sarin. 1997. Learning through reinforcement and replicator dynamics. Journal of Economic Theory 77:1–14.

Borkar, V. S. 2008. Stochastic approximation: a dynamical systems viewpoint. Cambridge University Press, Cambridge.

Boyd, R., and P. J. Richerson. 1988. Culture and the Evolutionary Process. University of Chicago Press.

Brown, G. W. 1951. Iterative solution of games by fictitious play. In Activity analysis of production and allocation, pages 374–376. Wiley, New York.

Bush, R. R., and F. Mostelller. 1951. A mathematical model for simple learning. Psychological Review 58:313–323.

Camerer, C., and T. H. Ho. 1999. Experienced-weighted attraction learning in normal form games. Econometrica 67:827–874.

Camerer, C. F. 2003. Behavioral game theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L. L., and M. W. Feldman. 1983. Cultural versus genetic adaptation. Proceedings of the National Academy of Sciences of the United States of America 80:4993–4996.

Chalmeau, R. 1994. Do chimpanzees cooperate in a learning task? Primates 35:385–392.

Charnov, E. L. 1976. Optimal foraging, the marginal value theorem. Theoretical Population Biology 9:129–136.

Chasparis, G., J. Shamma, and A. Arapostathis. 2010. Aspiration learning in coordination games. In 2010 49th IEEE Conference on Decision and Control (CDC), pages 5756–5761.

Cho, I.-K., and A. Matsui. 2005. Learning aspiration in repeated games. Journal of Economic Theory 124:171–201.

Cournot, A. A. 1838. Recherches sur les principes mathématiques de la théorie des richesses. L. Hachette.

Dayan, P., and L. F. Abbott. 2005. Theoretical Neuroscience: Computational And Mathematical Modeling of Neural Systems. MIT Press.

Dickinson, A. 1980. Contemporary Animal Learning Theory. Cambridge University Press, Cambridge, UK.

Dijker, A. 2011. Physical constraints on the evolution of cooperation. Evolutionary Biology 38:124–143.

Dorosheva, E. A., I. K. Yakovlev, and Z. I. Reznikova. 2011. An innate template for enemy recognition in red wood ants. Entomological Review 91:274–280.

Dridi, S., and L. Lehmann. 2013. On learning dynamics underlying the evolution of learning rules. Theoretical Population Biology (in press).

Dugatkin, L. A. 2010. Principles of Animal Behavior. 2 edition. WW Norton & Co.

Dukas, R. 2004. Evolutionary biology of animal cognition. Annual Review of Ecology, Evolution, and Systematics 35:347–374.

Dunlap, A. S., and D. W. Stephens. 2009. Components of change in the evolution of learning and unlearned preference. Proceedings of the Royal Society B: Biological Sciences 276:3201–3208.

Emery, N. J., and N. S. Clayton. 2001. Effects of experience and social context on prospective caching strategies by scrub jays. Nature 414:443–446.

———. 2004. The mentality of crows: Convergent evolution of intelligence in corvids and apes. Science 306:1903–1907.

———. 2009. Comparative social cognition. Annual Review of Psychology 60:87–113.

Enquist, M. E., and S. Ghirlanda. 2005. Neural Networks And animal Behavior. Princeton University Press.

Erev, I., and A. E. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88:848–881.

Fawcett, T. W., S. Hamblin, and L.-A. Giraldeau. 2013. Exposing the behavioral gambit: the evolution of learning and decision rules. Behavioral Ecology 24:2–11.

Feldman, M., K. Aoki, and J. Kumm. 1996. Individual versus social learning: evolutionary analysis in a fluctuating environment. Anthropological Science 104:209–232.

Foster, D. P., and H. Young. 2003. Learning, hypothesis testing, and nash equilibrium. Games and Economic Behavior 45:73–96.

Fudenberg, D., and D. K. Levine. 1998. The Theory of Learning in Games. MIT Press, Cambridge, MA.

Fudenberg, D., and S. Takahashi. 2011. Heterogeneous beliefs and local information in stochastic fictitious play. Games and Economic Behavior 71:100–120.

Gale, J., K. G. Binmore, and L. Samuelson. 1995. Learning to be imperfect: The ultimatum game. Games and Economic Behavior 8:56–90.

Gavrilets, S., and S. M. Scheiner. 1993. The genetics of phenotypic plasticity. VI. theoretical predictions for directional selection. Journal of Evolutionary Biology 6:49–68.

Gavrilets, S., and A. Vose. 2006. The dynamics of machiavellian intelligence. Proceedings of the National Academy of Sciences 103:16823–16828.

Geisler, W. S., and R. L. Diehl. 2002. Bayesian natural selection and the evolution of perceptual systems. Philosophical Transactions: Biological Sciences 357:419–448.

Giraldeau, L.-A., and T. Caraco. 2000. Social Foraging Theory. Princeton University Press.

Gomulkiewicz, R., and M. Kirkpatrick. 1992. Quantitative genetics and the evolution of reaction norms. Evolution 46:390–411.

Gong, Z. 2012. Innate preference in drosophila melanogaster. Science China Life Sciences 55:8–14.

Grimmett, G. R., and D. R. Stirzaker. 2001. Probability and Random Processes. 3 edition. Oxford University Press.

Groß, R., A. I. Houston, E. J. Collins, J. M. McNamara, F.-X. Dechaume-Moncharmont, and N. R. Franks. 2008. Simple learning rules to cope with changing environments. Journal of The Royal Society Interface 5:1193 –1202.

Hamblin, S., and L.-A. Giraldeau. 2009. Finding the evolutionarily stable learning rule for frequency-dependent foraging. Animal Behaviour 78:1343–1350.

Hammerstein, P., and J. R. Stevens, eds. 2012. Evolution and the Mechanisms of Decision Making. MIT Press, Cambridge, MA.

Harley, C. B. 1981. Learning the evolutionarily stable strategy. Journal of Theoretical Biology 89:611–633.

Hart, S., and A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. Econometrica 68:1127–1150.

Heller, D. 2004. An evolutionary approach to learning in a changing environment. Journal of Economic Theory 114:31–55.

Herrnstein, R. J. 1970. On the law of effect. Journal of the Experimental Analysis of Behavior 13:243–266.

Hirsch, M. W., S. Smale, and R. L. Devaney. 2004. Differential equations, dynamical systems, and an introduction to chaos. Academic Press.

Ho, T. H., C. F. Camerer, and J.-K. Chong. 2007. Self-tuning experience weighted attraction learning in games. Journal of Economic Theory 133:177–198.

Hofbauer, J., and W. H. Sandholm. 2002. On the global convergence of stochastic fictitious play. Econometrica 70:2265–2294.

Hofbauer, J., and K. Sigmund. 1998. Evolutionary games and population dynamics. Cambridge University Press.

———. 2003. Evolutionary game dynamics. Bulletin of the American Mathematical Society 40:479–519.

Hollis, K., and L. Guillette. 2011. Associative learning in insects: Evolutionary models, mushroom bodies, and a neuroscientific conundrum. Comparative Cognition & Behavior Reviews 6:25–46.

Hollis, K. L., M. J. Dumas, P. Singh, and P. Fackelman. 1995. Pavlovian conditioning of aggressive behavior in blue gourami fish (trichogaster trichopterus): Winners become winners and losers stay losers. Journal of Comparative Psychology 109:123–133.

Hopkins, E. 2002. Two competing models of how people learn in games. Econometrica 70:2141–2166.

Houston, A., and J. M. McNamara. 1999. Models of adaptive behaviour. Cambridge University Press.

Houston, A. I. 1983. Comments on "Learning the evolutionarily stable strategy". Journal of Theoretical Biology 105:175–178.

Houston, A. I., and B. H. Sumida. 1987. Learning rules, matching and frequency dependence. Journal of Theoretical Biology 126:289–308.

Humphrey, N. K. 1976. The social function of intellect. In P. P. G. Bateson, and R. A. Hinde, eds., Growing Points in Ethology. Cambridge University Press, Oxford, England.

Izquierdo, L. R., S. S. Izquierdo, N. M. Gotts, and J. G. Polhill. 2007. Transient and asymptotic dynamics of reinforcement learning in games. Games and Economic Behavior 61:259–276.

Johnston, T. D. 1982. Selective costs and benefits in the evolution of learning. In R. A. H. Jay S. Rosenblatt, ed., Advances in the Study of Behavior, volume Volume 12, pages 65–106. Academic Press.

Jordan, J. S. 1991. Bayesian learning in normal form games. Games and Economic Behavior 3:60–81.

Josephson, J. 2008. A numerical analysis of the evolutionary stability of learning rules. Journal of Economic Dynamics and Control 32:1569–1599.

Karlin, S., and H. E. Taylor. 1975. A First Course in Stochastic Processes. Academic Press, San Diego, CA.

Katsnelson, E., U. Motro, M. W. Feldman, and A. Lotem. 2011. Evolution of learned strategy choice in a frequency-dependent game. Proceedings of the Royal Society B: Biological Sciences .

Kerr, B. 2007. Niche construction and cognitive evolution. Biological Theory 2:250–262.

Kerr, B., and M. W. Feldman. 2003. Carving the cognitive niche: Optimal learning strategies in homogeneous and heterogeneous environments. Journal of Theoretical Biology 220:169–188.

Kraemer, P. J., and J. M. Golding. 1997. Adaptive forgetting in animals. Psychonomic Bulletin & Review 4:480–491.

Krebs, J. R., N. B. Davies, and S. A. West. 1993. An Introduction to Behavioural Ecology. 3 edition. Wiley-Blackwell.

Kushner, H. J., and G. G. Yin. 2003. Stochastic Approximation and Recursive Algorithms and Applications. 2nd edition. Springer.

Lahkar, R., and R. M. Seymour. 2013. Reinforcement learning in population games. Games and Economic Behavior 80:10–38.

Laland, K. N. 2004. Social learning strategies. Learning & behavior 32:4–14.

Leimar, O., and P. Hammerstein. 2001. Evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society B: Biological Sciences 268:745–753. PMID: 11321064 PMCID: PMC1088665.

Leslie, D. S., and E. J. Collins. 2005. Individual q-learning in normal form games. SIAM Journal on Control and Optimization 44:495–514.

Ljung, L. 1977. Analysis of recursive stochastic algorithms. IEEE Transactions on Automatic Control 22:551–575.

Lotem, A. 2013. Learning to avoid the behavioral gambit. Behavioral Ecology 24:13–13.

Luce, R. D. 1959. Individual Choice Behavior. Wiley, New York.

Luce, R. D., and H. Raiffa. 1989. Games and Decisions: Introduction and Critical Survey. New edition edition. Dover Publications Inc.

Macy, M. W., and A. Flache. 2002. Learning dynamics in social dilemmas. Proceedings of the National Academy of Sciences 99:7229–7236.

Mameli, M., and P. Bateson. 2006. Innateness and the sciences. Biology and Philosophy 21:155–188.

Maynard-Smith, J. 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.

Maynard-Smith, J. M., and G. R. Price. 1973. The logic of animal conflict. Nature 246:15–18.

McElreath, R. 2010. Baryplot 1.0.

McElreath, R., and R. Boyd. 2007. Mathematical Models of Social Evolution: A Guide for the Perplexed. University Of Chicago Press, Chicago, IL.

McKelvey, R. D., and T. R. Palfrey. 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10:6–38.

Mcnamara, J. M., and A. I. Houston. 1985. Optimal foraging and learning. Journal of Theoretical Biology 117:231–249.

McNamara, J. M., and A. I. Houston. 1987. Memory and the efficient use of information. Journal of Theoretical Biology 125:385–395.

———. 2009. Integrating function and mechanism. Trends in Ecology & Evolution 24:670–675.

Mengel, F. 2012. Learning across games. Games and Economic Behavior 74:601–619.

Mery, F., and T. J. Kawecki. 2002. Experimental evolution of learning ability in fruit flies. Proceedings of the National Academy of Sciences 99:14274–14279.

———. 2004. The effect of learning on experimental evolution of resource preference in drosophila melanogaster. Evolution 58:757–767.

Mohlin, E. 2012. Evolution of theories of mind. Games and Economic Behavior 75:299–318.

Niv, Y. 2009. Reinforcement learning in the brain. Journal of Mathematical Psychology 53:139–154.

Norman, M. F. 1968. Some convergence theorems for stochastic learning models with distance diminishing operators. Journal of Mathematical Psychology 5:61–101.

Nowak, M. A. 2006. Evolutionary dynamics: exploring the equations of life. Harvard University Press.

Pemantle, R. 1990. Nonconvergence to unstable points in urn models and stochastic approximations. The Annals of Probability 18:698–712.

Pigliucci, M. 2001. Phenotypic Plasticity: Beyond Nature and Nurture. Johns Hopkins University Press.

———. 2005. Evolution of phenotypic plasticity: Where are we going now? Trends in Ecology & Evolution 20:481–486.

Plotnik, J. M., R. Lair, W. Suphachoksahakun, and F. B. M. de Waal. 2011. Elephants know when they need a helping trunk in a cooperative task. Proceedings of the National Academy of Sciences .

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1:515–526.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rapoport, A., and A. M. Chammah. 1965. Prisoner's Dilemma. University of Michigan Press.

Rendell, L., R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. Laland. 2010. Why copy others? insights from the social learning strategies tournament. Science 328:208–213.

Rescorla, R. A., and A. R. Wagner. 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory, a. h. black and w. f. prokasy edition, pages 64–99. Appleton-Century-Crofts, New York (NY).

Riffell, J. A., R. Alarcón, L. Abrell, G. Davidowitz, J. L. Bronstein, and J. G. Hildebrand. 2008. Behavioral consequences of innate preferences and olfactory learning in hawkmoth–flower interactions. Proceedings of the National Academy of Sciences 105:3404–3409. PMID: 18305169.

Rodriguez-Gironés, M. A., and R. A. Vásquez. 1997. Density-ependent patch exploitation and acquisition of environmental information. Theoretical Population Biology 52:32–42.

Rogers, A. R. 1988. Does biology constrain culture? American Anthropologist 90:819–831.

Sandholm, W. H. 2011. Population Games and Evolutionary Dynamics. MIT Press.

Schlag, K. H. 1998. Why imitate, and if so, how? Journal of Economic Theory 78:130–156.

Shettleworth, S. J. 2009. Cognition, evolution, and behavior. Oxford University Press.

Shettleworth, S. J., J. R. Krebs, D. W. Stephens, and J. Gibbon. 1988. Tracking a fluctuating environment: a study of sampling. Animal Behaviour 36:87–105.

Stephens, D. W. 1991. Change, regularity, and value in the evolution of animal learning. Behavioral Ecology 2:77–89.

Stephens, D. W., and K. C. Clements. 1998. Game theory and learning. In Game Theory and Animal Behavior, pages 239–260. Oxford University Press, New York.

Sutton, R. S., and A. G. Barto. 1998. Reinforcement Learning: An Introduction. MIT Press.

Taylor, A. H., R. Miller, and R. D. Gray. 2012. New caledonian crows reason about hidden causal agents. Proceedings of the National Academy of Sciences page 201208724.

Thorndike, E. L. 1911. Animal Intelligence. Hafner, Darien, CT.

Tracy, N. D., and J. W. Seaman. 1995. Properties of evolutionarily stable learning rules. Journal of Theoretical Biology 177:193–198.

Tuyls, K., K. Verbeeck, and T. Lenaerts. 2003. A selection-mutation model for q-learning in multi-agent systems. In Proceedings of the second international joint conference on Autonomous agents and multiagent systems, pages 693–700. ACM, Melbourne, Australia.

van der Horst, W., M. van Assen, and C. Snijders. 2010. Analyzing behavior implied by EWA learning: An emphasis on distinguishing reinforcement from belief learning. Journal of Mathematical Psychology 54:222–229.

Villarreal, R., and M. Domjan. 1998. Pavlovian conditioning of social affirmative behavior in the mongolian gerbil (meriones unguiculatus). Journal of Comparative Psychology 112:26–35.

Wakano, J. Y., and N. Yamamura. 2001. A simple learning strategy that realizes robust cooperation better than pavlov in iterated prisoners' dilemma. Journal of Ethology 19:1–8.

Wakano, J. Y., K. Aoki, and M. W. Feldman. 2004. Evolution of social learning: a mathematical analysis. Theoretical Population Biology 66:249–258.

Walsh, P. T., M. Hansell, W. D. Borello, and S. D. Healy. 2011. Individuality in nest building: Do southern masked weaver (ploceus velatus) males vary in their nest-building behaviour? Behavioural Processes 88:1–6.

Weibull, J. W. 1997. Evolutionary game theory. MIT Press.

West-Eberhard, M. J. 1989. Phenotypic plasticity and the origins of diversity. Annual Review of Ecology and Systematics 20:249–278.

Wolfram Research, Inc. 2011. Mathematica, Version 8.0.4. Champaign, Illinois.

Young, H. P. 2004. Strategic learning and its limits. Oxford University Press, Oxford.

Appendix A

(Chapter 2)

A.1 Stochastic approximation

A.1.1 Expected motion of motivations

Here, we derive eq. 2.10 of the main text from eq. 2.7. To that end, we call $\mathbf{M}_{i,t} = (M_{i,t}(1), \ldots, M_{i,t}(m))$ the vector collecting the motivations of individual i at time t and $\mathbf{M}_t = (\mathbf{M}_{1,t}, \ldots, \mathbf{M}_{N,t})$ the vector of motivations in the whole population at time t. We also denote by $\mathbf{M}_{-i,t}$ the motivations of all individuals except individual i at time t. With this, the expectation of $R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t)$ (eq. 2.9) given current motivational state can be written as

$$\bar{R}_{i,t}(a, \mathbf{M}_t) = \mathbb{E}\left[R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t) \,|\, \mathbf{M}_t\right] = \sum_{h \in \mathscr{A}} \sum_{\mathbf{a}_{-i} \in \mathscr{A}^{N-1}} \sum_{\omega \in \Omega} R_i(a, h, \mathbf{a}_{-i}, \omega) \times p_{i,t}(h \,|\, \mathbf{M}_{i,t}) p_{-i,t}(\mathbf{a}_{-i} \,|\, \mathbf{M}_{-i,t}) \mu(\omega), \quad \text{(A.1)}$$

where $p_{i,t}(h \mid \mathbf{M}_{i,t})$ is the probability that individual i takes action h given its current motivations $\mathbf{M}_{i,t}$, $p_{-i,t}(\mathbf{a}_{-i} \mid \mathbf{M}_{-i,t})$ is the joint probability that the opponents of individual i play action profile $\mathbf{a}_{-i}$ when they have motivational state $\mathbf{M}_{-i,t}$, and $\mu(\omega)$ denotes the probability of state ω under the stationary distribution of environmental states (we will reason in terms of the long run behavior of the learning dynamics in the following).

For simplicity of presentation, we will use the notation of eq. 2.4, i.e., $p_{i,t}(k \mid \mathbf{M}_{i,t}) = p_{i,t}(k)$ and $p_{-i,t}(\mathbf{a}_{-i} \mid \mathbf{M}_{-i,t}) = p_{-i,t}(\mathbf{a}_{-i})$. Actions are taken independently by each individual in the population according to eq. 2.4, whereby

$$p_{-i,t}(\mathbf{a}_{-i}) = \prod_{i \neq j} p_{j,t}(a_j), \quad \text{(A.2)}$$

where a_j denotes the j-th element of the vector $\mathbf{a}_{-i}$.

With the above definitions, we can write eq. 2.7 as

$$M_{i,t+1}(a) - M_{i,t}(a) = \frac{1}{n_{t+1}} \left[-\epsilon_i M_{i,t}(a) + \bar{R}_{i,t}(a, \mathbf{M}_t) + U_{i,t+1}(a, \mathbf{a}_t, \omega_t) \right], \tag{A.3}$$

where the term $U_{i,t+1}(a, \mathbf{a}_t, \omega_t) = R_i(a, \mathbf{a}_t, \omega_t) - \bar{R}_{i,t}(a, \mathbf{M}_t)$ is called the "noise" term in stochastic approximation algorithm. The expression $U_{i,t+1}(a, \mathbf{a}_t, \omega_t)$ is subscribed by $t+1$ (and not t) in the stochastic approximation literature because it determines the value of the state variable at time $t+1$. It follows from the definition of the noise that $\{U_{i,t}(a_i, \mathbf{a}_t, \omega_t)\}_{t \geq 1}$ is a sequence of martingale differences adapted to the filtration generated by the random variables $\{\mathbf{M}_t\}_{t \geq 1}$. That is, $\mathbb{E}[U_{i,t+1}(a, \mathbf{a}_t, \omega_t) | \mathbf{M}_t] = 0$. Since the payoffs are bounded, we also have $\mathbb{E}[U_{i,t}(a, \mathbf{a}_t, \omega_t)^2] < \infty$. We further assume that the choice probability (eq. 2.4) is continuous in the motivations of the players, such that the expected reinforcement $\bar{R}_{i,t}(a, \mathbf{M}_t)$ is Lipschitz continuous in the motivations. With this, $-\epsilon_i M_{i,t}(a) + \bar{R}_{i,t}(a, \mathbf{M}_t)$ is a well-behaved vector field and standard results from stochastic approximation theory (Benaim, 1999; Benaim and El Karoui, 2005, p. 173) allow us to approximate the original stochastic process (eq. A.3) with the deterministic differential equation

$$\dot{M}_i(a) = -\epsilon_i M_i(a) + \bar{R}_i(a, \mathbf{M}). \tag{A.4}$$

The solutions of the original stochastic recursion (eq. 2.1) asymptotically track solutions of this differential equation. In particular, it has been established that the stochastic process almost surely converges to the internally chain recurrent set of the differential equation A.4 (Benaim, 1999, Prop. 4.1 and Th. 5.7). The simplest form of a chain recurrent set is the set of equilibrium points of the dynamics (the particular applications of our model that we study do not go beyond these cases). Note that in continuous time the equations are deterministic and we remove the subscript t to $\mathbf{M}_t$ for ease of presentation.

A.1.2 Differential equation in terms of mean payoff

Here, we show that it is possible to simplify the expression of the expected reinforcement $\bar{R}_{i,t}(a, \mathbf{M}_t)$ for our explicit learning model (eq. 2.1). First, recall from eq. 2.9 that for action a of player i, the realized reinforcement has the form

$$R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t) = \left[\delta_i + (1 - \delta_i) \mathbb{1}(a, a_{i,t}) \right] \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{A.5}$$

We see that

$$R_i(a,a_{i,t},\mathbf{a}_{-i,t},\omega_t)=\begin{cases}\pi_i(a,\mathbf{a}_{-i,t},\omega_t) & \text{if } a_{i,t}=a,\\ \delta_i\pi_i(a,\mathbf{a}_{-i,t},\omega_t) & \text{if } a_{i,t}\neq a.\end{cases} \tag{A.6}$$

In order to find an expression for the expected reinforcement $\bar{R}_{i,t}(a,\mathbf{M}_t)$, it is useful to rewrite eq. A.5 as

$$R_i(a,a_{i,t},\mathbf{a}_{-i,t},\omega_t)=\left[\delta_i+(1-\delta_i)\mathbb{1}(a,a_{i,t})\right]\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}\pi_i(a,\mathbf{a}_{-i},\omega_t)\mathbb{1}(\mathbf{a}_{-i},\mathbf{a}_{-i,t}), \tag{A.7}$$

since $\mathbb{1}(\mathbf{a}_{-i},\mathbf{a}_{-i,t})=1$ if $\mathbf{a}_{-i}=\mathbf{a}_{-i,t}$, 0 otherwise. Now, given that the event $\mathbf{a}=(a_1,\ldots,a_{i-1},a,a_{i+1},\ldots,a_N)$ occurs with probability $p_{i,t}(a)p_{-i,t}(\mathbf{a}_{-i})$ at time t, we deduce that the expected reinforcement of the motivation of action a is

$$\bar{R}_{i,t}(a,\mathbf{M}_t)=\sum_{\omega\in\Omega}\mu(\omega)\Big[p_{i,t}(a)\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}p_{-i,t}(\mathbf{a}_{-i})\pi_i(a,\mathbf{a}_{-i},\omega) + \delta_i(1-p_{i,t}(a))\{\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}p_{-i,t}(\mathbf{a}_{-i})\pi_i(a,\mathbf{a}_{-i},\omega)\}\Big]. \tag{A.8}$$

Factoring out, we have

$$\bar{R}_{i,t}(a,\mathbf{M}_t)=\sum_{\omega\in\Omega}\mu(\omega)[\{p_{i,t}(a)+\delta_i(1-p_{i,t}(a))\}\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}p_{-i,t}(\mathbf{a}_{-i})\pi_i(a,\mathbf{a}_{-i},\omega)]. \tag{A.9}$$

Define the average payoff

$$\bar{\pi}_i(a,\mathbf{a}_{-i})=\sum_{\omega\in\Omega}\mu(\omega)\pi_i(a,\mathbf{a}_{-i},\omega). \tag{A.10}$$

Taking expectation, then produces

$$\bar{R}_{i,t}(a,\mathbf{M}_t)=[p_{i,t}(a)+\delta_i(1-p_{i,t}(a))]\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}p_{-i,t}(\mathbf{a}_{-i})\bar{\pi}_i(a,\mathbf{a}_{-i}), \tag{A.11}$$

and substituting into eq. A.3 shows that we can write the differential equation for the motivations (eq. A.4) as

$$\dot{M}_i(a)=-\epsilon_i M_i(a)+[p_i(a)+\delta_i(1-p_i(a))]\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}p_{-i}(\mathbf{a}_{-i})\bar{\pi}_i(a,\mathbf{a}_{-i}). \tag{A.12}$$

A.1.3 Differential equation for the choice probabilities

Logit choice

Here, we derive the ODE for the choice probabilities (eq. 2.13) by combining the ODE for the motivations (eq. 2.10) with the choice rule (eq. 2.4), under the assumption that the choice rule is the logit choice function (eq. 2.6).

Differentiating the left and right member of eq. 2.4 with respect to time t, we have by the the chain rule

$$\dot{p}_i(a) = \sum_{k \in \mathscr{A}} \frac{\mathrm{d}p_i(a)}{\mathrm{d}M_i(k)} \dot{M}_i(k), \tag{A.13}$$

and substituting eq. 2.4 gives

$$\dot{p}_i(a) = \frac{\mathrm{d}f(M_i(a))}{\mathrm{d}M_i(a)} \frac{\dot{M}_i(a)}{\sum_{k \in \mathscr{A}} f(M_i(k))} - p_i(a) \sum_{k \in \mathscr{A}} \frac{\mathrm{d}f(M_i(k))}{\mathrm{d}M_i(k)} \frac{\dot{M}_i(k)}{\sum_{h \in \mathscr{A}} f(M_i(h))}. \tag{A.14}$$

Using $f(M) = \exp(\lambda_i M)$ in the choice function (eq. 2.4) gives eq. 2.6, which implies

$$\frac{\mathrm{d}f(M_i(a))}{\mathrm{d}M_i(a)} \times \frac{1}{\sum_{k \in \mathscr{A}} f(M_i(k))} = \lambda_i p_i(a), \tag{A.15}$$

whereby eq. A.14 can be written as

$$\dot{p}_i(a) = \lambda_i p_i(a) \left(\dot{M}_i(a) - \sum_{k \in \mathscr{A}} \dot{M}_i(k) p_i(k) \right). \tag{A.16}$$

Using the explicit expression for the differential equation of the motivations (eq. A.4), this is

$$\frac{1}{\lambda_i p_i(a)} \dot{p}_i(a) = \epsilon_i \left(\sum_{k \in \mathscr{A}} \{M_i(k) - M_i(a)\} p_i(k) \right) + \bar{R}_i(a) - \sum_{k \in \mathscr{A}} \bar{R}_i(k) p_i(k). \tag{A.17}$$

But, from the choice probabilities (eq. 2.6) we have the identity

$$\frac{p_i(k)}{p_i(a)} = \frac{\exp\left[\lambda_i M_i(k)\right]}{\exp\left[\lambda_i M_i(a)\right]}, \tag{A.18}$$

which gives

$$\log\left(\frac{p_i(k)}{p_i(a)}\right) = \lambda_i \left(M_i(k) - M_i(a)\right) \tag{A.19}$$

and on substitution into eq. A.17 produces

$$\dot{p}_i(a) = p_i(a) \left[\epsilon_i \sum_{k \in \mathscr{A}} \log\left(\frac{p_i(k)}{p_i(a)}\right) p_i(k) + \lambda_i \left(\bar{R}_i(a) - \sum_{k \in \mathscr{A}} \bar{R}_i(k) p_i(k) \right) \right]. \tag{A.20}$$

Power choice

Here, we perform the same derivation as in the last section but assume that $f(M) = M^{\lambda_i}$ in eq. 2.4. In this case, $\left[\mathrm{d}f(M_i(a))/\mathrm{d}M_i(a)\right] / \sum_{k \in \mathscr{A}} f(M_i(k)) = \left[\lambda_i M_i(a)^{\lambda_i - 1}\right] / \sum_{k \in \mathscr{A}} f(M_i(k)) = \lambda_i p_i(a)/M_i(a)$, whereby using $\dot{M}_i(a) = -\epsilon_i M_i(a) + \bar{R}_i(a)$ in eq. A.14 yields

$$\dot{p}_i(a) = \lambda_i p_i(a) \left[-\epsilon_i + \frac{\bar{R}_i(a)}{M_i(a)} - \sum_{k \in \mathscr{A}} p_i(k) \left(-\epsilon_i + \frac{\bar{R}_i(k)}{M_i(k)} \right) \right]. \tag{A.21}$$

Since $-\epsilon_i$ cancels from this equation and for non-negative motivations we have the equality $p_i(a)/p_i(k) = \left[M_i(a)/M_i(k)\right]^{\lambda_i}$, we can write

$$\dot{p}_i(a) = \left[\lambda_i p_i(a)/M_i(a)\right]\left[\bar{R}_i(a) - \sum_{k\in\mathcal{A}} p_i(k)\left(\frac{p_i(a)}{p_i(k)}\right)^{1/\lambda_i}\bar{R}_i(k)\right]. \tag{A.22}$$

A.2 Learning to play Hawk and Dove

Here, we analyze qualitatively the vector fields of eqs. 2.15–2.16 with an average Hawk-Dove game. We used Mathematica (Wolfram Research, Inc., 2011) to compute equilibria, eigenvalues and complicated algebraic expressions. We first study the interaction between two PRL and then the interaction between PRL and IL.

A.2.1 PRL vs. PRL

Pure Reinforcement Learning corresponds to $\delta_i = 0$. Thus, replacing $\delta_1 = \delta_2 = 0$ in eqs. 2.15–2.16 and using the payoffs of the Hawk-Dove game (Table 2.2) produces

$$\begin{aligned}
\dot{p}_1 &= p_1(1-p_1)\lambda_1\left[p_1 p_2\frac{B}{2} + (1-p_1)\{p_2 B + (1-p_2)\left(\frac{B}{2} - C\right)\}\right],\\
\dot{p}_2 &= p_2(1-p_2)\lambda_2\left[p_2 p_1\frac{B}{2} + (1-p_2)\{p_1 B + (1-p_1)\left(\frac{B}{2} - C\right)\}\right].
\end{aligned} \tag{A.23}$$

This dynamical system has eight different equilibria. In addition to the four at the corners of the state space $[(0,0),(1,1),(0,1),(1,0)]$, we have two interior equilibria and two symmetric (w.r.t. the line $p_1 = p_2$) equilibria on the edges $p_1 = 0$ and $p_2 = 0$ (Table A.1). Performing a linear stability analysis (Hirsch et al., 2004) near each equilibrium, we find that the vector field can be divided in three regions, each one being the basin of attraction of a locally stable equilibrium. The first one is the region where all trajectories tend to the equilibrium $(0,0)$. This equilibrium has negative eigenvalues. Its basin of attraction is delimited by the stable manifolds of the equilibria situated on the edges, precisely situated at $\left(0,\frac{1}{3}\right)$ and $\left(\frac{1}{3},0\right)$. The nullclines give the limits of a subset of this basin: the gray shaded area in Fig. 2.2A corresponds to all the points such that $\dot{p}_1 < 0$, $\dot{p}_2 < 0$, $p_{2,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$, $p_{2,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$. These are the points below the equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$ and where the vector field points south-west. Excluding this specific region, all points below the diagonal line $p_1 = p_2$ are in the basin of $(0,1)$ and all points above this line pertain to the basin of $(1,0)$. The points on this line

$p_1 = p_2$ (again excluding the points that are in the basin of $(0,0)$) are on the stable manifold of the interior equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$.

Table A.1: Local Stability analysis of the equilibria for the PRL vs. PRL interaction in the average Hawk-Dove game. (Expressions of the eigenvalues associated to the interior equilibria are too long to fit in the table.)

Equilibrium	Associated Eigenvalues	Eigenvalues' sign
$(0,0)$	$\left(-\frac{B}{2}, -\frac{B}{2}\right)$	$(-,-)$
$(0,1)$	$(-B,0)$	$(-,0)$
$(1,0)$	$(-B,0)$	$(-,0)$
$(1,1)$	$\left(-\frac{B}{2}+C, -\frac{B}{2}+C\right)$	$(+,+)$
$(0,\frac{1}{3})$	$\left(-\frac{B}{3}, \frac{B}{3}\right)$	$(-,+)$
$(\frac{1}{3},0)$	$\left(-\frac{B}{3}, \frac{B}{3}\right)$	$(-,+)$
$\left(\frac{B}{2B+\sqrt{2B(B-C)}}, \frac{B}{2B+\sqrt{2B(B-C)}}\right)$		$(+,+)$
$\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$		$(-,+)$

A.2.2 PRL vs. IL

Payoffs-Informed Learning (IL) corresponds to $\delta_i = 1$. Thus, replacing $\delta_1 = 0$ and $\delta_2 = 1$ in eqs. 2.15–2.16 and using the payoffs of the Hawk-Dove game (Table 2.2), one obtains the dynamical system describing learning between PRL (player 1) and IL (player 2) as

$$\begin{aligned}
\dot{p}_1 &= p_1(1-p_1)\lambda_1 \left[p_1 p_2 \frac{B}{2} - (1-p_1)\{p_2 B + (1-p_2)\left(\frac{B}{2} - C\right)\}\right], \\
\dot{p}_2 &= p_2(1-p_2)\lambda_2 \left[p_1 \frac{B}{2} - \{p_1 B + (1-p_1)\left(\frac{B}{2} - C\right)\}\right].
\end{aligned} \tag{A.24}$$

This determines six equilibria and three of them have at least one positive eigenvalue. We are left with $(0,1)$, $(1,0)$ and one interior at $(\frac{B}{2C}, \frac{3B-2C}{2B})$. The latter equilibrium has eigenvalues $(-B + \frac{3B^2}{8C} + \frac{C}{2}, -\frac{B(B-2C)}{4C})$ where the first one is always negative and the second one always positive. This equilibrium thus admits a stable manifold that splits the vector field in two regions: above the stable manifold, this is the basin of attraction of $(1,0)$ and below it trajectories go to $(0,1)$ (Fig. 2.2B).

A.3 Exploratory Reinforcement Learning

Here we analyze the equilibria of eq. 2.20 in the producer-scrounger model when λ_{E} is very large, in which case the second term in eq. 2.20 ($\bar{R}_{\mathrm{E}}(1) - \left[\bar{R}_{\mathrm{E}}(1)p_{\mathrm{E}} + \bar{R}_{\mathrm{E}}(2)(1-p_{\mathrm{E}})\right]$) dominates the first $[(1-p_{\mathrm{E}})\log([1-p_{\mathrm{E}}]/p_{\mathrm{E}})]$, which we neglect. We then find that there are three equilibria to this differential equation: $\hat{p}_{\mathrm{E}} = 0$, $\hat{p}_{\mathrm{E}} = 1$ and $\hat{p}_{\mathrm{E}} = V(2)/(V(1)+V(2))$. The interior equilibrium $[V(2)/(V(1)+V(2))]$ is unstable since

$$\frac{\mathrm{d}}{\mathrm{d}p_{\mathrm{E}}}\left\{p_{\mathrm{E}}\left(\bar{R}_{\mathrm{E}}(1) - \left[\bar{R}_{\mathrm{E}}(1)p_{\mathrm{E}} + \bar{R}_{\mathrm{E}}(2)(1-p_{\mathrm{E}})\right]\right)\right\}\Bigg|_{p_{\mathrm{E}}=V(2)/(V(1)+V(2))} > 0 \tag{A.25}$$

for $V(1) > V(2) \geq 0$. Thus, an ERL will learn to go on patch type 1 if its initial probability to go on it is greater than $V(2)/(V(1)+V(2))$, and it will learn to go on patch 2 otherwise. If one draws the initial condition at random from a uniform distribution on $[0,1]$, the expected equilibrium probability to go on patch type 1 for an ERL is

$$\hat{p}_{\mathrm{E}} = 0 \times \frac{V(2)}{V(1)+V(2)} + 1 \times \left(1 - \frac{V(2)}{V(1)+V(2)}\right) = \frac{V(1)}{V(1)+V(2)}. \tag{A.26}$$

More generally, eq. 2.20 is characterized by several stable equilibria and in order to define the expected equilibrium behavior of ERL we follow the same argument by having a distribution over the initial conditions. We call E the set of stable equilibria, $\hat{p}_{\mathrm{E}}^{e}$ the value of stable equilibrium e, and β_e the size of the basin of attraction of equilibrium e. Then, we define the expected probability to go on patch 1 of ERL as

$$\hat{p}_{\mathrm{E}} = \sum_{e \in E} \beta_e \hat{p}_{\mathrm{E}}^{e}, \tag{A.27}$$

where, by a slight abuse of notation, we still use $\hat{p}_{\mathrm{E}}$ to denote the average.

For instance, when we are in equilibrium regime I (Fig. 2.6), i.e., when there is one stable equilibrium, there is only one term in the sum of eq. A.27. When we are in equilibrium regime II and III, there are two terms in the sum.

A.4 Tit-for-Tat from EWA

Here, we derive the Tit-for-Tat strategy (Rapoport and Chammah, 1965; Axelrod, 1980; Axelrod and Hamilton, 1981) from EWA. This is not a learning rule, but it is interesting that it can be derived from the EWA framework by appealing to the concept of aspiration levels, which

are often used in learning models (Gale et al., 1995; Wakano and Yamamura, 2001; Macy and Flache, 2002; Cho and Matsui, 2005; Izquierdo et al., 2007; Chasparis et al., 2010). This provides a payoff-based (i.e., quantitative) version of TFT, which is easier to justify in terms of neuronal decision-making than the traditional version based on actions of opponent (which is more qualitative). To that aim, we need that the parameters are $\phi_i = 0$, $\rho_i = 0$, $\delta_i = 1$, $n_{i,1} = 1$, and $\lambda_i = \infty$ (Table 2.1), and we subtract aspiration levels to the original motivations, that is,

$$M_{i,t+1}(a) = \pi_i(a, \mathbf{a}_{-i,t}, \omega_t) - L_i(a), \tag{A.28}$$

where $L_i(a)$ is the aspiration level of individual i for action a.

In order to prove that eq. A.28 combined with eq. 2.6 are indeed Tit-for-Tat, consider an individual i who is engaged in the repeated play of the Prisoner's Dilemma with a fixed opponent playing $\mathbf{a}_{-i,t}$. The payoff matrix is

$$\begin{pmatrix} \mathscr{R} & \mathscr{S} \\ \mathscr{T} & \mathscr{P} \end{pmatrix},$$

with the traditional assumptions that $\mathscr{T} > \mathscr{R} > \mathscr{P} > \mathscr{S}$ and $(\mathscr{T} + \mathscr{S})/2 < \mathscr{R}$.

For eq. A.28 and eq. 2.6 with $\lambda_i = \infty$ to produce TFT behavior, one needs that

$$\begin{cases} M_{i,t+1}(\mathrm{C}) > M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ M_{i,t+1}(\mathrm{C}) < M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D}, \end{cases} \tag{A.29}$$

where $\mathbf{a}_{-i,t}$ denotes here the action of the single opponent of individual i. Substituting the definition of the motivations (eq. A.28) into eq. A.29, we have

$$\begin{cases} \pi_i(\mathrm{C},\mathrm{C}) - L_i(\mathrm{C}) > \pi_i(\mathrm{D},\mathrm{C}) - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ \pi_i(\mathrm{C},\mathrm{D}) - L_i(\mathrm{C}) < \pi_i(\mathrm{D},\mathrm{D}) - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D}. \end{cases} \tag{A.30}$$

where $L_i(\mathrm{C})$ is the aspiration level of individual i for cooperation, $L_i(\mathrm{D})$ its aspiration level for defection, and where we removed the dependence of the payoffs on the environmental state ω_t, because we consider a fixed game. Substituting the payoff from the payoff matrix, eq. A.28 produces TFT behavior if

$$\begin{cases} \mathscr{R} - L_i(\mathrm{C}) > \mathscr{T} - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ \mathscr{S} - L_i(\mathrm{C}) < \mathscr{P} - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D}, \end{cases} \tag{A.31}$$

which can be expressed as the single condition

$$\mathscr{T} - \mathscr{R} < L_i(\mathrm{D}) - L_i(\mathrm{C}) < \mathscr{P} - \mathscr{S}. \tag{A.32}$$

We remark that this payoff-based version of TFT needs that individual i has a bigger aspiration level for defection, i.e., individual i expects more of defection than of cooperation (because $\mathscr{T} - \mathscr{R} > 0$). Also, clearly not all Prisoner's Dilemma games satisfy condition A.32. More precisely, this condition entails that defection needs to risk dominate cooperation ($\mathscr{T} - \mathscr{P} < \mathscr{R} - \mathscr{S}$) for our version of TFT to be implementable.

Appendix B

(Chapter 3)

The reader might notice similarities between this Appendix and the Appendix for Chapter 2 above. The reason is that Chapter 2 and Chapter 3 are conceived as two independent articles submitted separately to two different journals, so that the papers could be read independently and their appendix must be self-contained.

B.1 Reinforcement and inference-based learning

In this appendix, we give the stochastic recursions for the motivations for the four learning rules studied in this paper. We have defined two forms of reinforcement learning: PRL and ERL, and two forms of inference-based learning: PIL and EIL.

B.1.1 Reinforcement learning

Dynamic learning rate

Substituting $\delta_i = 0$, $\phi_{i,t} = (1/t) + 1$ into eq. 3.3, we obtain PRL with updating rule

$$M_{i,t+1}(a) = M_{i,t}(a) + \frac{1}{t+1}\mathbb{1}(a, a_{i,t})\pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{B.1}$$

Every new experienced payoff is thus divided by the total number of previous interactions $(t+1)$ and added to the previous motivation. In the long run, the effect of new payoffs on motivations goes to zero. Note that when action a is not played, the motivation is not updated. Moreover, the learner does not forget information from the past. It is even the payoffs obtained in the first rounds of interaction that have the biggest effect on the motivations at time t.

Constant learning rate

Substituting $\delta_i = 0$ and $\phi_{i,t} = 1$ into eq. 3.3 gives ERL, which has updating rule

$$M_{i,t+1}(a) = \frac{t}{t+1} M_{i,t}(a) + \frac{1}{t+1} \mathbb{1}(a, a_{i,t}) \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{B.2}$$

This rule looks like a time average of the payoffs obtained when playing action a but it is actually a biased average. Indeed, in a non-biased average, the motivation of action a should not be updated when action a is not played at time t. However, here when action a is not played at time t, the motivation is still updated but it is as if the payoff obtained for action a at time t was zero. Hence, depending on the signs of the payoffs in the game, the non-played actions have a tendency to lose weight (e.g., when all payoffs in the game are positive) or gain weight (e.g., when all payoffs in the game are negative).

B.1.2 Inference-based learning

Dynamic learning rate

Substituting $\delta_i = 1$ and $\phi_{i,t} = (1/t) + 1$ into eq. 3.3 yields the PIL updating rule

$$M_{i,t+1}(a) = M_{i,t}(a) + \frac{1}{t+1} \pi_i(a, \mathbf{a}_{-i,t}, \omega_t), \tag{B.3}$$

where imagined payoffs have no effect on the motivations for large t.

Constant learning rate

Substituting $\delta_i = 1$ and $\phi_{i,t} = 1$ into eq. 3.3 gives EIL, which is the standard Belief-based learning rule (Camerer and Ho, 1999), with motivation updating given by

$$M_{i,t+1}(a) = \frac{t}{t+1} M_{i,t}(a) + \frac{1}{t+1} \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{B.4}$$

Contrary to the ERL rule, this equation represents a proper time average: this is the average payoff that would have been obtained by individual i if he was constantly playing action a up to time t, given the history of actions by his opponents $\{\mathbf{a}_{-i,\tau}\}_{\tau=1}^{t}$.

Belief-based learning

We now show that eq. B.4 can also be interpreted in terms of updating average payoffs given beliefs over the action play probabilities of partners like in Camerer and Ho (1999). For ease

of presentation, but without loss of generality, we consider that individual i interacts only with one other individual in the population, which plays action $a_{-i,t} \in \mathscr{A}$ at time t. We then write for $t \geq 2$

$$B_{-i,t+1}(a) = \frac{t}{t+1} B_{-i,t}(a) + \frac{1}{t+1} \mathbb{1}(a, a_{i,t}), \tag{B.5}$$

where $B_{-i,t+1}(a)$ is the frequency of times the partner of individual i has played action a up to time t, and which is the belief of individual i that its partner plays a at $t+1$ given the initial belief $B_{-i,1}$.

Let us now write

$$M_{i,t+1}(a) = \sum_{k \in \mathscr{A}} \sum_{\omega \in \Omega} \pi_i(a, k, \omega) \mu(\omega) B_{-i,t+1}(a), \tag{B.6}$$

which can be interpreted as the expected payoff to individual i given its beliefs over the action distribution of its partner and the current state of the environment. Substituting eq. B.5 into eq. B.6 shows that the motivation dynamics still satisfies eq. B.4. Hence, when individual is an inference-based learner and expresses action by using the logit choice rule (eq. 3.2), it behaves as if it tries to maximize its expected current reward given its beliefs.

B.2 Stochastic approximation

Here, we show the main steps to derive eqs. 3.5–3.6 from eqs. 3.2-3.3. First, an application of stochastic approximation theory (e.g., Benaim, 1999) shows that eqs. 3.2-3.3 can be approximated by the set of differential equations

$$\dot{p}_i(a) = p_i(a) \left[\epsilon_i \sum_{k \in \mathscr{A}} \log \left(\frac{p_i(k)}{p_i(a)} \right) p_i(k) + \lambda \left(\bar{R}_i(a) - \sum_{k \in \mathscr{A}} \bar{R}_i(k) p_i(k) \right) \right], \tag{B.7}$$

where

$$\bar{R}_i(a) = [p_i(a) + \delta_i(1 - p_i(a))] \sum_{\mathbf{a}_{-i} \in \mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i}) \bar{\pi}_i(a, \mathbf{a}_{-i}) \tag{B.8}$$

and

$$\bar{\pi}_i(a, \mathbf{a}_{-i}) = \sum_{\omega \in \Omega} \mu(\omega) \pi_i(a, \mathbf{a}_{-i}, \omega) \tag{B.9}$$

(Dridi and Lehmann, 2013, eqs. 11-13). Here, $\bar{R}_{i,t}(a, \mathbf{M}_t)$ is the expectation of the reinforcement to the motivation of action a, i.e., the expectation of the numerator of the second term of eq. 3.3 over the distribution of environmental states and the distribution of choice probabilities,

$p_{-i}(\mathbf{a}_{-i})$ is the probability of the joint action profile of other individuals than i, and $\bar{\pi}_i(a,\mathbf{a}_{-i})$ represents the payoff of the average game in which individual i is involved.

In our context, the parameter ϵ_i in eq. B.7 takes the value $\epsilon_i = 1 + t(1-\phi_{i,t})$ (Dridi and Lehmann, 2013, eq. 8 with $n_{i,t} = t$ and $\rho_i = 1$). Since we assumed that $\phi_{i,t} = (1/t)+1$, this gives $\epsilon_i = 0$ and the exploration term (the first term in square brackets in eq. B.7) cancels. Moreover, we are interested in 2×2 games so there are only two actions ($\mathscr{A} = \{1,2\}$) and two players. Let the two players be denoted i and j, p_i be the probability that individual i take action 1, and p_j be the probability that individual j take action 1. With this, we can write the differential equation for the probability that individual i takes action 1 using eq. B.7 as

$$\dot{p}_i = p_i\lambda\left(\bar{R}_i(1) - [\bar{R}_i(1)p_i + \bar{R}_i(2)(1-p_i)]\right), \tag{B.10}$$

where

$$\bar{R}_i(1) = [p_i + \delta_i(1-p_i)] \sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i})\bar{\pi}_i(1,\mathbf{a}_{-i}) \tag{B.11}$$

$$\bar{R}_i(2) = [(1-p_i) + \delta_i p_i] \sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i})\bar{\pi}_i(2,\mathbf{a}_{-i}). \tag{B.12}$$

Further, in the 2×2 games that we study here (one-shot matching model), the single opponent of individual i is individual j so $\mathbf{a}_{-i}\in\{1,2\}$ and $p_{-i}(1) = p_j$. Replacing these in eq. B.11, we can write eq. B.10 as

$$\begin{aligned}\dot{p}_i = p_i(1-p_i)\lambda[&\{p_j\bar{\pi}_i(1,1) + (1-p_j)\bar{\pi}_i(1,2)\}\{p_i + \delta_i(1-p_i)\} \\ &- \{p_j\bar{\pi}_i(2,1) + (1-p_j)\bar{\pi}_i(2,2)\}\{\delta_i p_i + (1-p_i)\}].\end{aligned} \tag{B.13}$$

Using the definition of the payoffs of the average game in Table 3.2, we have $\bar{\pi}_i(1,1) = \mathscr{R}$, $\bar{\pi}_i(1,2) = \mathscr{S}$, $\bar{\pi}_i(2,1) = \mathscr{T}$, $\bar{\pi}_i(2,2) = \mathscr{P}$, and on substitution into eq. B.13 yields

$$\begin{aligned}\dot{p}_i = p_i(1-p_i)\lambda[&\{p_j\mathscr{R} + (1-p_j)\mathscr{S}\}\{p_i + \delta_i(1-p_i)\} \\ &- \{p_j\mathscr{T} + (1-p_j)\mathscr{P}\}\{\delta_i p_i + (1-p_i)\}].\end{aligned} \tag{B.14}$$

For player j, the differential equation for its probability to take action 1 is the exact symmetric (because i is the single opponent of j), whereby

$$\begin{aligned}\dot{p}_j = p_j(1-p_j)\lambda[&\{p_i\mathscr{R} + (1-p_i)\mathscr{S}\}\{p_j + \delta_j(1-p_j)\} \\ &- \{p_i\mathscr{T} + (1-p_i)\mathscr{P}\}\{\delta_j p_j + (1-p_j)\}].\end{aligned} \tag{B.15}$$

B.3 Fecundity at behavioral equilibrium

In this section, we derive an expression for fecundity (eq. 3.7) under the assumption that the learning process (eqs. 3.5–3.6) has reached an equilibrium during lifespan. Indeed, if the individuals interact for a long enough time, the action probabilities $p_{i,t}(a)$ may reach an equilibrium for all i and a, and the fecundity of player i will be its average payoff at equilibrium. Then, the fecundity of individual i is

$$b_i = \sum_{\omega\in\Omega}\sum_{a\in\mathscr{A}}\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} \mu(\omega)\hat{p}_i(a)\hat{p}_{-i}(\mathbf{a}_{-i})\pi_i(a,\mathbf{a}_{-i},\omega), \tag{B.16}$$

where $\hat{p}_i(a)$ denotes the equilibrium probability with which individual i chooses action a, while $\hat{p}_{-i}(\mathbf{a}_{-i})$ is the equilibrium probability with which the opponents of individual i choose action profile $\mathbf{a}_{-i}$. This equilibrium is obtained by setting $\dot{p}_j(a) = 0$ in eq. B.7 for all j and a.

Equation B.16 should be understood as a long run average payoff in the game taken over three distributions. The first distribution gives the probability $\mu(\omega)$ to play game ω; the second distribution gives the equilibrium probability $\hat{p}_i(a)$ that player i takes action a; the third distribution tells the probability $\hat{p}_{-i}(\mathbf{a}_{-i})$ that the opponents of individual i take action profile $\mathbf{a}_{-i}$. The distribution $\mu(\omega)$ is already provided as a parameter of the model. The two other distributions have to be computed by studying the equilibria of the choice probabilities $\hat{p}_i(a)$ for all i in the population. Using eq. B.9, the average payoff can be simplified to

$$b_i = \sum_{a\in\mathscr{A}}\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} \hat{p}_i(a)\hat{p}_{-i}(\mathbf{a}_{-i})\bar{\pi}_i(a,\mathbf{a}_{-i}). \tag{B.17}$$

Since we are concerned with 2×2 games, and since the single opponent of individual i is individual j (eqs. 3.5–3.6), we have $a\in\{1,2\}$ and $\mathbf{a}_{-i}\in\{1,2\}$. If we further call $\hat{p}_i$ the probability that individual i plays action 1 at a behavioral equilibrium and $\hat{p}_j$ the corresponding probability for individual j, eq. B.17 can be developed as

$$b_i = \hat{p}_i\hat{p}_j\bar{\pi}_i(1,1) + \hat{p}_i(1-\hat{p}_j)\bar{\pi}_i(1,2) + (1-\hat{p}_i)\hat{p}_j\bar{\pi}_i(2,1) + (1-\hat{p}_i)(1-\hat{p}_j)\bar{\pi}_i(2,2). \tag{B.18}$$

Factoring out and replacing the the average payoffs, $\bar{\pi}_i(\cdot,\cdot)$, by their values in Table 3.2, we finally obtain

$$b_i = \hat{p}_i(\hat{p}_j\mathscr{R} + (1-\hat{p}_j)\mathscr{S}) + (1-\hat{p}_i)(\hat{p}_j\mathscr{T} + (1-\hat{p}_j)\mathscr{P}), \tag{B.19}$$

where in the main text we used $b_{ij} = b_i$ in order to emphasize that in the one-shot matching case, the payoff of individual i depends only on its single opponent (j).

B.4 Qualitative analysis for the one-shot matching model

Here, we carry out the stability analysis of the equilibrium points of the learning dynamics presented in the main text (eqs. 3.5–3.6). Before starting the analysis, let us make a technical remark. The dynamical systems we will analyze can display hyperbolic equilibria that admits stable manifolds (with one positive and one negative eigenvalue). We will completely discard the possibility that an initial condition is on a stable manifold because doing so leads to a locally unstable equilibrium, which is not robust to small perturbations under the original stochastic process (Pemantle, 1990). Equilibria and eigenvalues were calculated using the Mathematica software. The payoffs of the average games are defined in Table 3.2.

B.4.1 Prisoner's Dilemma

• *PRL vs. PRL*. The dynamical system obtained by setting $g_i = 0$, $g_j = 0$, $R = B - C$, $S = -C$, $T = B$, and $P = 0$ into eqs. 3.5–3.6 admits seven equilibria (Table B.1). Four are at the corners of the state space, two are on the edges and one is completely interior. The first egde equilibrium is situated on the line $p_i = 1$ and the other one is symmetric with respect to the line $p_j = p_i$. The interior equilibrium is $\left(\hat{p}_i = \frac{B+C}{2B}, \hat{p}_j = \frac{B+C}{2B}\right)$.

Evaluating the Jacobian matrix at each equilibrium and computing the eigenvalues (Table B.1) reveals that all equilibria are characterized by at least one positive eigenvalue, except the two equilibria $(0,0)$ (both players defect) and $(1,1)$ (both players cooperate). These two latter equilibria are thus the only possible endpoints of the learning dynamics (Hirsch et al., 2004). Fig. 3.1A shows some solution orbits for this dynamical system. There, we can see that the interior equilibrium $\left(\frac{B+C}{2B}, \frac{B+C}{2B}\right)$ is an unstable node, i.e., its eigenvalues are both positive. Moreover, the two equilibria on the edges admit a stable and unstable manifold because they have one positive eigenvalue and one negative eigenvalue. All this implies that the equilibria $(0,0)$ and $(1,1)$ have a basin of attraction that is delimited by these stable manifolds. Since solutions along the nullcline defined by $\dot{p}_i = 0$ verify $\dot{p}_j > 0$ and solutions along the other nullcline ($\dot{p}_j = 0$) verify $\dot{p}_i > 0$, solving the inequalities $\dot{p}_i > 0, \dot{p}_j > 0$ for p_i and p_j (the region above the nullclines) gives a subset of the basin of attraction. In other words, trajectories are increasing in this region and cannot escape it. Using Mathematica, we find that these inequalities are satisfied in two cases:

$$\begin{cases} \left(\frac{B}{2B-C} < p_{j,1} \leq \frac{B+C}{2B} \quad \text{and} \quad \frac{C p_{j,1}}{-B+2B p_{j,1}} < p_{i,1} < 1\right) & \text{or} \\ \left(\frac{B+C}{2B} < p_{j,1} < 1 \quad \text{and} \quad \frac{B p_{j,1}}{-C+2B p_{j,1}} < p_{i,1} < 1\right). \end{cases} \tag{B.20}$$

Table B.1: Local Stability analysis of the equilibria in the PRL vs. PRL case when the average game is the Prisoner's Dilemma.

Equilibrium	Associated Eigenvalues	Eigenvalues' sign
$(0,0)$	$(0,0)$	$(0,0)$
$(0,1)$	$(-B,C)$	$(-,+)$
$(1,0)$	$(-B,C)$	$(-,+)$
$(1,1)$	$(C-B,C-B)$	$(-,-)$
$(1,\frac{B}{2B-C})$	$\left(C+\frac{B^2}{C-2B},B+\frac{B^2}{C-2B}\right)$	$(-,+)$
$(\frac{B}{2B-C},1)$	$\left(C+\frac{B^2}{C-2B},B+\frac{B^2}{C-2B}\right)$	$(-,+)$
$\left(\frac{B+C}{2B},\frac{B+C}{2B}\right)$	$\left(\frac{(B-C)^2(B+C)}{4B^2},\frac{(B-C)(B+C)^2}{4B^2}\right)$	$(+,+)$

• *PIL vs. PIL*. The dynamical system obtained by setting $g_i = 1$, $g_j = 1$, and the PD game payoffs into eqs. 3.5–3.6 admits the four corners of the state space $[(0,1),(1,1),(1,0),(1,1)]$ as equilibria. The only locally stable equilibrium is the point $(0,0)$ because it has negative eigenvalues $(-C,-C)$. Hence, two players using belief learning will end up always defecting (Fig. 3.1B).

• *PRL vs. PIL*. Setting $g_i = 0$, $g_j = 1$, and the PD game payoff into eqs. 3.5–3.6, we find five equilibria. The four corner equilibria and one on the edge $p_{\mathbf{I}} = 1$ situated at $\left(\hat{p}_{\mathbf{R}} = \frac{B}{2B-C}, \hat{p}_{\mathbf{I}} = 1\right)$. The linearization shows that all equilibria are characterized by at least one positive eigenvalue, except the equilibrium $(0,0)$ which has eigenvalues $(-C,0)$, implying, by elimination, that it is the only stable equilibrium. Both players will tend to defect in the long run (Fig. 3.1C).

B.4.2 Hawk-Dove Game

• *PRL vs. PRL*. In this case ($g_i = 0$, $g_j = 0$, $R = B/2-C$, $S = B$, $T = 0$, and $P = B/2$), eqs. 3.5–3.6 has eight different equilibria. In addition to the four at the corners, we have two interior equilibria and two symmetric (w.r.t. the line $p_i = p_j$) equilibria on the edges $p_i = 0$ and $p_j = 0$ (Table A.1). The vector field can be divided in three regions, each one being the basin of attraction of an asymptotically stable equilibrium. The first one is the region where all trajectories tend to the equilibrium $(0,0)$. This equilibrium has negative eigenvalues. Its basin of attraction is delimited by the stable manifolds of the equilibria situated on the edges, precisely situated at $\left(0,\frac{1}{3}\right)$ and $\left(\frac{1}{3},0\right)$. The nullclines give a good approximation of the limits of this basin: the gray shaded area in Fig. 3.1D corresponds to all the points such that $\dot{p}_i < 0$, $\dot{p}_j < 0$, $p_{j,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$,

$p_{j,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$. These are the points below the equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$ and where the vector field points south-west. Excluding this specific region, all points below the diagonal line $p_j = p_i$ are in the basin of $(0,1)$ and all points above this line pertain to the basin of $(1,0)$. The points on this line $p_i = p_j$ (again excluding the points that are in the basin of $(0,0)$) are on the stable manifold of the interior equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$.

• *PIL vs. PIL*. Here, eqs. 3.5–3.6 admits only two equilibria $(0,1)$ and $(1,0)$, which are asymptotically stable in the region below the diagonal line $p_i = p_j$ and above this line, respectively. This represents the stable manifold of the equilibrium $(\frac{B}{2C}, \frac{B}{2C})$. On $p_i = p_j$, we have the single population replicator dynamics (Fig. 3.1E), hence the stable point on this line is the ESS of the Hawk-Dove game, $(\frac{B}{2C}, \frac{B}{2C})$.

• *PRL vs. PIL*. We have six equilibria and three of them have at least one positive eigenvalue. We are left with $(0,1)$, $(1,0)$ and one interior at $(\frac{B}{2C}, \frac{3B-2C}{2B})$. The latter equilibrium has eigenvalues $(-B + \frac{3B^2}{8C} + \frac{C}{2}, -\frac{B(B-2C)}{4C})$, where the first one is always negative and the second one always positive. This equilibrium thus admits a stable manifold that splits the vector field in two regions: above the stable manifold, this is the basin of attraction of $(1,0)$ and below it trajectories go to $(0,1)$ (Fig. 3.1F). This stable manifold is a curve passing through the points $(0, \frac{1}{3})$, $(\frac{B}{2C}, \frac{3B-2C}{2B})$, and $(1,1)$.

B.4.3 Coordination Game

This game provides the simplest dynamics, where the equilibria $(0,0)$ and $(1,1)$ are always the only two asymptotically stable states.

• *PRL vs. PRL*. Here, we set $g_i = 0$, $g_j = 0$, $R = B$, $S = 0$, $T = 0$, and $P = B$ in eqs. 3.5–3.6, which then admits four trivial corner equilibria plus all the points on the line $p_j = 1 - p_i$. The equilibria $(0,0)$ and $(1,1)$ both have negative eigenvalues and are thus locally stable. The two other corners equilibria ($(0,1)$ and $(1,0)$) have eigenvalues $(0,0)$. The equilibria on the line $p_j = 1 - p_j$ have eigenvalues $(0, 2B\,p_i(1-p_i))$, where the second eigenvalue is 0 when $p_i = 0$ or $p_i = 1$ and positive otherwise. This all implies that the equilibrium $(0,0)$ is asymptotically stable in the region below the line $p_j = 1 - p_i$ and the equilibrium $(1,1)$ is asymptotically stable above this line (Fig. 3.1G).

As mentioned in the main text, the unstable line $p_j = 1 - p_i$ is not an interesting set of initial conditions.

• *PIL vs. PIL.* The system eqs. 3.5–3.6 admits five equilibria in this situation: the four corners and one interior at $\left(\frac{1}{2},\frac{1}{2}\right)$. The equilibria $(0,0)$ and $(1,1)$ are both asymptotically stable because they both have eigenvalues $(-B,-B)$ while the equilibria $(0,1)$ and $(1,0)$ have eigenvalues (B,B). The interior equilbrium $\left(\frac{1}{2},\frac{1}{2}\right)$ is a saddle with eigenvalues $\left(-\frac{B}{2},\frac{B}{2}\right)$ and consequently admits a stable and an unstable manifold. It is easy to see in Fig. 3.1H that the stable manifold is the diagonal line $p_j = 1 - p_i$ while the unstable manifold is the other diagonal $p_j = p_i$. Every trajectory starting above $p_j = 1 - p_i$ will tend to $(1,1)$ while if it starts below this line it will tend to $(0,0)$.

• *PRL vs. PIL.* Here, eqs. 3.5–3.6 has again five equilibria: the four corners plus one interior at $\left(\frac{1}{2},\frac{1}{2}\right)$. The points $(0,0)$ and $(1,1)$ are both asymptotically stable having eigenvalues $(-B,-B)$. The interior equilibrium is a saddle with eigenvalues $\left(-\frac{B}{4},\frac{B}{2}\right)$. Hence a stable manifold passing through this saddle splits the vector field in two regions, which correspond respectively to the basin of attraction of $(0,0)$ and $(1,1)$. Here the stable manifold is situated no more on the diagonal because we lost the symmetry property of the PIL vs. PIL case (compare Fig. 3.1I with Fig. 3.1H).

B.5 Simulations

B.5.1 Individual based simulations

Here, we present the algorithm of our individual based simulations. Each individual $i \in \{1,2,...,N\}$ take a genotypic value $\delta_i \in \{0,1\}$. In each generation, each individual i obtains fecundity $b_i = \sum_{t=1}^{T} \pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \omega_t)$ (eq. 3.1, but we do not use the normalization factor $\frac{1}{T}$ in the simulations), where the actions $(a_{i,t}, \mathbf{a}_{-i,t})$, which are random variables, are calculated by implementing eqs. 3.2–3.3. The environmental state in each period (ω_t) is drawn from a uniform distribution, which entails that $\mu(\text{PD}) = \mu(\text{HD}) = \mu(\text{CG}) = 1/3$. The average game is parametrized according to the B and C parameters as in Table 3.2 (we always used $B = 5$ and $C = 3$), but the payoffs of the three sub-games (PD, HD, and CG) are randomly generated at the beginning of each generation so that they average up to the desired average game and satisfy the inequalities described in Table 3.2. This implies that there are between-generation fluctuations, and we used them to represent the conditions where a learning ability gives an advantage over innate behavior. Under one-shot matching, individuals were paired only at the beginning of the generation ($t = 1$) and each pair played together until T, while under repeated matching

individuals were rematched at each time period $t = 1, 2, \ldots, T$.

The next generation is sampled with replacement according to the relative fecundity of individuals (i.e., $b_i / \sum_{i=1}^{N} b_i$, a Wright-Fisher process). An offspring inherits the genotype of its parent with probability $1 - \eta$ or mutates with probability η to the other genotype. The mutation rate was set to $\eta = 10^{-3}$. We ran simulations with $N = 1000$ and we use a value of T bigger than the average time needed for learning to converge to a stable value (where this average time was computed separately for each game and initial condition, and we did not use values smaller than $T = 500$). Each case described in the main text (i.e., each game and each type of initial preferences of the learners) was run in three different replicates: one with an initial population (at the first generation) composed of half PRLs and half PILs; one with an initial population of only PRLs; one with an initial population of only PILs. This was to check that our simulation results are independent of initial conditions. Moreover, we waited for each run that the dynamic mean frequency of types converge to a stable value, at which point we stopped the simulation. Our convergence criterion was met when the time average of the types' frequencies did not change by more than 10^{-6} for 100 successive generations. We used $k = 0$ and varied λ from 10^{-1} to 10^{3} (see below).

We also carried out simulations only of the learning phase. To that aim, we implemented, in the one-shot matching case, pairs of individuals playing outside a population setting using the same parameters as in the full evolutionary simulations. We shall remark at this point that, contrary to the evolutionary simulations (which implement a Markov Chain admitting a stationary distribution), the learning dynamics does not admit a stationary distribution; stochastic approximation theory shows that the dynamics of learning will converge to one of the equilibrium points under the deterministic differential equation (see for example Borkar, 2008, Chap. 2, Corollary 4). In particular, the stochastic process will not necessarily converge to a linearly stable equilibrium point (this is the criterion of stability we used in the above analysis). To understand the simulations, just note that the learning dynamics should converge to one of the equilibria but that it need not converge to the same equilibrium for two different replicates of a simulation, hence the necessity to run many replicates of the same parameter set.

For repeated matching, learning simulations consisted of running only one generation but setting the frequencies of PRL and PIL manually. We took the same cases (i.e., the three different average games (PD, HD, and CG) and the different initial preferences over actions of the learners) as in the one-shot matching model and simulated learning behavior for 11 different values

of the frequency q of PRL in the population ranging from 0 to 1 by steps of 0.1.

B.5.2 Tuning parameters

In order to find parameters that reproduce our analytical results, the process of running simulations consisted of two steps. First, we ran simulations between pairs of learners (PRL vs. PRL, PRL vs. PIL, and PIL vs. PIL) for each game, in order to establish the accuracy of the approximation of the equilibrium action play probabilities. It has been previously found that the time t needed for the simulated learning process to converge to the predicted equilibrium critically depends on the sensitivity to motivations, λ, and on the initial difference between motivations of action 1 and 2, $\Delta M_{i,1} \equiv M_{i,1}(1) - M_{i,1}(2)$ at time $t = 1$ (Dridi and Lehmann, 2013). Since our analytic prediction is asymptotic, it tells nothing about the values of λ, $\Delta M_{i,1}$, and T. Hence, we first ran several simulations of learning with different values of λ and $\Delta M_{i,1}$, and waited for the learning process to "converge" (based on a numeric convergence criterion). This gives us the parameter T needed for convergence to happen during lifespan. We then compared the equilibrium behavior in these simulations with the predicted equilibrium, and chose the values of $\Delta M_{i,1}$ and λ that give the best match to prediction. In order to reduce our search in parameter space, we fixed the value of $p_{i,1}(a)$ to 0.85 for the initially preferred action a (see below) and only vary λ. This automatically changes the initial value of motivations $\Delta M_{i,1}$, such that we do not require to vary them explicitly. We used five different values of λ of the form 10^{α} for $\alpha = -1, 0, 1, 2, 3$. In the second step, we simulated the evolutionary process of selection on PRL vs. PIL. To this end, we used the values of λ, ΔM, and T found in step 1 which give the best match between stochastic learning process and deterministic approximation. The idea is that, since we chose parameters of learning where the approximation works well, the evolutionary simulations should also match the evolutionary predictions based on our approximation of learning. In order to further investigate the model under the alternative condition that ϕ_i is constant, we also performed the same set of simulations but with $\phi_i = 1$. This changes the type of the learning rules, which are now the counterpart to PRL and PIL when $\phi_i = 1$; that is, Exploratory Reinforcement Learning (ERL) and Belief-based learning BL). We used the parameter values for λ and $\Delta M_{i,1}$ applied in the case with dynamic learning rate.

B.6 Detailed simulation results

In this appendix, we describe in greater detail the results of our individual-based simulations

B.6.1 One-shot matching

Prisoner's Dilemma: dynamic learning rate

Both players initially prefer Cooperation. Here, the simulations give results close to what is expected under our approximation when $\lambda = 10$. Regarding learning, we find the following results for the three possible interactions between types. When PRL plays against PRL, most pairs learn to cooperate, but some pairs learn to defect (Fig. 3.2A): this is not surprising because under the stochastic learning process, individuals can escape basins of attractions and reach the locally stable equilibrium where both players defect. In the interaction between PILs, all pairs learn to defect (Fig. 3.2B), as predicted by the analysis. In the interaction between PRL and PIL, both types learn to defect: PRL does not get exploited by PIL (Fig. 3.2C). Overall this gives an advantage to PRL because this type cooperates with itself but defects with the defector PIL. As a consequence, in the evolutionary simulations, PRL fixes in the population irrespective of the initial composition of the population (Table 3.3).

Both players initially prefer Defection. In this case, we also find that $\lambda = 10$ gives the best fit to deterministic analysis. Surprisingly at first sight, we also observe that PRL individuals sometimes learn to cooperate when paired with themselves (Fig. 3.2D; while the analysis predicts that they will always defect in this case). This is actually perfectly possible, and is explained (as above) by the fact that initial conditions do not constrain absolutely the equilibrium behavior: individuals with initial preference for defection can still learn to cooperate because this is also a stable equilibrium for the dynamics. PIL individuals do not deviate from perfect defection (Fig. 3.2E). For the interaction between PRL and PIL, we find again that both types learn to defect (Fig. 3.2F). Since learning behavior is somehow similar to the case where individuals prefer cooperation, we find as expected that PRL fixes in the population when we simulate natural selection. This is due to the tendency of PRL to sometimes cooperate with itself (Table 3.3).

Prisoner's Dilemma: constant learning rate

Both players initially prefer Cooperation. In this situation, the results are very similar to the case with a dynamic learning rate. Namely, ERL is able to learn to cooperate against itself, a behavior that BL cannot express, and this gives a fitness advantage to ERL because ERL learns to defect against BL (Fig. B.1A,B,C). As a consequence, the frequency of ERL at an evolutionary equilibrium is very close to 1 (Table 3.3).

Both players initially prefer Defection. Here ERL individuals display the same behavior as with a dynamic learning rate when paired together: namely some pairs learn to defect and some other pairs learn to cooperate (Fig. B.1D). BL individuals on the other hand, still learn to defect whatever their opponent is (Fig. B.1E). The interactions between ERL and BL display a different outcome than previously. Here we observe that ERL individuals converge to a state where they have a positive probability to cooperate, and hence get exploited on some interaction rounds (Fig. B.1F).

As a consequence of this learning behavior, the evolutionary simulations show that BL fixes in the population in the long run when they are initially in high frequency but otherwise this is ERL that invades (Table 3.3). Such a result is possible if there is an interior unstable equilibrium in the evolutionary dynamics: when ERL are common in the population, they have a tendency to increase in frequency; when they are in low frequency they have a tendency to further decrease in frequency. This situation corresponds to observed learning behavior: while ERL individuals have the advantage of cooperating with themselves, this does not seem to compensate the fitness loss due to sometimes cooperating against the defector BL when BL constitutes most of the population.

Hawk Dove game: dynamic learning rate

PRL initially prefers to play Hawk and PIL prefers Dove. In the analysis, we predict that with these initial preferences, PRL individuals will learn to play half of the time Hawk and half of the time Dove when paired against themselves. However, in the simulations, we observe that a high proportion of PRL learned to play Dove (Fig. 3.3A; the best value found for sensitivity is here $\lambda = 10$). As before, this can be explained by the possibility of escaping a basin of attraction: the outcome (Dove, Dove) is also an equilibrium for the dynamics and some pairs of individuals converge to this equilibrium. When a PIL plays against a PIL, we find as predicted that approxi-

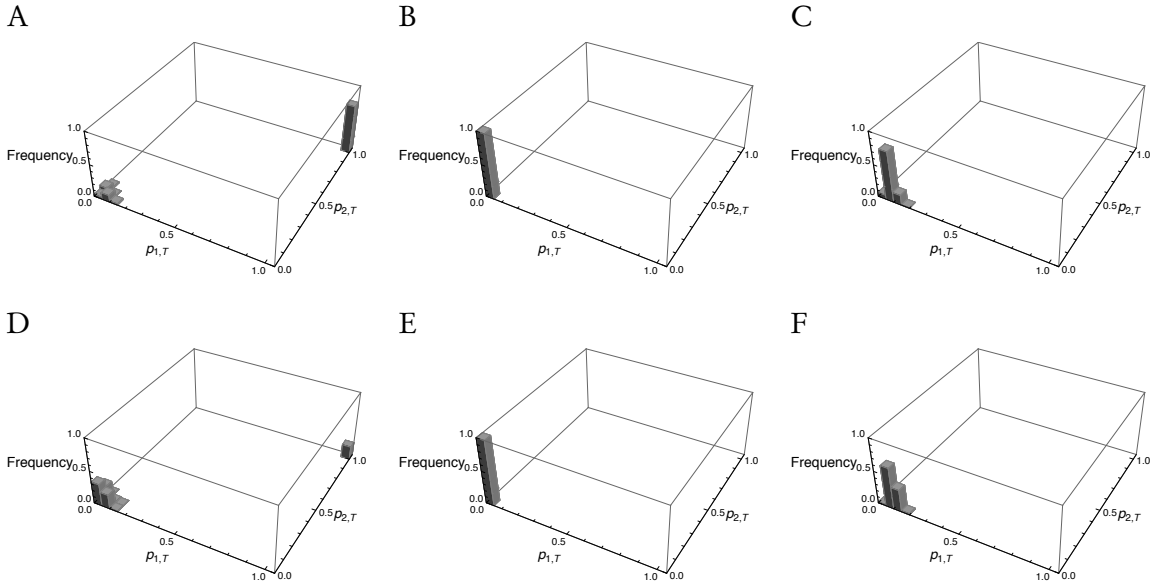

Figure B.1: Distribution of behavior at equilibrium of learning in the average Prisoner's Dilemma for the one-shot matching model with constant learning rate for pairs of opponents. This represents the frequency of pairs having reached the probability to play action 1 $(p_{1,T}, p_{2,T})$ at the end of lifespan, T. We used a total of 1000 individuals of each type in each simulation. First line: initial preference for Cooperation ($p_{i,1} = 0.85$). Second line: initial preference for Defection ($p_{i,1} = 0.15$). Left column: interaction between two PRLs. Middle column: interaction between two PILs. Right Column: interaction between PRL (player 1) and PIL (player 2).

matively half PILs learn Hawk and half learn Dove (Fig. 3.3B). Finally, for PRL vs. PIL, things go as predicted with a vast majority of PRL learning Hawk and a vast majority of PIL learning Dove: PIL gets "exploited" by PRL here (Fig. 3.3C).

When we run simulations of natural selection in a population of PRL and PIL we obtain that PRL fixes for all initial compositions of the population (Table 3.3). Since learning behavior is in conformity with our analytic prediction, this is not a surprise. The important interaction between PRL and PIL turns to the advantage of PRL. The latter is more prompt to learn the Hawk strategy and PIL is penalized by its initial preference for Dove. One unpredicted outcome of learning, namely the fact that PRL will learn to play Dove against itself even if it initially prefers Hawk, gives no special advantage to PRL with the payoff structure of our Hawk-Dove game.

PRL initially prefers to play Dove and PIL prefers Hawk. This initial condition is mirroring the previous case, and the analysis thus predicts that PIL should be the one that learns to play Hawk against PRL (the sensitivity that gives the best match to prediction is $\lambda = 100$ for this case). This is indeed what we observe (Fig. 3.3F). For the interactions between the same types (PRL vs. PRL and PIL vs. PIL), we have the same behavior as in paragraph (a): most PRL learn to play Dove (Fig. 3.3D), and PIL learn half of the time to play Hawk and half of the time to play Dove (Fig. 3.3E). With this learning behavior, PIL invades and fixes in the population for all initial compositions of the population, and this happens very fast (in the first generations; (Table 3.3)).

Hawk Dove game: constant learning rate

PRL initially prefers to play Hawk and PIL prefers Dove. In this situation, we observe qualitatively the same learning behaviors as with a dynamic learning rate. In particular, ERL learns Hawk against BL and the latter learns Dove (Fig. B.2A,B,C). As a consequence, ERL fixes to a frequency of 1 at an evolutionary equilibrium (Table 3.3).

PRL initially prefers to play Dove and PIL prefers Hawk. The result is again very similar to the case with dynamic learning rate. Namely, in ERL vs. BL interactions, ERL learn Dove and BL learn Hawk (Fig. B.2D,E,F) and this implies that BL rapidly take over and fixes in the population under the evolutionary simulations (Table 3.3).

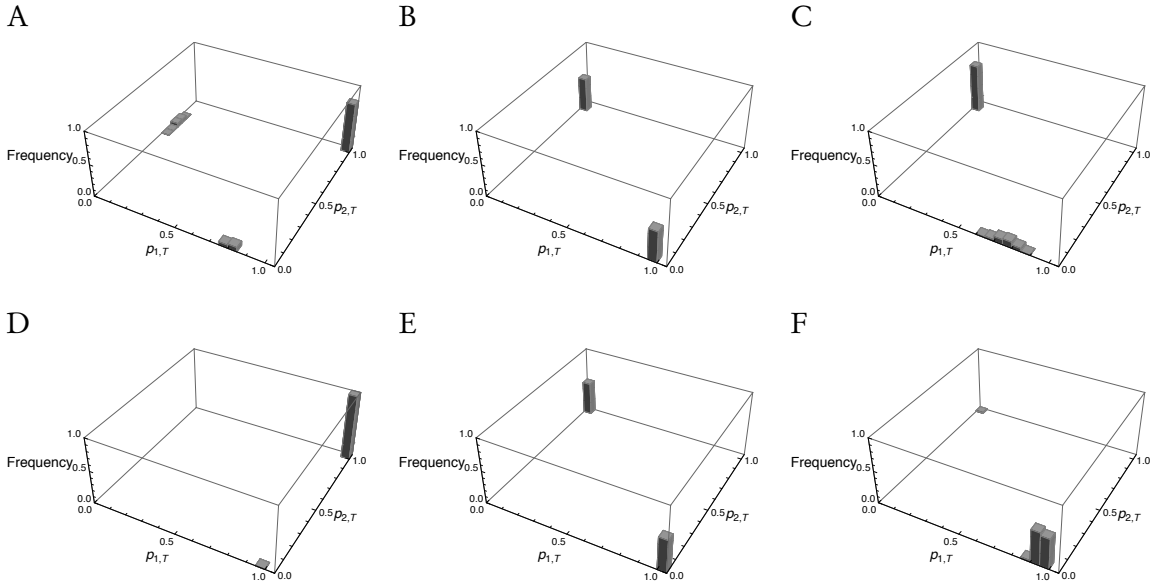

Figure B.2: Distribution of behavior at equilibrium of learning in the average Hawk-Dove game for the one-shot matching model with constant learning rate for pairs of opponents. This represents the frequency of pairs having reached the probability to play action 1 $(p_{1,T}, p_{2,T})$ at the end of lifespan, T. We used a total of 1000 individuals of each type in each simulation. First line: RL initially prefer Hawk ($p_{\mathbf{R},1} = 0.15$) and IL prefers Dove ($p_{\mathbf{I},1} = 0.85$). Second line: RL initially prefer Dove ($p_{\mathbf{R},1} = 0.85$) and IL prefers Hawk ($p_{\mathbf{I},1} = 0.15$). Left column: interaction between two RLs. Middle column: interaction between two ILs. Right Column: interaction between RL (player 1) and IL (player 2).

Coordination game: dynamic learning rate

In this average game, we give to all individuals a preference for the "Left" action and find that all individuals succeed in coordinating at time T. Pairs of PRL individuals coordinate mostly on action 2 ("Right") while pairs of PIL coordinate mostly on action 1 ("Left") (Fig. 3.4A,B). The heterogeneous pairs (PRL vs PIL) coordinate mostly on the Right action (Fig. 3.4C). This result is difficult to explain because individuals had an initial preference for "Left" but since the analysis demonstrates that (Right, Right) is also a stable equilibrium of the associated deterministic system, this result does not contradict our qualitative analysis.

Interestingly, the evolutionary simulations give an outcome different than what we expected. While the above learning behavior suggests that both types should co-exist in equal frequency in the long-run, we find that the population converges to a mixed state with domination of PIL individuals (Table 3.3). Again, it is difficult to know why this happened, but a possible reason for this might be that some PRL individuals converge more slowly to the equilibrium. Even if we chose T big enough, our criterion was based on the average time needed for all individuals to converge in the population. Some individuals might converge more slowly than in T time steps, and fail to coordinate at this time, giving an advantage to PIL.

Coordination game: constant learning rate

The results of learning dynamics under constant learning rate are very similar to the above, with ERL pairs coordinating on Right, BL pairs coordinating on Left and heterogeneous pairs coordinating on Right (Fig. B.3A,B,C).

Regarding evolution, the result is in conformity to our analysis since the population converges to a mixed state where the frequency of both types is close to 0.5 but with a slight domination of ERL individuals ($q \approx 0.56$; Table 3.3).

B.6.2 Repeated matching

Prisoner's Dilemma: dynamic learning rate

All individuals initially prefer Cooperation. In the average PD game, we use the value $\lambda = 10$ that gives the best correspondence between analysis and simulation in the one-shot matching (OM) model. We find that PIL learns to defect irrespective of the frequency of the types in

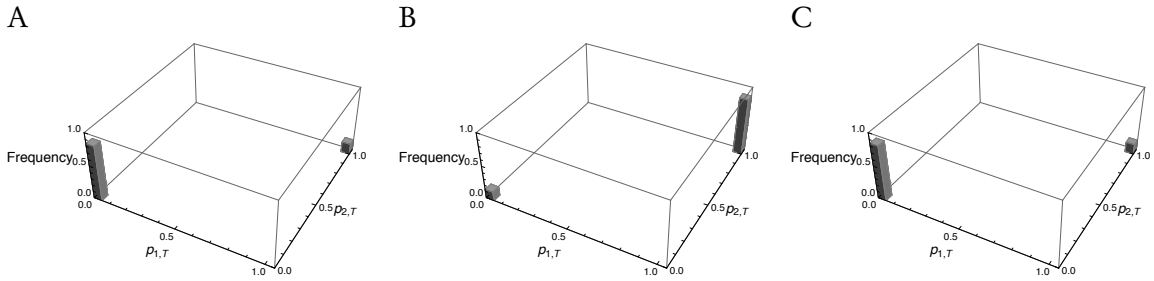

Figure B.3: Distribution of behavior at equilibrium of learning in the average Coordination game for the one-shot matching model with constant learning rate for pairs of opponents. This represents the frequency of pairs having reached the probabilities to play action 1 $(p_{1,T}, p_{2,T})$ at the end of lifespan, T. We used a total of 1000 individuals of each type in each simulation. Left column: interaction between two RLs. Middle column: interaction between two ILs. Right Column: interaction between RL (player 1) and IL (player 2).

the population. However, PRL individuals can learn to cooperate when in sufficiently high frequency in the population (precisely, the average probability of cooperating at equilibrium of learning of PRL is above 0 for $q \geq 0.8$; Fig. 3.5A). This result is not surprising given the OM results. Indeed, when in high frequency, there is a positive probability that an PRL individual meets only other PRLs and the dynamics of those individuals will likely be similar to the dynamics of pairs of PRL in the OM model.

This learning behavior implies that evolution leads to an interior equilibrium with the coexistence of PRL and PIL. Indeed, when PRL are in lower frequency than 0.8, evolution is neutral because everybody defects. However, when the population reaches a state where the frequency of PRL is above 0.8, PRL starts to cooperate so the latter has a disadvantage compared to PIL and have thus a tendency to decrease in frequency. The population thus visits all the states such that the frequency of PRL is less than 0.8 equally often, and in our evolutionary simulations we obtain a stable average frequency of PRL around $q \approx 0.19$ (Table 3.4).

All individuals initially prefer Defection. In this game and with these initial preferences, individuals of both types learn to defect for all frequencies of PRL in the population (Fig. 3.5B).

This leads to think that evolution is neutral here and that the evolutionary simulations should converge to a state where $q = 0.5$. Our simulations give a result close to this prediction but we

also find that PIL slightly dominates the population at an evolutionary equilibrium (Table 3.4).

Prisoner's Dilemma: constant learning rate

All individuals initially prefer Cooperation. At all frequencies, ERL individuals converge to a state where their average probability of cooperation is above 0, which gives a complete advantage to the defector BL in this repeated matching model (Fig. 3.5C). As a consequence, the evolutionary dynamics display a state where the domination of BL is almost total as they nearly fix in the population at an evolutionary equilibrium (Table 3.4).

All individuals initially prefer Defection. In the same vein as when individuals initially prefer Cooperation, ERL players also converge to an average probability of Cooperation that is positive for all values of q, while BL always learn to defect (Fig. 3.5D). This situation makes BL almost fix in the population in an evolutionary long run (Table 3.4).

Hawk Dove game: dynamic learning rate

ERL initially prefers to play Hawk and BL prefers Dove. Even though we implement an initial preference for Hawk to PRL individuals, we find here that, in their learning behavior, PRL converge to a state where their probability of playing Dove is always higher than that of PIL. Moreover, PRL have a tendency to increase their probability of playing Dove as they increase in frequency while we observe the inverse tendency among PIL individuals, who decrease their learned probability of playing Dove as the frequency of PRL increases (Fig. 3.5E).

As a result, the evolutionary simulations of natural selection give the result that both types co-exist in the long-run with a domination of PIL individuals. Indeed, we observe that at low frequencies q, PIL are playing Hawk a little more often than prescribed by the ESS (which is here $1 - B/2C \approx 0.17$ because we use $B = 5$ and $C = 3$), and PRL are playing Dove with a high average probability. This gives an advantage to PRL at low frequencies since they perform better against PIL than PIL perform against themselves (because they play Hawk too often). However, when the frequency of PRL increases, the latter gets exploited more often by PIL because PRL plays more and more Dove while PIL plays more and more Hawk, which gives an advantage to PIL. Hence at high enough frequencies of PRL, PIL has a higher fitness than PRL. This explains the interior equilibrium with a domination of PIL in the evolutionary simulations (Table 3.4).

ERL initially prefers to play Dove and BL prefers Hawk. This initial condition was favoring PIL in the one-shot matching model, and we have the same situation here. We observe that PRL learns to play Dove with high average probability (not smaller than 0.6) for all q, while PIL learns to play Hawk with a high average probability, and this probability even decreases as PRL increases in frequency (Fig. 3.5F).

Hence, PIL obtains a higher payoff against PRL than PRL against PRL which gives it an advantage for high q. However, for low q, PRL against PIL cannot obtain a much better payoff than PIL against PIL since the latter plays close to the ESS at low q. This is why we observe in the evolutionary simulations that PIL dominates the population in an interior equilibrium (Table 3.4).

Hawk Dove game: constant learning rate

ERL initially prefers to play Hawk and BL prefers Dove. The results for this case resemble the case with dynamic learning rate above. Namely, ERL individuals have for all q a learned probability of playing Dove above that of PIL. Moreover, the probability to play Dove of ERL increases as q but this probability decreases for q (Fig. 3.5G).

For the same reason as in the case with dynamic learning rate, natural selection leads to a interior equilibrium with a large domination of PIL (Table 3.4).

ERL initially prefers to play Dove and BL prefers Hawk. Again this situation is very similar to the one with dynamic learning rate above. ERL always has an average probability to play Dove higher than PIL, while the latter plays close to the ESS when frequent in the population and plays almost always Hawk when rare (Fig. 3.5H).

This learning behavior favors BL over ERL and simulations of evolution confirm this by showing that the frequency of ERL at evolutionary equilibrium is around $q \approx 0.05$ (Table 3.4).

Coordination game: dynamic learning rate

In this game, PRL has difficulties in coordinating on the same equilibrium as PIL for all frequencies q, while PIL always succeeds in learning to coordinate on a single action (that can be Left or Right depending on stochastic events in the simulations; Fig. 3.5I).

This learning behavior implies that PRL has lower fitness for all q. Indeed simulations of

natural selection show that PIL fix in the population in the long run (Table 3.4).

Coordination game: constant learning rate

In this case, the learning behavior of BL is similar to the situation with dynamic learning rate, namely, all BL learn to coordinate on a single action. On the other hand, ERL have still difficulties to coordinate for sufficiently high q, but coordinate efficiently for low q (Fig. 3.5J).

This learning behavior implies that evolution should be neutral for low q but should favor BL for higher q. This is indeed consistent with what we obtain when we simulate natural selection, where we observe an interior equilibrium with a large domination of BL ($q \approx 0.06$; Table 3.4).