Theoretical Population Biology 91 (2014) 20–36

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

On learning dynamics underlying the evolution of learning rules

Slimane Dridi, Laurent Lehmann [*]

Department of Ecology and Evolution, University of Lausanne, Switzerland

ARTICLE INFO

Article history:
Available online 17 September 2013

Keywords:
Fluctuating environments
Evolutionary game theory
Stochastic approximation
Reinforcement learning
Fictitious play
Producer–scrounger game

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 agent-based computer simulations. In this paper, we show that stochastic approximation theory can help to qualitatively understand learning dynamics and formulate analytical models for the evolution of learning rules. We consider a population of individuals repeatedly interacting during their lifespan, and 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.

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 of unlimited powers of computation), which curtails the direct application of standard optimality approaches for studying dynamic behavior such as optimal control

[*] Corresponding author.
E-mail address: laurent.lehmann@unil.ch (L. Lehmann).

0040-5809/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.tpb.2013.09.003

theory and dynamic programming. Indeed, the dynamics of even the simplest learning rule, such as reinforcement learning 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, 1985; 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, 2011b; 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 (Großet al., 2008; Josephson, 2008; Hamblin and Giraldeau, 2009; Arbilly et al., 2010, 2011a,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; Benaïm and Hirsch, 1999a; 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 equations 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. Model

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. 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 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 $\mathcal{A} = \{1, \ldots, m\}$. 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 the environment 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.

2.3. 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.

2.3.1. Dynamics of motivations

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

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

where

$$n_{i,t+1} = 1 + \rho_i n_{i,t} \tag{2}$$

is individual i's count of the number of steps of play. The initial conditions of Eqs. (1) and (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. (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. (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. (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) 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. (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

$$\mathbb{1}(a, a_{i,t}) = \begin{cases} 1, & \text{if } a_{i,t} = a, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

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.

2.3.2. 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

$$p_{i,t}(a) = \frac{f(M_{i,t}(a))}{\sum_{k\in\mathcal{A}} f(M_{i,t}(k))}, \tag{4}$$

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

The choice rule (Eq. (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

$$p_{i,t}(a) = \Pr\{a = \operatorname*{argmax}_{b\in\mathcal{A}}\ [M_{i,t}(b) + \varepsilon(b)]\} \tag{5}$$

where $(\varepsilon(b))_{b\in\mathcal{A}}$ 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. (5) restricts the possibilities for the form taken by f. In fact, the only function satisfying at the same time both Eqs. (4)–(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. (4), we obtain that an organism chooses its actions according to the so-called logit choice function

$$p_{i,t}(a) = \frac{\exp[\lambda_i M_{i,t}(a)]}{\sum_{k\in\mathcal{A}} \exp[\lambda_i M_{i,t}(k)]}, \tag{6}$$

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, 2011b).

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 $a^* = \operatorname{argmax}_{b\in\mathcal{A}}\ [M_{i,t}(b)]$ 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.

Table 1
Special cases of EWA learning (Eqs. (1)–(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 D 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	$M_{i,t}(a) + \frac{1}{t+1}\pi_i(a, \mathbf{a}_{-i,t}, \omega_t)$	$\propto \exp[\lambda_i M_{i,t}(a)]$

Note: The expression $\propto \exp[\lambda_i M_{i,t}(a)]$ refers to the logit choice rule Eq. (6).

2.3.3. *Learning rules in the EWA genotype space*

In the EWA model, individuals differ by the value of the four parameters $\phi_{i,t}$, ρ_i, δ_i, λ_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 1, we retrieve from the model (Eq. (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 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).

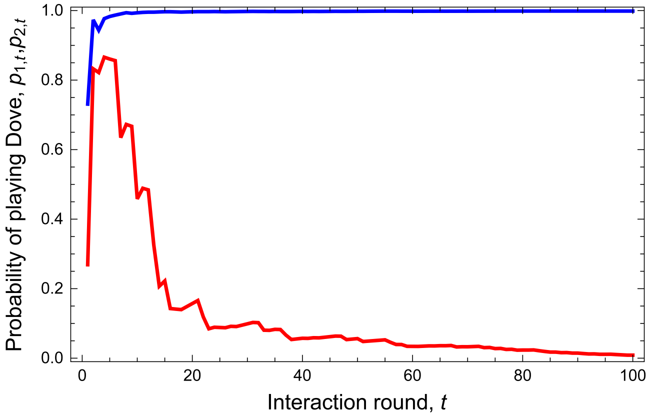

Fig. 1. Example of learning dynamics for two interacting individuals (1 and 2) in a 2 × 2 Hawk–Dove game with $\pi(1,1) = B/2$, $\pi(1,2) = 0$, $\pi(2,1) = B$, $\pi(2,2) = B/2 - C$, where $B = 5$ and $C = 3$. The blue line represents the probability $p_{1,t}$ to play Dove for individual 1 and the red 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 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 $p_{2,1} \approx 0.27$).

In EWA, the learning dynamics of an individual (Eqs. (1)–(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. 1, we show a simulation of a typical learning dynamics of two interacting individuals who learn according to the EWA model (Eqs. (1)–(4)) in a repeated Hawk–Dove game, and with actions play probabilities following the logit choice rule (Eq. (6)). Is it possible to approximate this dynamics in order to obtain a qualitative understanding of the change of play probabilities?

2.4. *Stochastic approximation*

2.4.1. *Differential equations for motivations*

We now use stochastic approximation theory (Ljung, 1977; Benveniste et al., 1991; Benaïm, 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. (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 describe very closely the long-run stochastic dynamics of the motivations (see Benaïm, 1999, for a standard reference, and Hopkins, 2002, for an application of this principle to learning). To that aim, we write Eq. (1) as

$$M_{i,t+1}(a) - M_{i,t}(a) = \frac{1}{n_{i,t+1}}\left[-\epsilon_{i,t} M_{i,t}(a) + R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t)\right] \tag{7}$$

where

$$\epsilon_{i,t} = 1 + n_{i,t}(\rho_i - \phi_{i,t}) \tag{8}$$

is a decay rate and

$$R_i(a, a_{i,t}, \mathbf{a}_{-i,t}, \omega_t) = \left[\delta_i + (1-\delta_i)\mathbb{1}(a, a_{i,t})\right]\pi_i(a, \mathbf{a}_{-i,t}, \omega_t) \tag{9}$$

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 $\lim_{t\to\infty}(1/n_{i,t}) = 0$ (Benaïm, 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). We further assume a constant value of ϵ_i from now on (this is the case for all rules in Table 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 1) with ϕ_i close to 1 (see Benaïm and Hirsch, 1999b; Izquierdo et al., 2007, for results on processes with constant step-sizes).

With these assumptions, we show in Appendix A (Eqs. (A.1)–(A.12)) that the differential equation arising from taking the expected motion of the stochastic dynamics in Eq. (7) is

$$\dot{M}_i(a) = -\epsilon_i M_i(a) + \bar{R}_i(a), \tag{10}$$

where

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

and

$$\bar{\pi}_i(a, \mathbf{a}_{-i}) = \sum_{\omega\in\Omega} \mu(\omega)\pi_i(a, \mathbf{a}_{-i}, \omega). \tag{12}$$

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 $\mathcal{A}^{N-1}$ 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. (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. (10)–(12) define a bona fide autonomous system of differential equations.

Eq. (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. (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.

2.4.2. Differential equations for action play probabilities

Using the logit choice rule (Eq. (6)) and the dynamics of motivations (Eq. (10)), we can derive a differential equation for the choice probability for each action a of individual i

$$\dot{p}_i(a) = p_i(a)\left[\epsilon_i \sum_{k\in\mathcal{A}} \log\left(\frac{p_i(k)}{p_i(a)}\right) p_i(k) + \lambda_i\left(\bar{R}_i(a) - \sum_{k\in\mathcal{A}} \bar{R}_i(k)p_i(k)\right)\right], \tag{13}$$

(Appendix A, Eqs. (A.13)–(A.20)). Because $\bar{R}_i(a)$ depends on the action play probabilities, Eq. (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. (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. (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. (1)–(4)) asymptotically follows very closely the deterministic path of the differential equation (13). For instance, if the limit set of Eq. (13) consists of isolated equilibria, the stochastic process (Eqs. (1)–(4)) will converge to one of these equilibria almost surely (Benaïm, 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. (13). For instance, when the choice rule is the so-called power choice, the dynamics of actions will also depend on the dynamics of motivations (i.e., $f(M) = M^{\lambda_i}$ in Eq. (4), which gives rise to $\dot{p}_i(a) = [\lambda_i p_i(a)/M_i(a)]\left[\bar{R}_i(a) - \sum_{k\in\mathcal{A}} \bar{R}_i(k)p_i(k)\{p_i(a)/p_i(k)\}^{1/\lambda_i}\right]$, Eq. (A.22)). 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. (13)).

3. Applications

3.1. Pure reinforcement vs. payoff-informed learning

We now apply our main result (Eq. (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: 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 $\mathcal{R}_\omega$ the payoff obtained when the game is ω and both players play action 1 (see Table 2 for the description of the payoffs for each game ω), so that the average payoff obtained when both players play action 1 is $\mathcal{R} = \mu(\text{PD})\mathcal{R}_{\text{PD}} + \mu(\text{CG})\mathcal{R}_{\text{CG}} + \mu(\text{HD})\mathcal{R}_{\text{HD}}$. Likewise, one can evaluate the payoffs $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{P}$ 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).

Table 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 $\mathcal{T}_{\rm PD} > \mathcal{R}_{\rm PD} > \mathcal{P}_{\rm PD} > \mathcal{S}_{\rm PD}$ and $(\mathcal{T}_{\rm PD} + \mathcal{S}_{\rm PD})/2 < \mathcal{R}_{\rm PD}$. In the Hawk–Dove (Middle), we have $\mathcal{T}_{\rm HD} > \mathcal{R}_{\rm HD}$, $\mathcal{S}_{\rm HD} > \mathcal{P}_{\rm HD}$, $\mathcal{P}_{\rm HD} > \mathcal{R}_{\rm HD}$. In the Coordination Game (Right), $\mathcal{R}_{\rm CG} > \mathcal{S}_{\rm CG}$, $\mathcal{R}_{\rm CG} = \mathcal{P}_{\rm CG}$, $\mathcal{S}_{\rm CG} = \mathcal{T}_{\rm CG}$.

$\bar{G}$	**Dove**	**Hawk**						
Dove	$\mathcal{R} = B/2$	$\mathcal{S} = 0$						
Hawk	$\mathcal{T} = B$	$\mathcal{P} = B/2 - C$						
PD	**Cooperate**	**Defect**	HD	**Dove**	**Hawk**	CG	**Left**	**Right**
Cooperate	$\mathcal{R}_{\rm PD}$	$\mathcal{S}_{\rm PD}$	**Dove**	$\mathcal{R}_{\rm HD}$	$\mathcal{S}_{\rm HD}$	**Left**	$\mathcal{R}_{\rm CG}$	$\mathcal{S}_{\rm CG}$
Defect	$\mathcal{T}_{\rm PD}$	$\mathcal{P}_{\rm PD}$	**Hawk**	$\mathcal{T}_{\rm HD}$	$\mathcal{P}_{\rm HD}$	**Right**	$\mathcal{T}_{\rm CG}$	$\mathcal{P}_{\rm CG}$

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

$$\phi_{i,t} = 1 + \frac{1}{t} \quad \text{and} \quad \rho_i = 1 \tag{14}$$

so that when $\delta_i = 0$ we obtain a form of reinforcement learning, which we call Pure Reinforcement Learning (PRL: see Table 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 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. (14) into Eq. (8) gives $\epsilon_{i,t} = 0$ (since $n_t = t$) and thus $\epsilon_i = 0$ in Eq. (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. (8), the above assumptions, and Table 2, that the action play probabilities satisfy the dynamics

$$\begin{aligned}\dot{p}_1 \;=\; & p_1(1-p_1)\lambda_1[\{p_2\mathcal{R} + (1-p_2)\mathcal{S}\}\{p_1 + \delta_1(1-p_1)\} \\ & - \{p_2\mathcal{T} + (1-p_2)\mathcal{P}\}\{\delta_1 p_1 + (1-p_1)\}],\end{aligned} \tag{15}$$

$$\begin{aligned}\dot{p}_2 \;=\; & p_2(1-p_2)\lambda_2[\{p_1\mathcal{R} + (1-p_1)\mathcal{S}\}\{p_2 + \delta_2(1-p_2)\} \\ & - \{p_1\mathcal{T} + (1-p_1)\mathcal{P}\}\{\delta_2 p_2 + (1-p_2)\}].\end{aligned} \tag{16}$$

In order to compare the dynamics predicted by Eqs. (15)–(16) to that obtained from iterating Eq. (1) with logit choice function Eq. (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 $\mu(\text{PD}) = \mu(\text{CG}) = \mu(\text{HD}) = 1/3$.

3.1.1. PRL vs. PRL

When two PRL play against each other (Eqs. (15)–(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. 2A): the two pure asymmetric Nash equilibria, (Dove, Hawk) and (Hawk, Dove), and the Pareto efficient outcome where both individuals play Dove (Appendix B). 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. 2A).

In Fig. 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. (15)–(16) to that obtained from Eq. (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 more likely to play Dove initially since its probability to play Dove is bigger than 0.5, while the player is more likely to play Hawk if the difference is negative ($\Delta M_{i,1} < 0$, which entails a probability to play Dove lower than 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. 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 a probability to play Dove that is bigger than 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 a probability to play Dove lower than 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, which is very close to the lower-left corner of the state space in Fig. 2A. Preference reversal may then take a very long time, and Fig. 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. 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. 3).

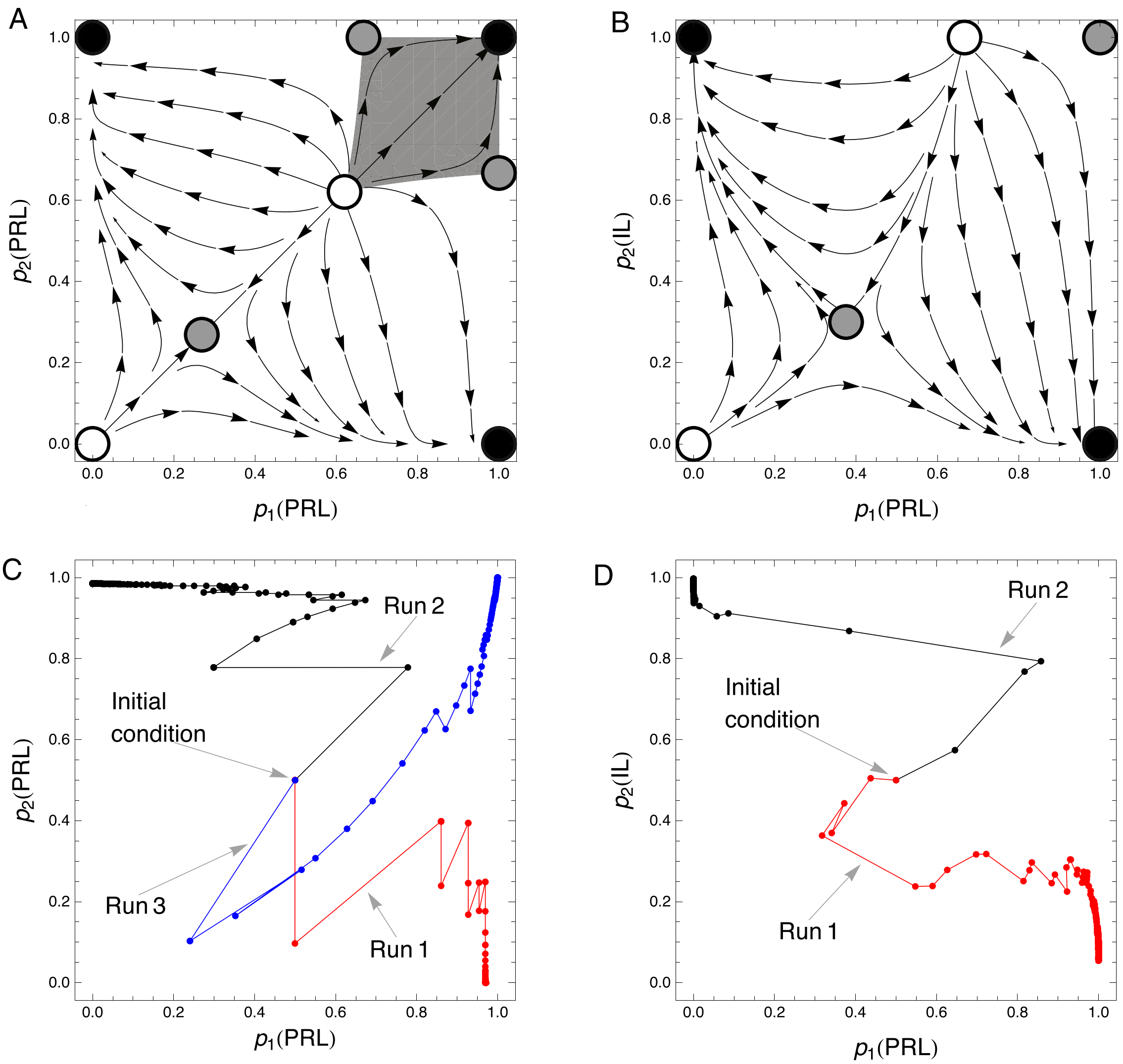

Fig. 2. Solution orbits for the deterministic dynamics (panels A and B) and sample paths for the stochastic dynamics (panels C and D) for two learners in the average Hawk–Dove game ($B = 5$ and $C = 3$), where the x and y axis represent the probabilities of playing Dove by player 1 and 2, respectively. 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.

3.1.2. PRL vs. IL

When a PRL plays against an IL (Eqs. (15)–(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 B, Eq. (B.2)).

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. 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. (15)–(16)) generally approximates qualitatively well the quasi-equilibrium probabilities of playing action obtained under the stochastic learning processes (Eqs. (1)–(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. 4). Now that we have a feeling 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.

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

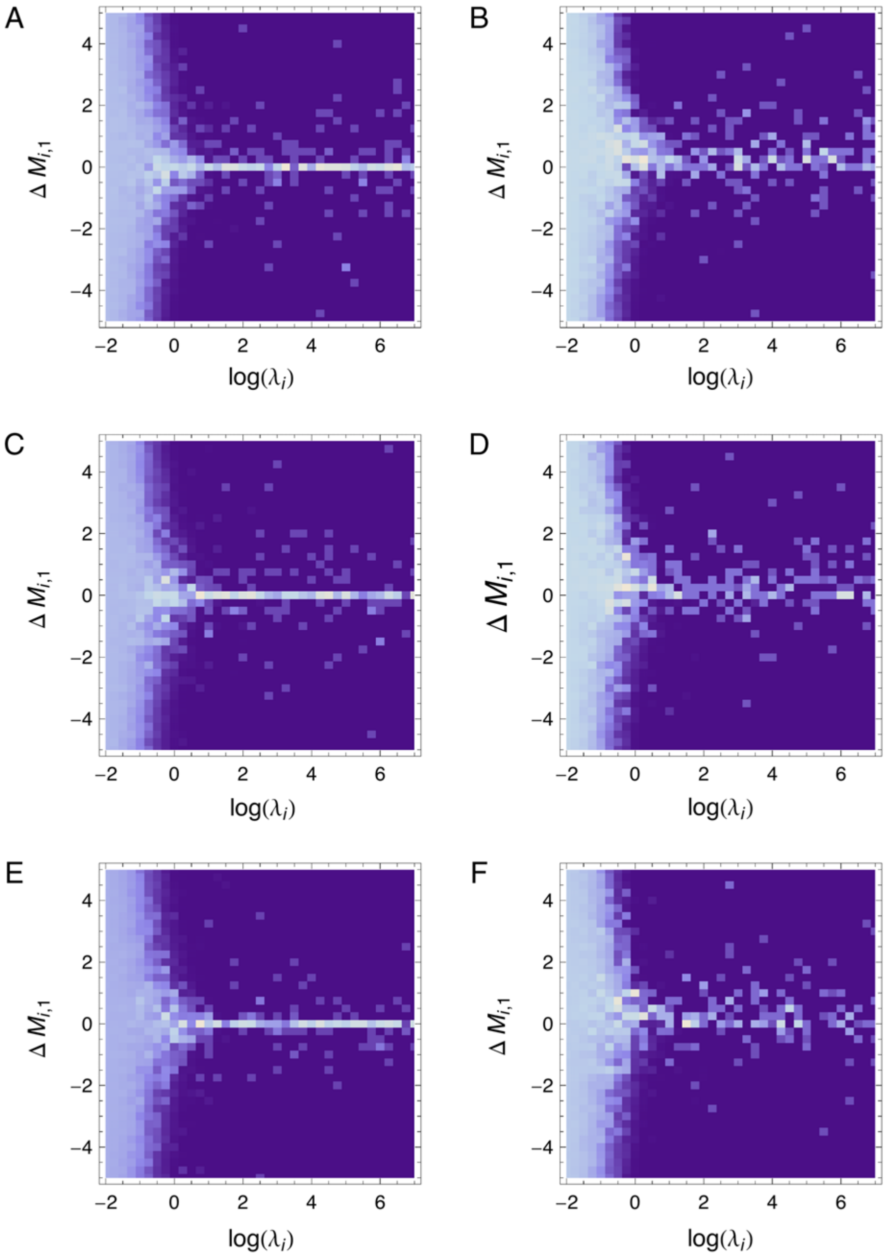

Fig. 3. Density plot of the average distance between the probability of playing Dove obtained from the equilibrium of the deterministic model (Eqs. (15)–(16)) and the stochastic learning dynamics, as a function of λ_i and $\Delta M_{i,1}$. Each data pixel is the average over 5 simulation runs. Lightly shaded regions indicates a big euclidean distance ($\sim\sqrt{2}$) between simulations and analytic prediction, while dark regions indicates a small distance ($\sim$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 $T = 1000$.

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 is competition between learning rules.

3.2.1. *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, for example, 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 its value with the scrounger so its payoff is $\pi_i(a_{i,t}, 1) = V(a_{i,t})/2$.

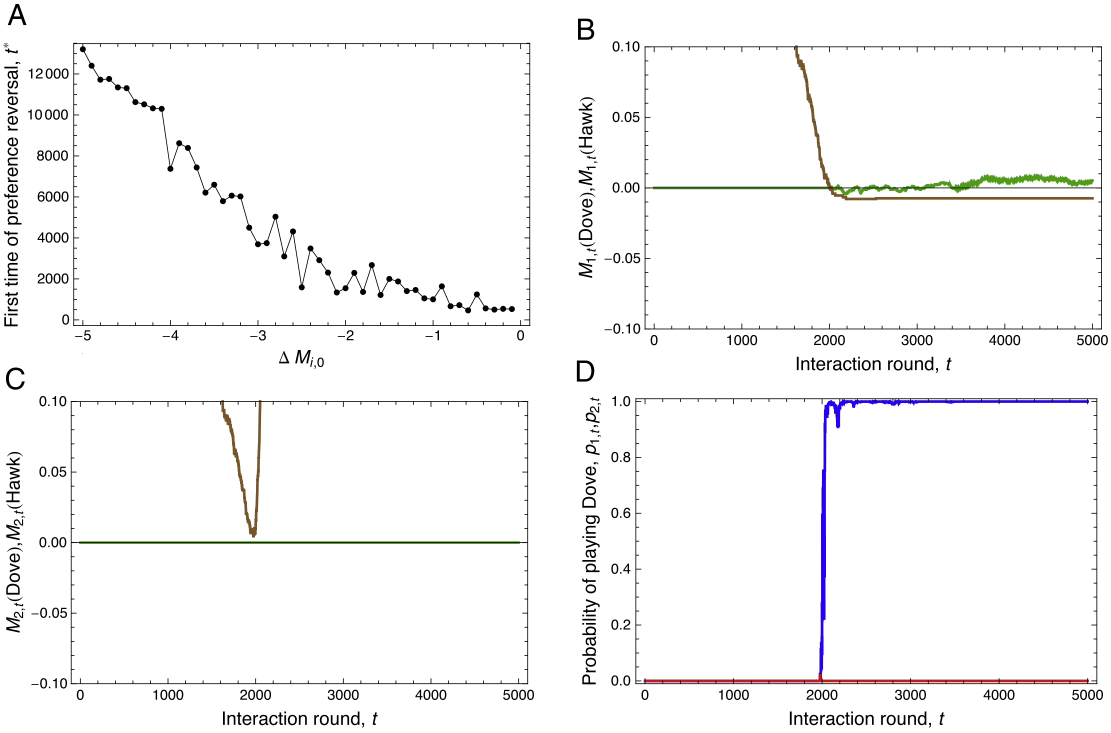

Fig. 4. Preference reversal for two PRL playing the average Hawk–Dove game with same initial preferences for Hawk ($\Delta M_{i,1} < 0$) and a high sensitivity to motivations ($\lambda_i = 1000$). In panel A, we graph the first time t^* that a preference reversal occurs (i.e., first time that $\Delta M_{i,t^*} > 0$ 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).

We assume that there are three types of individuals in this population: Scroungers (S), Fictitious Players (FP), and Exploratory Reinforcement Learners (ERL, see Table 1), where both learners (ERL and FP) are exploratory ($\epsilon = 1$ in Eq. (13)). The life-cycle of these individuals is as described above (Section 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 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.

3.2.2. Fictitious play

Substituting $\epsilon = 1$ into Eq. (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

$$\dot{p}_\mathbf{F} = p_\mathbf{F}\left[\log\left(\frac{1-p_\mathbf{F}}{p_\mathbf{F}}\right)(1-p_\mathbf{F}) + \lambda_\mathbf{F}\left(\bar{R}_\mathbf{F}(1)\right.\right.$$
$$\left.\left. - \left[\bar{R}_\mathbf{F}(1)p_\mathbf{F} + \bar{R}_\mathbf{F}(2)(1-p_\mathbf{F})\right]\right)\right]. \tag{17}$$

For this model of patch choice, the expected reinforcement to an FP (Eq. (11)) when foraging on patch type a is

$$\bar{R}_\mathbf{F}(a) = (1-s)V(a) + s\frac{V(a)}{2}$$
$$= \left(1-\frac{s}{2}\right)V(a), \quad a = 1, 2, \tag{18}$$

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

$$\hat{p}_\mathbf{F} = \frac{1}{1+\exp\left(\frac{\lambda_\mathbf{F}[2-s][V(2)-V(1)]}{2}\right)}. \tag{19}$$

Fig. 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.

3.2.3. Exploratory reinforcement learning

For exploratory reinforcement learning, substituting $\epsilon = 1$ into Eq. (13), we obtain that the probability that an ERL visits patch 1 obeys

$$\dot{p}_\mathbf{E} = p_\mathbf{E}\left[\log\left(\frac{1-p_\mathbf{E}}{p_\mathbf{E}}\right)(1-p_\mathbf{E}) + \lambda_\mathbf{E}\left(\bar{R}_\mathbf{E}(1)\right.\right.$$
$$\left.\left. - \left[\bar{R}_\mathbf{E}(1)p_\mathbf{E} + \bar{R}_\mathbf{E}(2)(1-p_\mathbf{E})\right]\right)\right], \tag{20}$$

where the expected reinforcements (Eq. (11)) of going on patch types 1 and 2 are, respectively, given by

$$\bar{R}_\mathbf{E}(1) = p_\mathbf{E}\left(1-\frac{s}{2}\right)V(1),$$
$$\bar{R}_\mathbf{E}(2) = (1-p_\mathbf{E})\left(1-\frac{s}{2}\right)V(2). \tag{21}$$

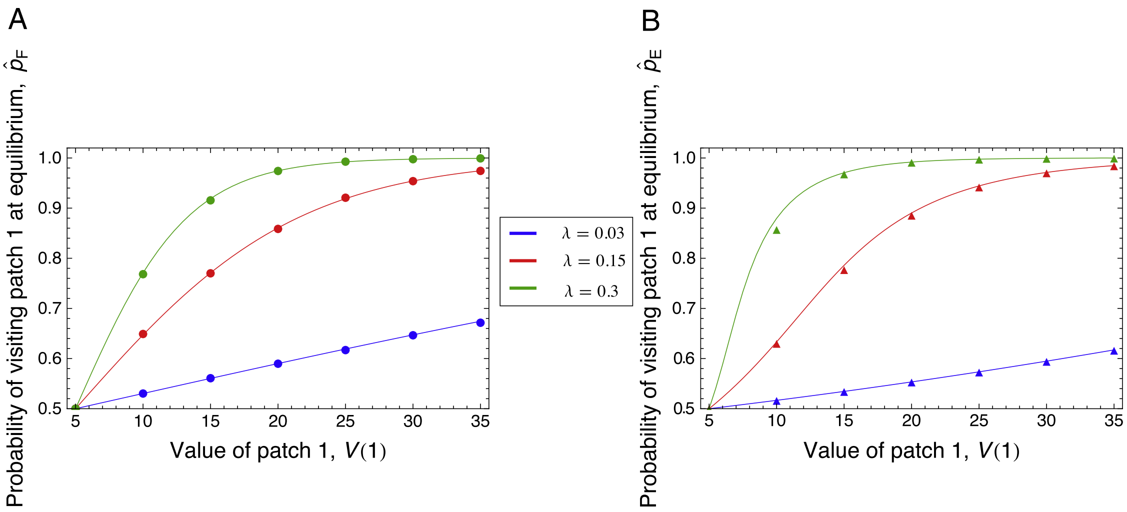

Fig. 5. Panel A: equilibrium probability $\hat{p}_{\mathbf{F}}$ of visiting patch 1 for a FP (Eq. (19)) graphed as a function of the value of patch 1, $V(1)$, for different values of λ_i. We fixed the value of patch 2 at $V(2) = 5$, the frequency of scroungers to $s = (3 - \sqrt{5})/2$ and $\lambda_i = 0.03$ for the blue line, $\lambda_i = 0.15$ for the red line, and $\lambda_i = 0.3$ 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 $\hat{p}_{\mathbf{E}}$ of visiting patch 1 for a ERL (obtained by solving for the equilibrium of Eqs. (20)–(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 ($p_{\mathbf{E},1}$) ranging from 0 to 1.

The equilibria of Eq. (20) cannot be solved analytically (this is a transcendental equation in $p_{\mathbf{E}}$). We thus relied on a numerical analysis to obtain its fixed points ($\dot{p}_{\mathbf{E}} = 0$) and focused on the variation of $\lambda_{\mathbf{E}}$ values by performing a bifurcation analysis (with Newton's method using Mathematica, Wolfram Research, Inc., 2011). Fig. 6 shows that the phase line passes through three different regimes as $\lambda_{\mathbf{E}}$ increases. These regimes are separated by two critical values of $\lambda_{\mathbf{E}}$, which we will call $\lambda_{\mathbf{E}}^{\mathrm{crit}_1}$ and $\lambda_{\mathbf{E}}^{\mathrm{crit}_2}$, and provide the following cases.

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

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

(III) When $\lambda_{\mathbf{E}} > \lambda_{\mathbf{E}}^{\mathrm{crit}_2}$, there is only one interior equilibrium that is unstable. As $\lambda_{\mathbf{E}}$ 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. 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 D, Eq. (C.3)), which allows us to obtain an expected equilibrium probability that an ERL visits patch 1. When $\lambda_{\mathbf{E}}$ becomes large, this expectation is the average over visiting only patch 1 or 2, which gives

$$\hat{p}_{\mathbf{E}} = \frac{V(1)}{V(1) + V(2)} \tag{22}$$

(Appendix D). 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.

3.2.4. *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 ($\hat{p}_{\mathbf{F}}$ given Eq. (19) for FP and $\hat{p}_{\mathbf{E}}$ given by the expectation over the various equilibria like in Eq. (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

$$b_{\mathbf{F}} = \alpha + \hat{p}_{\mathbf{F}} \left(\frac{2-s}{2}\right) V(1) + (1 - \hat{p}_{\mathbf{F}}) \left(\frac{2-s}{2}\right) V(2) - k,$$

$$b_{\mathbf{E}} = \alpha + \hat{p}_{\mathbf{E}} \left(\frac{2-s}{2}\right) V(1) + (1 - \hat{p}_{\mathbf{E}}) \left(\frac{2-s}{2}\right) V(2) - k, \tag{23}$$

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

$$b_{\mathbf{S}} = \alpha + \frac{1-s}{s} \left[\frac{q}{1-s} \left(\hat{p}_{\mathbf{F}} \frac{V(1)}{2} + (1 - \hat{p}_{\mathbf{F}}) \frac{V(2)}{2} \right) \right. \\ \left. + \frac{1-q-s}{1-s} \left(\hat{p}_{\mathbf{E}} \frac{V(1)}{2} + (1 - \hat{p}_{\mathbf{E}}) \frac{V(2)}{2} \right) \right]. \tag{24}$$

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 $\hat{p}_i[V(1)/2] + [1 - \hat{p}_i][V(2)/2]$.

3.2.5. *ESS analysis*

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

$$\Delta q = q \left(b_{\mathbf{F}} - \bar{b} \right) / \bar{b}$$

$$\Delta s = s \left(b_{\mathbf{S}} - \bar{b} \right) / \bar{b}, \tag{25}$$

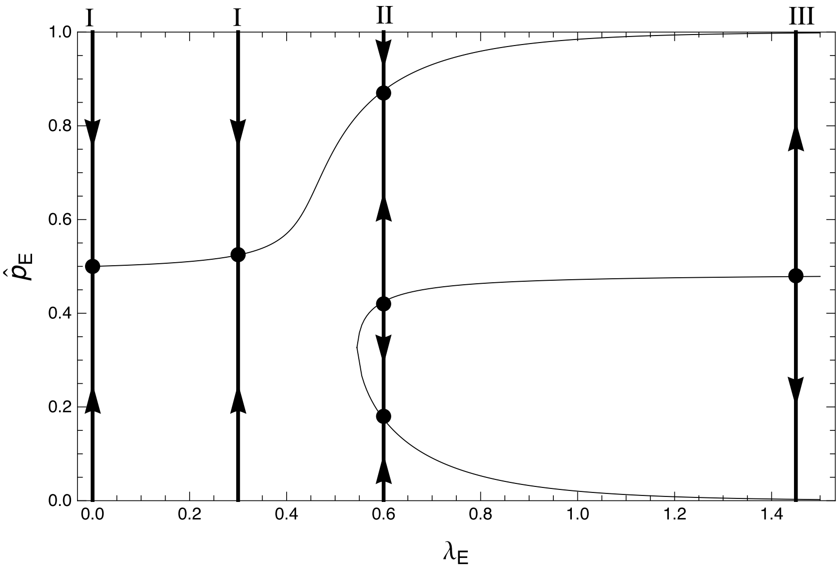

Fig. 6. Bifurcation diagram for the differential equation (20) that describes the learning behavior of ERL in the producer–scrounger model as a function of $\lambda_{\mathbf{E}}$. The thin curves describe the equilibrium values of $\hat{p}_{\mathbf{E}}$ and the thick vertical lines are phase lines at the corresponding values of $\lambda_{\mathbf{E}}$. Dots on the phase lines denote interior equilibria. Our numerical exploration suggests that there are three possible phase lines depending on the value of $\lambda_{\mathbf{E}}$ (indicated by I, II, and III). Parameter values: $s = (3 - \sqrt{5})/2$, $V(1) = 5.3$, $V(2) = 5$.

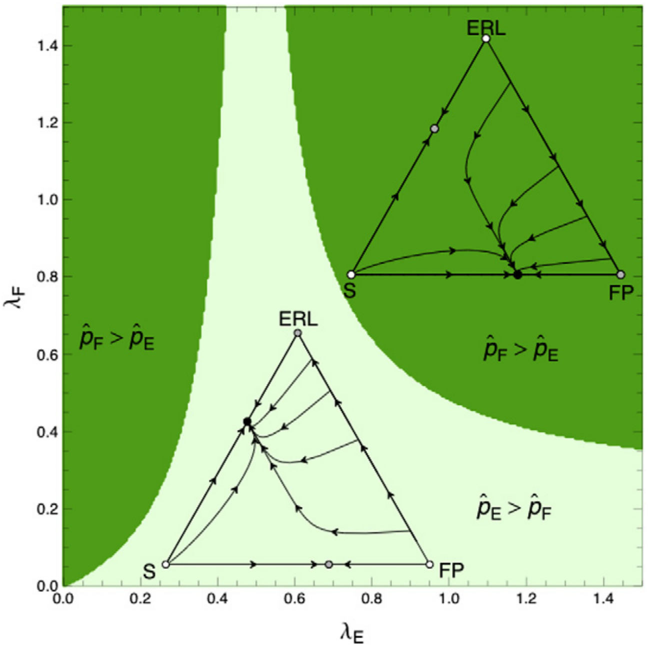

Fig. 7. Solution orbits of the evolutionary dynamics (Eq. (25)) in the producer–scrounger model on the 3-strategy simplex as a function of $\lambda_{\mathbf{E}}$ and $\lambda_{\mathbf{F}}$. In the light shaded region, ERL is the most performant ($\hat{p}_{\mathbf{E}} > \hat{p}_{\mathbf{F}}$), while in the dark shaded region, FP is the most performant ($\hat{p}_{\mathbf{F}} > \hat{p}_{\mathbf{E}}$). The simplex drawn in the light region is plotted for $\hat{p}_{\mathbf{E}} > \hat{p}_{\mathbf{F}}$ 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: $s = \left(3 - \sqrt{5}\right)/2$, $V(1) = 5.3$, $V(2) = 5$.

where $\bar{b} = qb_{\mathbf{F}} + sb_{\mathbf{S}} + (1-q-s)b_{\mathbf{E}}$ is the mean reproductive output in the population. The evolutionary dynamics (Eq. (25) with Eqs. (23)–(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, $(\frac{1}{2}(\sqrt{5} - 1), \frac{1}{2}(3 - \sqrt{5}), 0)$, and one with a coexistence between ERL and scroungers at the same frequency as in the previous case: $(0, \frac{1}{2}(3 - \sqrt{5}), \frac{1}{2}(\sqrt{5} - 1))$. 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. (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

$$\hat{p}_{\mathbf{E}} > \hat{p}_{\mathbf{F}}. \tag{26}$$

This invasion condition is not necessarily satisfied when $\lambda_{\mathbf{E}} > \lambda_{\mathbf{F}}$, and in Fig. 7 we display the regions of values of $\lambda_{\mathbf{F}}$ and $\lambda_{\mathbf{E}}$ 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 $\lambda_{\mathbf{E}}$ becomes very large, it is possible to have an exact invasion condition by substituting Eqs. (19) and (22) into Eq. (27), which implies that ERL invades the stable mix of FP and S if and only if

$$\lambda_{\mathbf{F}} < \frac{2\log\left(\frac{V(1)}{V(2)}\right)}{[V(1) - V(2)]\left(1 + \sqrt{5}\right)/2}. \tag{27}$$

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 ($\lambda_{\mathbf{F}}$ and $\lambda_{\mathbf{E}}$). 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 $\lambda_{\mathbf{E}}$ for ERL does not always leads to a higher probability of visiting patch 1. When $\lambda_{\mathbf{E}}$

is relatively small, this is actually true (in regime I of the learning dynamics of ERL, Fig. 6) but when $\lambda_{\mathbf{E}}$ 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 ($\hat{p}_{\mathbf{E}} < 0.5$). This makes ERL less performant than FP for high values of λ_i (the upper-right region in Fig. 7). Further, when λ_i is very small (the lower-left region in Fig. 7), ERL seems to be less sensitive than FP to an increase in λ_i.

4. Discussion

In this paper, we used stochastic approximation theory (Ljung, 1977; Benveniste et al., 1991; Fudenberg and Levine, 1998; Benaïm, 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. (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. (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. (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 replicator 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, Benaïm, 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. 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 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. (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

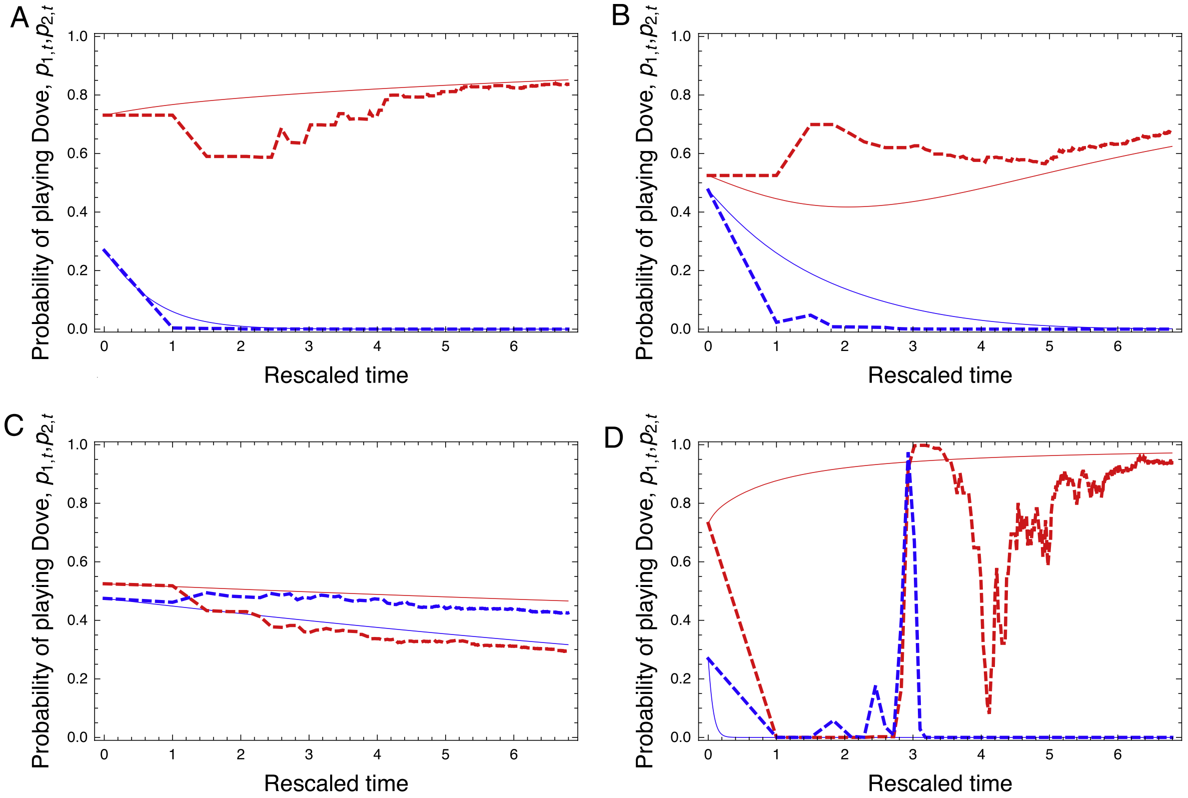

Fig. 8. Comparison between the deterministic (thin, plain lines) and the stochastic (thick, dashed lines) time dynamics of playing Dove for a PRL meeting an IL, and for different values of λ_i and $\Delta M_{i,1}$. The blue line is for the PRL while the red line is for the IL. We always set opposed initial 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 Benaïm, 1999), which is used to plot the numerical solution of the differential equations.

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.

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.

Appendix A. Stochastic approximation

A.1. Expected motion of motivations

Here, we derive Eq. (10) of the main text from Eq. (7). To that end, we call $\mathbf{M}_{i,t} = (M_{i,t}(1), \dots, M_{i,t}(m))$ the vector collecting the motivations of individual i at time t and $\mathbf{M}_t = (\mathbf{M}_{1,t}, \dots, \mathbf{M}_{N,t})$ 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. (9) given current motivational state can be written as

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

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. (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. (4), whereby

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

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

With the above definitions, we can write Eq. (7) as

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

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. (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 (Benaïm, 1999; Benaïm 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. (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) (Benaïm, 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.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. (1)). First, recall from Eq. (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

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

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

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

Factoring out, we have

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

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

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

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

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

A.3. Differential equation for the choice probabilities

A.3.1. Logit choice

Here, we derive the ODE for the choice probabilities (Eq. (13)) by combining the ODE for the motivations (Eq. (10)) with the choice rule (Eq. (4)), under the assumption that the choice rule is the logit choice function (Eq. (6)).

Differentiating the left and right member of Eq. (4) with respect to time t, we have by the chain rule

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

and substituting Eq. (4) gives

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

Using $f(M) = \exp(\lambda_i M)$ in the choice function (Eq. (4)) gives Eq. (6), which implies

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

whereby Eq. (A.14) can be written as

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

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

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

But from the choice probabilities (Eq. (6)) we have the identity

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

which gives

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

and on substitution into Eq. (A.17) produces

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

A.3.2. Power choice

Here, we perform the same derivation as in the last section but assume that $f(M) = M^{\lambda_i}$ in Eq. (4). In this case, $[\mathrm{d}f(M_i(a))/\mathrm{d}M_i(a)] / \sum_{k\in\mathcal{A}} f(M_i(k)) = \left[\lambda_i M_i(a)^{\lambda_i-1}\right] / \sum_{k\in\mathcal{A}} f(M_i(k)) = \lambda_i p_i(a)/M_i(a)$, whereby using $\dot{M}_i(a) = -\epsilon_i M_i(a) + \bar{R}_i(a)$ in Eq. (A.14) yields

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

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

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

Appendix B. Learning to play Hawk and Dove

Here, we analyze qualitatively the vector fields of Eqs. (15)–(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.

B.1. PRL vs. PRL

Pure Reinforcement Learning corresponds to $\delta_i = 0$. Thus, replacing $\delta_1 = \delta_2 = 0$ in Eqs. (15)–(16) and using the payoffs of the Hawk–Dove game (Table 2) produces

$$\begin{aligned}\dot{p}_1 &= p_1(1-p_1)\lambda_1 \\ &\quad\times\left[p_1p_2\frac{B}{2} + (1-p_1)\left\{p_2B + (1-p_2)\left(\frac{B}{2}-C\right)\right\}\right], \\ \dot{p}_2 &= p_2(1-p_2)\lambda_2 \\ &\quad\times\left[p_2p_1\frac{B}{2} + (1-p_2)\left\{p_1B + (1-p_1)\left(\frac{B}{2}-C\right)\right\}\right].\end{aligned} \tag{B.1}$$

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 3). 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. 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 $(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}})$ 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 $(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}})$.

Table 3
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
$(0, 0)$	$\left(-\frac{B}{2}, -\frac{B}{2}\right)$	$(-, -)$
$(0, 1)$	$(-B, 0)$	$(-, 0)$
$(1, 0)$	$(-B, 0)$	$(-, 0)$
$(1, 1)$	$\left(-\frac{B}{2} + C, -\frac{B}{2} + C\right)$	$(+, +)$
$(0, \frac{1}{3})$	$\left(-\frac{B}{3}, \frac{B}{3}\right)$	$(-, +)$
$(\frac{1}{3}, 0)$	$\left(-\frac{B}{3}, \frac{B}{3}\right)$	$(-, +)$
$\left(\frac{B}{2B+\sqrt{2B(B-C)}}, \frac{B}{2B+\sqrt{2B(B-C)}}\right)$		$(+, +)$
$\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$		$(-, +)$

B.2. PRL vs. IL

Payoff-Informed Learning (IL) corresponds to $\delta_i = 1$. Thus, replacing $\delta_1 = 0$ and $\delta_2 = 1$ in Eqs. (15)–(16) and using the payoffs of the Hawk–Dove game (Table 2), one obtains the dynamical system describing learning between PRL (player 1) and IL (player 2) as

$$\begin{aligned}\dot{p}_1 &= p_1(1-p_1)\lambda_1 \\ &\quad\times\left[p_1p_2\frac{B}{2} - (1-p_1)\left\{p_2B + (1-p_2)\left(\frac{B}{2}-C\right)\right\}\right], \\ \dot{p}_2 &= p_2(1-p_2)\lambda_2 \\ &\quad\times\left[p_1\frac{B}{2} - \left\{p_1B + (1-p_1)\left(\frac{B}{2}-C\right)\right\}\right].\end{aligned} \tag{B.2}$$

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. 2B).

Appendix C. Exploratory reinforcement learning

Here we analyze the equilibria of Eq. (20) in the producer–scrounger model when $\lambda_{\mathbf{E}}$ is very large, in which case the second term in Eq. (20) $(\bar{R}_{\mathbf{E}}(1) - [\bar{R}_{\mathbf{E}}(1)p_{\mathbf{E}} + \bar{R}_{\mathbf{E}}(2)(1-p_{\mathbf{E}})])$ dominates the first $[(1-p_{\mathbf{E}})\log([1-p_{\mathbf{E}}]/p_{\mathbf{E}})]$, which we neglect. We then find that there are three equilibria to this differential equation: $\hat{p}_{\mathbf{E}} = 0$, $\hat{p}_{\mathbf{E}} = 1$ and $\hat{p}_{\mathbf{E}} = V(2)/(V(1)+V(2))$. The interior equilibrium $[V(2)/(V(1)+V(2))]$ is unstable since

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

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

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

More generally, Eq. (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, $\hat{p}^e_{\mathbf{E}}$ 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

$$\hat{p}_{\mathbf{E}} = \sum_{e \in E} \beta_e \hat{p}^e_{\mathbf{E}}, \tag{C.3}$$

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

For instance, when we are in equilibrium regime I (Fig. 6), i.e., when there is one stable equilibrium, there is only one term in the sum of Eq. (C.3). When we are in equilibrium regime II and III, there are two terms in the sum.

Appendix D. 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 1), and we subtract aspiration levels to the original motivations, that is,

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

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

In order to prove that Eq. (D.1) combined with Eq. (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

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

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

For Eqs. (D.1) and (6) with $\lambda_i = \infty$ to produce TFT behavior, one needs that

$$\begin{cases} M_{i,t+1}(\mathrm{C}) > M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ M_{i,t+1}(\mathrm{C}) < M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D}, \end{cases} \tag{D.2}$$

where $\mathbf{a}_{-i,t}$ denotes here the action of the single opponent of individual i. Substituting the definition of the motivations (Eq. (D.1)) into Eq. (D.2), we have

$$\begin{cases} \pi_i(\mathrm{C},\mathrm{C}) - L_i(\mathrm{C}) > \pi_i(\mathrm{D},\mathrm{C}) - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ \pi_i(\mathrm{C},\mathrm{D}) - L_i(\mathrm{C}) < \pi_i(\mathrm{D},\mathrm{D}) - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D} \end{cases} \tag{D.3}$$

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. (D.1) produces TFT behavior if

$$\begin{cases} \mathcal{R} - L_i(\mathrm{C}) > \mathcal{T} - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{C}, \\ \mathcal{S} - L_i(\mathrm{C}) < \mathcal{P} - L_i(\mathrm{D}), & \text{if } \mathbf{a}_{-i,t} = \mathrm{D}, \end{cases} \tag{D.4}$$

which can be expressed as the single condition

$$\mathcal{T} - \mathcal{R} < L_i(\mathrm{D}) - L_i(\mathrm{C}) < \mathcal{P} - \mathcal{S}. \tag{D.5}$$

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 $\mathcal{T} - \mathcal{R} > 0$). Also, clearly not all Prisoner's Dilemma games satisfy condition D.5. More precisely, this condition entails that defection needs to risk dominate cooperation ($\mathcal{T} - \mathcal{P} < \mathcal{R} - \mathcal{S}$) for our version of TFT to be implementable.

References

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

Anderson, S.P., de Palma, A., Thisse, J.-F., 1992. Discrete Choice Theory of Product Differentiation. The MIT Press, Cambridge, MA.

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

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

Arbilly, M., Motro, U., Feldman, M.W., Lotem, A., 2011a. Evolution of social learning when high expected payoffs are associated with high risk of failure. Journal of The Royal Society Interface 8, 1604–1615.

Arbilly, M., Motro, U., Feldman, M.W., Lotem, A., 2011b. Recombination and the evolution of coordinated phenotypic expression in a frequency-dependent game. Theoretical Population Biology 80, 244–255.

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

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

Benaïm, M., 1999. Dynamics of stochastic approximation algorithms. In: Azéma, J., et al. (Eds.), Séminaire de Probabilités XXXIII, vol. 1709. Springer, Berlin, pp. 1–68.

Benaïm, M., El Karoui, N., 2005. Promenade Aléatoire: Chaînes de Markov et Simulations; Martingales et Stratégies. Editions de l'Ecole Polytechnique, Palaiseau, France.

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

Benaïm, M., Hirsch, M.W., 1999b. Stochastic approximation algorithms with constant step size whose average is cooperative. The Annals of Applied Probability 9, 216–241.

Benveniste, A., Metivier, M., Priouret, P., 1991. Adaptive Algorithms and Stochastic Approximations. Springer-Verlag, New York (NY).

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

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

Börgers, T., Sarin, R., 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, MA.

Boyd, R., Richerson, P.J., 1985. 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. Wiley, New York, pp. 374–376.

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

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

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

Cavalli-Sforza, L.L., Feldman, M.W., 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.

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

Cho, I.-K., Matsui, A., 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. Hachette, L., Paris, France.

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

Dugatkin, L.A., 2010. Principles of Animal Behavior, second ed. WW Norton & Co, New York (NY).

Dunlap, A.S., Stephens, D.W., 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., Clayton, N.S., 2004. The mentality of crows: Convergent evolution of intelligence in corvids and apes. Science 306, 1903–1907.

Enquist, M.E., Ghirlanda, S., 2005. Neural Networks and Animal Behavior. Princeton University Press, Princeton, NJ.

Erev, I., Roth, A.E., 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., Hamblin, S., Giraldeau, L.-A., 2013. Exposing the behavioral gambit: the evolution of learning and decision rules. Behavioral Ecology 24, 2–11.

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

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

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

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

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

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

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

Grimmett, G.R., Stirzaker, D.R., 2001. Probability and Random Processes, third ed. Oxford University Press, Oxford.

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

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

Hammerstein, P., Stevens, J.R. (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.

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

Hirsch, M.W., Smale, S., Devaney, R.L., 2004. Differential Equations, Dynamical Systems, and an Introduction to Chaos. Academic Press, San Diego, CA.

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

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

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge, MA.

Hofbauer, J., Sigmund, K., 2003. Evolutionary game dynamics. Bulletin of the American Mathematical Society 40, 479–519.

Hollis, K.L., Dumas, M.J., Singh, P., Fackelman, P., 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.I., 1983. Comments on "Learning the evolutionarily stable strategy". Journal of Theoretical Biology 105, 175–178.

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

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

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

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

Kushner, H.J., Yin, G.G., 2003. Stochastic Approximation and Recursive Algorithms and Applications, second ed. Springer, New York (NY).

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

Leslie, D.S., Collins, E.J., 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.

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

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

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

McElreath, R., 2010. Baryplot 1.0.

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

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

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

McNamara, J.M., Houston, A.I., 2009. Integrating function and mechanism. Trends in Ecology & Evolution 24, 670–675.

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.

Plotnik, J.M., Lair, R., Suphachoksahakun, W., De Waal, F.B.M., 2011. Elephants know when they need a helping trunk in a cooperative task. Proceedings of the National Academy of Sciences 108, 5116–5121.

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

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

Rescorla, R.A., Wagner, A.R., 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York (NY), pp. 64–99.

Rodriguez-Gironés, M.A., Vásquez, R.A., 1997. Density-dependent 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, Cambridge, MA.

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, New York (NY).

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

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

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. MIT Press, Cambridge, MA.

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

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

Tuyls, K., Verbeeck, K., Lenaerts, T., 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. ACM, Melbourne, Australia, pp. 693–700.

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

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

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

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

Walsh, P.T., Hansell, M., Borello, W.D., Healy, S.D., 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, Cambridge, MA.

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

Young, H.P., 2004. Strategic Learning and its Limits. Oxford University Press, Oxford.