Adaptive learning in large populations

Misha Perepelitsa

Received: 27 November 2018 / Revised: 26 June 2019 / Published online: 14 September 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
We consider the adaptive learning rule of Harley (J Theor Biol 89:611–633, 1981) for behavior selection in symmetric conflict games in large populations. This rule uses organisms’ past, accumulated rewards as the predictor for future behavior, and can be traced in many life forms from bacteria to humans. We derive a partial differential equation for the distribution of agents in the space of stimuli to select a particular strategy which describes the evolution of learning in heterogeneous populations. We analyze the solutions of the PDE model for symmetric 2 × 2 games. It is found that in games with small residual stimuli, adaptive learning rules with larger memory factor converge faster to the optimal outcome.

Keywords Adaptive learning · Relative payoff sum · Symmetric games

Mathematics Subject Classification 35Q91 · 91A05 · 91A20

1 Introduction
The seminal paper Maynard Smith and Price (1973) introduced the concepts of game theory into the study of animal behavior to explain the evolution of behavioral traits. To this end, a notion of evolutionarily stable strategy (ESS) was developed to describe stable outcomes of natural selection, by defining it as being uninvadable by pure or mixed strategies players. A dynamical process leading to an ESS can be formalized by the following model. Consider a situation where each individual in a population consistently uses one of the available behavioral traits $t_1, \ldots, t_k$ to interact with other members and no mutations take place. Assume the choice of the behavior (action) is inherited, and the population consists of groups that use a particular behavior. Over a long period of time involving large number of interactions, the average fitness per game for individuals using $t_i$, denoted by $W_i$, is compared to the population average
fitness $\bar{W}$, and the individuals in the $t_i$-group are reproduced at rate proportional to $W_i / \bar{W}$. If the changes in the frequencies of $t_i$ are approximately continuous Taylor and Jonker (1978), Zeeman (1981) derived the replicator dynamics equations for the frequencies $m_i$ of individuals acting according to $t_i$:

$$\frac{dm_i}{dt} = km_i(W_i - \bar{W}), \quad i = 1, \ldots, N$$  \hspace{1cm} (1)

and showed that ESS’s are the asymptotically stable fixed points of this system of equations.

Considered from the point of view of game theory, an evolutionarily stable strategy is a refinement of a Nash equilibrium, which describes an optimal choice of actions in games. This way, natural selection is a mechanism for implementing rational decision-making in the evolution of species. There is another way by which organisms, even without complex cognition, can discover optimal actions. It can be achieved through their ability to regulate behaviors depending on their experience, in particular, through the tendency to repeat positive and avoid negative experiences. This is known as the law of effect, first formulated by Thorndike (1898), and generally accepted as one of the main paradigms of animal behavior, see for example, Ferster and Skinner (1957), Herrstein (1961, 1970), Catania (1963), Chung and Herrstein (1967), Domjan and Burkhard (1984).

The law of effect is the basis for reinforcement learning models. They were introduced by Bush and Mosteller (1955), and since then have been applied to problems in such diverse areas as psychology, biology, economics, and engineering. Some representative examples of the extensive literature on this subject can be found in Luce (1959), Harley (1981), Cross (1983), Roth and Erev (1995), Erev and Roth (1998), Sutton and Barto (1998), Sandholm (2010), Nax and Perc (2015).

Interestingly enough, Borgers and Sarin (1997) demonstrated that models of learning also lead to the replicator dynamics equations similar to (1). For that, they considered repeated plays of a $2 \times 2$ game between two agents who adjust their probabilities for actions according to a reinforcement learning model of Cross (1983), and derived the replicator dynamics equation in the limit of small payoffs. Reinforcement learning in more general types of games and its relation to replicator-like equations were further discussed in Fudenberg and Levine (1998), Rustichini (1999), Fudenberg and Takahashi (2011), and Mertikopoulos and Sandholm (2016).

The law of effect can be expressed in many different ways, depending on which decision-making facilities are reinforced and on the specific rules of reinforcement. In the model of Cross (1983), it is the probability to play a particular action that undergoes reinforcement. In an alternative model motivated by bio-chemical processes in neural circuits, Harley (1981) proposed the relative payoff sum (RPS) algorithm as a decision making mechanism. The RPS learning rule assumes the ability of an organism to maintain a record of cumulative rewards from previous experiences, which at epoch $t$ is given by vector $S(t) = (S_1(t), \ldots, S_k(t))$. $S(t)$ is the predictor for future behavior and can be interpreted as a vector of “motivations” or “stimuli” to engage in a corresponding action. From the current stimuli an agent computes the probability to play $t_i$:
If action $t_{i_0}$ is chosen and it brings payoff $P_{i_0}(t)$, which is assumed to be non-negative, then stimuli are updated according to the rule

$$S_{i_0}(t + 1) = (1 - \bar{\mu})S_{i_0}(t) + \bar{\mu}r_{i_0} + P_{i_0}(t),$$

and for $k \neq i_0$,

$$S_k(t + 1) = (1 - \bar{\mu})S_k(t) + \bar{\mu}r_k.$$  

Positive parameter $\bar{\mu}$ expresses a memory effect: payoff $P_i(t - n)$ from $n$ previous plays will appear with the weight $(1 - \bar{\mu})^{n+1}$ in the expression for $S_i(t + 1)$. Parameter $r_i$ is some default level (residual) of stimuli $S_i$. For example, the residuals might represent the genetic preference for the action. A similar learning model was introduced by Roth and Erev (1995).

The goal of this paper is to address the question of behavior of RPS learning agents in large populations, where agents are randomly matched in pairwise encounters, i.e. learning in heterogeneous populations. Unlike the evolutionary dynamics framework, RPS learning agents cannot be identified with some particular strategy they use all the time. The RPS rule will in general prescribe a new strategy every time an agent plays.

In these situations, the natural quantity to describe the state of agents is the distribution of agents according to their current stimuli. In large populations, the probability density function of this distribution can be approximated by a continuous function and its changes can be described by a non-linear Fokker–Planck equation. The derivation of this equation, which generalizes the replicator dynamics equation for heterogeneous populations, is contained in the “Appendix” of this paper. The assumptions needed for the derivation of the equation are: large population size, large number of plays of the game, and incremental (infinitesimal) structure payoffs. This approach is a well-known method for modeling multi-agent systems in the problems of physics, biology, economics and sociology, see for example, Risken (1992) and Pareschi and Toscani (2014). A similar method has been used by Traulsen et al. (2005, 2006), in the analysis of the evolutionary selection by the Moran process.

After deriving the equations we apply them to determine the behavior of RPS learners in symmetric $2 \times 2$ games. It should be kept in mind, however, that the time asymptotic behavior based on the Fokker–Planck equation and the time asymptotic of the original discrete-time stochastic process are not, in general, the same. Moreover, the PDE model derived in this paper is a leading approximation of a continuous-time stochastic process. Thus, any statement claimed in this paper about the convergence of the system to a particular state should be understood as a statement that the system gets close to this state within the limits of the validity of the PDE approximation.

In game theory, the method of stochastic approximation by Benaim and Hirsch (1999) is typically used to determine the long-time behavior of interacting agents. The method relies on the stability analysis of an averaged (deterministic) system of ODEs. As the dimension of a system is of the order (the number of agents $\times$ (the number of
The analysis of the PDE model derived in this paper shows that for games with a single Nash equilibrium, the strategies of all agents converge to the dominant strategy, when the RPS rule has no memory factor, or with a memory factor and zero residuals. For learning models with a memory factor the convergence is faster than for the models with perfect memory. Additionally, the learning time in the former case varies inversely with the size of the memory factor. For games with a mixed Nash equilibrium, the learning process converges to a state in which the population mean probabilities equal the equilibrium values. As the population mean probability approaches its limit, the “strength” of learning decreases, and the individual probabilities do not, in general, converge to the equilibrium value. In this case, the population remains heterogeneous. Finally, if the memory factor is present and residuals are not zeros, an agent’s strategies converge to some mixed strategy, for a generic $2 \times 2$ symmetric game.

The problems with more than two pure strategies are of the great interest. However, the corresponding PDE model cannot be easily solved by hand and one must rely, instead, on the numerical solutions. The analysis of such problems is postponed for future work.

## 2 The model

We consider a series of plays of a game between randomly selected individuals in a large population. The payoff matrix of the game is given in Table 1. The behaviors are labeled A and B. We will analyze the RPS learning in the games that a) have a single pure Nash equilibrium with $a > c, d > b$ (or $a < c, d < b$); b) two Nash equilibria with $a > c, b > d$; and c) a mixed Nash equilibrium when $a < c, d > b$. In the latter case, there are also two asymmetric Nash equilibria $(A, B)$ and $(B, A)$.

Now, let there be a group of $N$ individuals where each individual is characterized by vector $X^t_i = (S^t_{i,1}, S^t_{i,2})$ representing the accumulated stimuli to play A and B, respectively, at epoch $t$. Suppose that two agents $i$ and $j$ are selected at random to play the game. Agents play with the probabilities to cooperate $S^t_{1,i}/(S^t_{1,i} + S^t_{2,i})$ and $S^t_{1,j}/(S^t_{1,j} + S^t_{2,j})$, for agent $i$ and agent $j$, respectively. Denote the outcomes $(A,A), (A,B), (B,A), (A,A)$, where the first is the action chosen by agent $i$. As the result of the interaction agent $i$ increments his/her states according to the rule

| Table 1 | The incremental payoff matrix for the game |
|---------|-----------------|
|         | A               | B               |
| A       | (ah, ah)        | (dh, ch)        |
| B       | (ch, dh)        | (bh, bh)        |

Parameters $a, b, c, d, h$ are positive.
Our main interest is in the distribution of agents in the stimuli-space \( x = (s_1, s_2) \), \( s_1, s_2 \geq 0 \), described by the density function (PDF) \( f(x, t) \). In this space the straight lines through the origin represent the sets of stimuli of constant probability to cooperate when an agent is playing a mixed strategy. The probability \( m \) is related to slope \( k \) by \( m = (1 + k)^{-1} \), so that the stimuli with preference for C are located closer to the \( s_1 \)-axis. For any subset \( \Omega \) in the stimuli space, \( \int_{\Omega} f(x, t) \, dx \) represents the proportion of agents with their stimuli in the set \( \Omega \) at time \( t \).

The following equation is found to be a leading order approximation for the process, when parameters \( h, \delta \) are small and the number of players is large;

\[
\frac{\partial f}{\partial t} + \text{div}(u(x, t) f) = 0, \tag{5}
\]

and velocity \( u \)

\[
u(x, t) = \frac{1}{s_1 + s_2} \left[ (a\bar{m}(t) + d(1 - \bar{m}(t))) s_1 + \mu(r_1 - s_1)(s_1 + s_2) \right]. \tag{6}
\]

where

\[
\bar{m}(t) = \int \frac{s_1}{s_1 + s_2} f(x, t) \, dx.
\]

In a special case of \( \mu = 0 \), the analysis of the equation can be reduced to the behavior of the system of ODEs:

\[
\frac{dx}{dt} = u(x, t), \tag{7}
\]

and the equation for \( \bar{m}(t) \) that follows from (5):

\[
\frac{d\bar{m}}{dt} = ((a - c)\bar{m} + (d - b)(1 - \bar{m})) \int \frac{s_1 s_2}{(s_1 + s_2)^2} f(x, t) \, dx. \tag{8}
\]

The derivation of Eqs. (5)–(8) is contained in the “Appendix”.
Velocity $u$ represents the rates of change of the stimuli of agents whose current state is given by $x$. The rates are proportional to the group average payoffs for corresponding actions, and “penalized” by memory for large deviations from the default residual levels $(r_1, r_2)$. A more convenient equation for analyzing the games with pure Nash equilibrium is (8). For the game with a mixed equilibrium, described by the frequency $m_*$ to play A, more convenient is equation

$$
\frac{d(\bar{m} - m_*)}{dt} = (c - a + d - b)(\bar{m} - m_*) \int \frac{s_1 s_2}{(s_1 + s_2)^2} f(x, t) \, dx.
$$

The nonlinear equation (5) is the first order approximation of the stochastic, learning process. The next order approximation contains a diffusion term, with the diffusion coefficients of order $h$. The diffusion generally prevents the convergence of learning of the group to a single strategy (fixation). For example, the convergence of the group learning to a single strategy, based on Eq. (5), only indicates that the distribution of agents’ strategies gets close to that particular strategy, within the limits of validity of Eq. (5).

Now we consider the dynamics of learning in symmetric $2 \times 2$ games. First we consider models with no memory factor, $\mu = 0$.

### 2.2 Pure Nash equilibrium

Let $a > c$ and $d > b$. The characteristic property of this regime is the positive sign of $\bar{m}'(t)$ in Eq. (8), for any distribution function $f$. This reflects the fact that action C is your best choice, no matter what your opponent does.

It is shown in “Appendix” that $\bar{m}(t)$ increases to its maximum value 1 and the support of function $f$ is transported to infinity along the trajectories of ODE (7). In particular, for large values of $t$, the stimuli of all agents will be located below any straight line of positive slope through the origin. That is, asymptotically all agents learn to play the equilibrium strategy “always A”. The inclusion of the second-order effects does not change the asymptotic picture.

### 2.3 Mixed Nash equilibrium

Let $a < c$ and $d > b$. The equilibrium mixed strategy is $m_* = \frac{d - b}{c - a + d - b}$ and the group average probability to play A evolves according to Eq. (9).

It is shown in the “Appendix” that the dynamics of Eq. (5) implies that $\bar{m}(t)$ converges to the equilibrium density $m_*$. The population average probability to play A asymptotically coincides with the probability $m_*$ at the Nash equilibrium. Unlike the pure Nash equilibrium case, in general, agents keep playing with different strategies. This can be seen from the following example. If the initial data $f_0(x)$ describes the population of agents playing different strategies and is such that

$$
\int \frac{s_1}{s_1 + s_2} f_0(x) \, dx = m_*,
$$
then the trajectories of the flow generated by $u$ are straight lines and for any $t > 0$,

$$\int \frac{s_1}{s_1 + s_2} f(x, t) \, dx = m_*.$$

Clearly, for all $t$, there is non-diminishing spread in the distribution function $f(x, t)$. To put it differently, there is no learning when the group average probability to play A equals $m_*$, because players' expected payoffs from A and B are equal.

### 2.4 Two Nash equilibria

For this type of game $a > c$ and $d < b$. We change the notation for $m_*$ to $\frac{b-c}{a-c+b-d}$. As in the previous case, the group average probability to play A evolves according to Eq. (9). If initially $\bar{m}(0) > m_*$, then the velocity carries the support of $f$ to a region with values of $s_1$ much larger than $s_2$. As the average $\bar{m}(t)$ increases, the dynamics is consistent and leads to learning of the Nash equilibrium A. With $\bar{m}(0) < m_*$ learning converges to the other equilibrium. The borderline case $\bar{m}(0) = m_*$, is unstable: stimuli increase along the straight lines, and agents retain their initial probabilities to play A and B, but any perturbations will deviate the system to A or the B equilibrium.

### 2.5 Memory factor RPS

Consider now models with $\mu > 0$. It can be seen from Eq. (6) that a sufficiently large box $[0, \hat{s}_1] \times [0, \hat{s}_2]$ is invariant under the flow of (7). This can be seen from the sign of the velocity components. We show in the “Appendix” that with residuals $r_1, r_2 > 0$, RPS learning will approach an asymptotically stable point in the stimuli space, for any positive payoff rates $a, b, c, d$. All agents will tend to play a mixed strategy $m_* = \frac{s_1}{s_1 + s_2}$. When (A,A) is a Nash equilibrium, the ratio of stimuli $s_1^*/s_2^* > r_1/r_2$, and agents favor action A after learning more than at their default levels.

There is an interesting limiting case of zero residuals $r_1 = r_2 = 0$. For such RPS models, when the game has (A,A) as the single Nash equilibrium, all agents learn this optimal strategy, as function $f(x, t)$ converges to a delta mass supported at the point $(a/\mu, 0)$.

In this case, the equation for mean $\bar{m}(t)$ is closely approximated by the replicator dynamics equation (11) given below. The factor $(s_1(t) + s_2(t))^{-1}$ on the right-hand side of that equation, in this case, is of order 1. The convergence to the optimal strategy is faster than in the case of learning with perfect memory ($\mu = 0$), for which the factor is of the order $(1 + t)^{-1}$.

Moreover, among the models with a memory factor, the learning period appears to be shorter for larger values of $\mu$, see Fig. 1 and explanations in the “Appendix”.

Thus it appears that RPS models with small residuals and large values of $\mu$ should be preferred by natural selection. In such models agents act predominantly on the basis of the last few payoffs. In this context it is worth mentioning that one of the postulates of prospect theory of Kahneman and Tversky (1984), is the statement that peoples’ actions (in games with monetary payoffs) are directed by the increase in their total
wealth, rather than the total accumulated wealth, which shows a tendency to use short memory and suggests that RPS learning might be at work.

2.6 Relation to the replicator dynamics equation (RDE)

If one postulates that all agents have the same, or approximately the same, stimuli

$$X_i^t = (s_1(t), s_2(t)), \quad \forall i = 1, \ldots, N,$$

(10)

for some $s_1(t), s_2(t)$, so that $f(s, t)$ is represented by a delta function supported at $(s_1(t), s_2(t))$, then Eq. (5) leads to the replicator dynamics equation for the probability to cooperate $m(t) = s_1(t) / (s_1(t) + s_2(t))$:

$$\frac{dm}{dt} = \frac{1}{s_1(t) + s_2(t)} \frac{m(1-m)}{s_1(t) + s_2(t)} ((a-c)m + (d-b)(1-m)).$$

(11)

Notice the positive factor on the right-hand side of the equation. For a learning processes in which stimuli increase the learning rate slows down. The extent to which hypothesis (10) is consistent with the dynamics of (7) is limited only to the cases when the latter has a single asymptotically stable fixed point.

Acknowledgements The author wishes to thank the anonymous referees for patient reading of the manuscript and detailed comments that helped to improve it in so many ways.
Appendix: A PDE model

Fokker–Planck equation

Consider a group of $N$ individuals acting according to the RPS learning rule described in Sect. 2. Let $X^i = (X^i_1, \ldots, X^i_N)$ represent the vector of pairs of stimuli for all the members at epoch $t$. Each component of this vector is 2-dimensional: $X^i = (s^i_1, s^i_2)$. By $w_h(x, t)$, where $x \in [0, 1]^2N$, we denote the PDF for the distribution of $X^i$. We will write $\bar{x} = (x_1, \ldots, x_N)$, where each $x_i = (s^i_1, s^i_2)$. The probability to play A will be denoted as $\lambda_i = s^i_1 / (s^i_1 + s^i_2)$.

Suppose that member $i$ and $j$ are selected for the interaction. There will be only one game played during the period from $t$ to $t + \delta$. The matrix of payoffs is described in Table 1. The range of parameters $\delta$, $h$, $N$ will be restricted later on.

Conditioned on the event $X^t = \bar{x}$, the agent probabilities for the next period are set according to the RPS rule (4), which in the notation of the stochastic process are

$$X^t_{i+\delta} = \begin{cases} (1-\mu h)s^i_1 + r_1 \mu h + ah, (1-\mu h)s^i_2 + r_2 \mu h) \text{ Prob} = \lambda_i \lambda_j \\ (1-\mu h)s^i_1 + r_1 \mu h + dh, (1-\mu h)s^i_2 + r_2 \mu h) \text{ Prob} = \lambda_i (1-\lambda_j) \\ (1-\mu h)s^i_1 + r_1 \mu h, (1-\mu h)s^i_2 + r_2 \mu h + ch) \text{ Prob} = (1-\lambda_i) \lambda_j \\ (1-\mu h)s^i_1 + r_1 \mu h + bh, (1-\mu h)s^i_2 + r_2 \mu h + bh) \text{ Prob} = (1-\lambda_i) (1-\lambda_j) \end{cases}$$

and symmetrically for $X^t_{j+\delta}$. For all other agents, $X^t_{k+\delta} = X^t_k$ for $k \neq i, j$. The definition of $X^t$ makes it a discrete-time Markov process. We proceed by writing down the integral form of the Chapman–Kolmogorov equations and approximate its solution by a solution of the Fokker–Planck equation (forward Kolmogorov’s equation), for small values of $\delta$, $h$ and large $N$.

Change of $w_h(\bar{x}, t)$ from $t$ to $t + \delta$, can be described in the following way.

$$\int \phi(\bar{x}) w_h(\bar{x}, t + \delta) \, d\bar{x} = \mathbb{E}[\phi(X^{t+\delta})]$$

$$= \sum_{i \neq j} (N(N-1))^{-1} \int \begin{pmatrix} \lambda_i \lambda_j \phi(\bar{x}) \\ \lambda_i (1-\lambda_j) \phi(\bar{x}) \\ (1-\lambda_i) \lambda_j \phi(\bar{x}) \\ (1-\lambda_i) (1-\lambda_j) \phi(\bar{x}) \end{pmatrix} \begin{pmatrix} x_i = (1-\mu h)s^i_1 + r_1 \mu h + ah, (1-\mu h)s^i_2 + r_2 \mu h \\ x_j = (1-\mu h)s^i_1 + r_1 \mu h + dh, (1-\mu h)s^i_2 + r_2 \mu h \\ x_i = (1-\mu h)s^i_1 + r_1 \mu h, (1-\mu h)s^i_2 + r_2 \mu h + ch \\ x_j = (1-\mu h)s^i_1 + r_1 \mu h + bh, (1-\mu h)s^i_2 + r_2 \mu h + bh \end{pmatrix} w_h(\bar{x}, t) \, d\bar{x}.$$  

This equation can be written in a slightly different way:
\[
\int \phi(\bar{x}) w_h(\bar{x} , t + \delta) \, d\bar{x} = \int \phi(\bar{x}) w_h(x , t) \, d\bar{x} \\
+ \sum_{i \neq j} (N(N - 1))^{-1} \\
\int \left( \lambda_i \lambda_j \left[ \phi(\bar{x}) \bigg| x_i = ((1 - \mu) x_{i1} + r_1 \mu h + ah, (1 - \mu) x_{i2} + r_2 \mu h) - \phi(\bar{x}) \bigg| x_j = ((1 - \mu) x_{j1} + r_1 \mu h + ah, (1 - \mu) x_{j2} + r_2 \mu h) \right] \\
+ \lambda_i (1 - \lambda_j) \left[ \phi(\bar{x}) \bigg| x_i = ((1 - \mu) x_{i1} + r_1 \mu h + dh, (1 - \mu) x_{i2} + r_2 \mu h + ch) - \phi(\bar{x}) \bigg| x_j = ((1 - \mu) x_{j1} + r_1 \mu h, (1 - \mu) x_{j2} + r_2 \mu h + ch) \right] \\
+ (1 - \lambda_i) \lambda_j \left[ \phi(\bar{x}) \bigg| x_i = ((1 - \mu) x_{i1} + r_1 \mu h, (1 - \mu) x_{i2} + r_2 \mu h + ch) - \phi(\bar{x}) \bigg| x_j = ((1 - \mu) x_{j1} + r_1 \mu h + dh, (1 - \mu) x_{j2} + r_2 \mu h) \right] \\
+ (1 - \lambda_i) (1 - \lambda_j) \right[ \phi(\bar{x}) \bigg| x_i = ((1 - \mu) x_{i1} + r_1 \mu h, (1 - \mu) x_{i2} + r_2 \mu h + bh) - \phi(\bar{x}) \bigg| x_j = ((1 - \mu) x_{j1} + r_1 \mu h + rh, (1 - \mu) x_{j2} + r_2 \mu h + bh) \right] \right) w_h(\bar{x} , t) \, d\bar{x}.
\]

(13)

The above equation can be used to obtain a 2N-dimensional ODE approximation of the stochastic process by evaluating \( \lim_{h \to 0} \left( \frac{\mathbb{E} [X^{t+h} | X^t] - X^t}{h} \right) = F(X^t) \). This approach was implemented in the method of stochastic approximation developed by Benaim and Hirsch (1999) and applied to the study of convergence of stochastic fictitious play processes. The method guarantees the convergence of the process \( X^t \) under certain stability conditions for the dynamics of the associated ODE.

The large dimension of that dynamical system is an obstacle for further analysis. In contrast, we would like to obtain an equation for the distribution of a large number \( N \) of agents in a 2-dimensional stimuli space. For this, denote the PDF of the distribution by

\[
f_h(x, t) = \sum_k N^{-1} \int w_h(\bar{x}) \bigg| x_k = x_k d\bar{x}, \quad x \in \mathbb{R}^2,
\]

where \( \bar{x}_k \) is a 2N - 2 dimensional vector of all coordinates, excluding \( x_k \). In statistical physics this function is also called a one-particle distribution. In the formulas to follow we need to use a two-particle distribution function

\[
g_h(x, y, t) = \sum_{i \neq j} (N(N - 1))^{-1} \int w_h(\bar{x}) \bigg| x_i = x, x_j = y d\bar{x}_{ij},
\]

where \( \bar{x}_{ij} \) is the 2N - 4 dimensional vector of all coordinates excluding \( x_i \) and \( x_j \). Function \( g_h \) is symmetric in \( (x, y) \) and is related to \( f_h \) by the formulas

\[
f_h(x, t) = \int g_h(x, y, t) \, dx = \int g_h(x, y, t) \, dy.
\]
The moments of function $f_h$ and $g_h$ are computed from the moment of $w_h$:

$$\int \psi(x) f_h(x, t) \, dx = \sum_k N^{-1} \int \psi(x_k) w_h(\bar{x}) \, d\bar{x},$$

and

$$\int \omega(x, y) g_h(x, y, t) \, dx \, dy = \sum_{i \neq j} (N(N - 1))^{-1} \int \omega(x_i, x_j) w_h(\bar{x}) \, d\bar{x}.$$

This follows from the definition of these functions.

Now we use (13) to obtain an integral equation of the change of function $f_h$. For that select $\phi(\bar{x}) = \psi(x_k)$, sum over $k$ and take average. We get

$$\int \psi(x) f_h(x, t + \delta) \, dx = \int \psi(x) f_h(x, t) \, dx + N^{-1} \sum_{i \neq j} (N(N - 1))^{-1}
\int \left( \lambda_i \lambda_j \left[ \psi((1 - \mu) s_{1,1} + r_1 \mu h + ah, (1 - \mu) s_{2,i} + r_2 \mu h) - \psi(x_i) \right]
+ \psi((1 - \mu) s_{1,j} + r_1 \mu h + ah, (1 - \mu) s_{2,j} + r_2 \mu h) - \psi(x_j) \right]
+ \lambda_i (1 - \lambda_j) \left[ \psi((1 - \mu) s_{1,1} + r_1 \mu h + dh, (1 - \mu) s_{2,i} + r_2 \mu h) - \psi(x_i) \right]
+ \psi((1 - \mu) s_{1,j} + r_1 \mu h + dh, (1 - \mu) s_{2,j} + r_2 \mu h + ch) - \psi(x_j) \right]
+ (1 - \lambda_i) \lambda_j \left[ \psi((1 - \mu) s_{1,1} + r_1 \mu h, (1 - \mu) s_{2,i} + r_2 \mu h + ch) - \psi(x_i) \right]
+ \psi((1 - \mu) s_{1,j} + r_1 \mu h, (1 - \mu) s_{2,j} + r_2 \mu h) - \psi(x_j) \right]
+ \lambda_i (1 - \lambda_j) \left[ \psi((1 - \mu) s_{1,1} + r_1 \mu h, (1 - \mu) s_{2,i} + r_2 \mu h + bh) - \psi(x_i) \right]
+ \psi((1 - \mu) s_{1,j} + r_1 \mu h, (1 - \mu) s_{2,j} + r_2 \mu h + bh) - \psi(x_j) \right) \, w_h(\bar{x}, t) \, d\bar{x}.$$ (14)

The right-hand side can be conveniently expressed in terms of a two-particle function $gh$:

$$\int \psi(x) f_h(x, t + \delta) \, dx = \int \psi(x) f_h(x, t) \, dx
+ 2N^{-1} \int \left( \lambda(x) \lambda(y) \left[ \psi((1 - \mu) s_{1,1} + r_1 \mu h + ah, (1 - \mu) s_{2,x} + r_2 \mu h) - \psi(x) \right]
+ \psi((1 - \mu) s_{1,y} + r_1 \mu h + ah, (1 - \mu) s_{2,y} + r_2 \mu h) - \psi(y) \right) \, \, d\bar{x}.$$
\[ + \lambda(x)(1 - \lambda(y))[\psi((1 - \mu h)s^x_1 + r_1\mu h + dh, \\
(1 - \mu h)s^x_1 + r_2\mu h) - \psi(x) \\
+ \psi((1 - \mu h)s^y_1 + r_1\mu h, \\
(1 - \mu h)s^y_2 + r_2\mu h + ch) - \psi(y)] + (1 - \lambda(x))\lambda(y)[\psi((1 - \mu h)s^x_1 + r_1\mu h, \\
(1 - \mu h)s^y_2 + r_2\mu h + ch) - \psi(x) \\
+ \psi((1 - \mu h)s^y_1 + r_1\mu h + dh, \\
(1 - \mu h)s^y_2 + r_2\mu h) - \psi(y)] + (1 - \lambda(x))(1 - \lambda(y))[\psi((1 - \mu h)s^x_1 + r_1\mu h, \\
(1 - \mu h)s^y_2 + r_2\mu h + bh) - \psi(x) \\
+ \psi((1 - \mu h)s^y_1 + r_1\mu h, \\
(1 - \mu h)s^y_2 + r_2\mu h + bh) - \psi(y)] \} g_h(x, y, t) \, dx \, dy. \quad (15) \]

where \( x = (s^x_1, s^x_2), \ y = (s^y_1, s^y_2), \) and \( \lambda(x) = s^x_1/(s^x_1 + s^x_2), \) and similar for \( \lambda(y). \)

In the processes with large number of agents and random binary interactions, the two-particle distribution function can be factored into two independent distributions:

\[ g_h(x, y, t) = f_h(x, t) f_h(y, t). \]

With this relation, (15) becomes a family of non-linear integral relations for the next time step distribution \( f_h(x, t + \delta). \) Taking the Taylor expansions up to the first order for the increment of the test function \( \psi, \) we obtain integral equations:

\[
\frac{N}{2h} \int \psi(x)(f_h(x, t + \delta) - f_h(x, t)) \, dx \\
= 2\tilde{m}_h(t) \int (\lambda(x)(a + \mu(r_1 - s_1))\partial_{s_1} \psi(x) \\
+ \lambda(x)\mu(r_2 - s_2)\partial_{s_2} \psi(x)) f_h(x, t) \, dx \\
+ 2(1 - \tilde{m}_h(t)) \int (\lambda(x)(d + \mu(r_1 - s_1))\partial_{s_1} \psi(x) \\
+ \lambda(x)\mu(r_2 - s_2)\partial_{s_2} \psi(x)) f_h(x, t) \, dx \\
+ 2\tilde{m}_h(t) \int ((1 - \lambda(x))(\mu(r_1 - s_1))\partial_{s_1} \psi(x) \\
+ (1 - \lambda(x))(c + \mu(r_2 - s_2))\partial_{s_2} \psi(x)) f_h(x, t) \, dx \\
+ 2(1 - \tilde{m}_h(t)) \int ((1 - \lambda(x))(\mu(r_1 - s_1))\partial_{s_1} \psi(x) \\
+ (1 - \lambda(x))(b + \mu(r_2 - s_2))\partial_{s_2} \psi(x)) f_h(x, t) \, dx
\]

where \( \tilde{m}_h(t) = \int \lambda(x) f_h(x, t) \, dx. \)
Combining various terms on the right-hand side of the equation we get

\[
\frac{N\delta}{4h} \int \psi(x) \left( \frac{f_h(x, t + \delta) - f_h(x, t)}{\delta} \right) \, dx = \int \left( v_1 \partial_s_1 \psi + v_2 \partial_s_2 \psi \right) f_h(x, t) \, dx, 
\]

with

\[
v_1 = \lambda(x)(\bar{m}_h(t)(a\bar{m}(t) + (1 - \bar{m}_h(t))d) + \mu(r_1 - s_1),
\]

\[
v_2 = (1 - \lambda(x))(c\bar{m}_h(t) + (1 - \bar{m}_h(t))b) + \mu(r_2 - s_2).
\]

By integrating the right-hand side by parts, and assuming that \(h, \delta\) are small and \(N\) is large, in such a way that \(4h\delta N = 1\), we obtain the integral relation:

\[
\int \psi(x) \left( \frac{\partial f}{\partial t} + \text{div}(u(x, t) f) \right) \, dx = 0.
\]

Since the test function \(\psi\) is an arbitrary, the Fokker–Planck equation follows:

\[
\frac{\partial f}{\partial t} + \text{div}(u(x, t) f) = 0, \tag{16}
\]

where \(x = (s_1, s_2), s_1, s_2 > 0\) and the drift velocity is given by the formula

\[
u(x, t) = \frac{1}{s_1 + s_2} \left[ \begin{array}{c}
(a\bar{m}(t) + d(1 - \bar{m}(t)) s_1 + \mu(r_1 - s_1)(s_1 + s_2)) \\
(c\bar{m}(t) + b(1 - \bar{m}(t)) s_2 + \mu(r_2 - s_2)(s_1 + s_2))
\end{array} \right],
\]

with

\[
\bar{m}(t) = \int \frac{s_1}{s_1 + s_2} f(x, t) \, dx. \tag{17}
\]

Equation (8) is obtained from (16) by multiplying it by \(\lambda(x) = s_1/(s_1 + s_2)\), and integrating by parts:

\[
\frac{d}{dt} \int \frac{s_1}{s_1 + s_2} f(x, t) \, dx = \int \left( u_1(x, t) \frac{s_2}{(s_1 + s_2)^2} - u_2(x, t) \frac{s_1}{(s_1 + s_2)^2} \right) f(x, t) \, dx.
\]

Here we are assuming that the support of \(f\) is contained in the interior of the first quadrant, so that \(f\) is zero on the boundary. That is to say that all agents play mixed strategies. This is a natural hypothesis, as nothing else can be learned if an agent chooses an action with certainty.
Substituting formulas for $u_1$ and $u_2$ from the previous page (with $\mu = 0$) we obtain (8).

Consider now learning from playing a symmetric $n \times n$ game. Let the payoff to playing the $i$th action against the $j$th be equal to $a_{ij} h$. Denote the stimulus vector $x = (s_1, \ldots, s_n)$, and the population average probability to play $i$ by

$$\tilde{m}_i(t) = \int \frac{s_i}{\sum_j s_j} f(x, t) \, dx, \quad i = 1, \ldots, n.$$ 

The first-order approximation of the RPS learning process is given by the Fokker–Planck equation

$$\frac{\partial f}{\partial t} + \text{div}(u(x, t) f) = 0,$$

on the domain with $s_i \geq 0, \ i = 1, \ldots, n$. In this equation, the velocity vector $u = (u_1, \ldots, u_n)$ is given by its components

$$u_i(x, t) = \frac{1}{\sum_j s_j} \left[ \left( \sum_j a_{ij} \tilde{m}_j(t) \right) s_i + \mu (r_i - s_i) \sum_j s_j \right], \quad i = 1, \ldots, n.$$ 

Now we consider in some detail learning in $2 \times 2$ games. Much of the analysis of Eq. (16) is derived from the behavior of the trajectories of ODE (7). The solution $f$ of (16) is obtained by transporting the support of $f_0$ along trajectories of the dynamical system (7) and changing the values $f_0$ so that the “mass” (measured by the density function $f$) of any “fluid element” remains constant. In fact one can write the formula for $f$ in terms of $u$ and prove that the solution $f$ of the non-linear problem exists and is unique. This can be done by standard methods of PDEs, but it is outside of the scope of the present paper. Here, we will be interested in the long time, qualitative asymptotic for $f(x, t)$.

Equation (16) is considered in the first quadrant $s_1, s_2 > 0$. For $\mu = 0$, the boundary of the domain is invariant under the flow of (7). For the model with fading memory, $\mu > 0$, the velocity $u$, at the boundary, is directed into the flow domain. In either case, we will assume that the function $f_0(x)$ is zero on the boundary. Then, this property will hold for all $t > 0$. Additionally, in all of the analysis below, we assume that $f_0$ is a continuously differentiable function with compact support (zero outside some bounded set).

Consider the case of the pure Nash equilibrium $(a > c, d > b)$ and no memory effects, $\mu = 0$. The velocity is

$$u(x, t) = \frac{1}{s_1 + s_2} M(t)x,$$
where matrix

\[
M(t) = \begin{bmatrix}
    a\bar{m}(t) + d(1 - \bar{m}(t)) & 0 \\
    0 & c\bar{m}(t) + b(1 - \bar{m}(t))
\end{bmatrix}.
\]

For any \( t \), the origin is an unstable node with two positive eigenvalues; the eigenvalue corresponding to the \( s_1 \)-direction is the dominant one. From the phase portrait of the ODE it is clear that the flow transports the support of \( f \) into the region where \( s_1 \gg s_2 \), which correspond to the case of all agents asymptotically in \( t \) adopting choice \( C \) in the game.

In the case of the mixed Nash equilibrium \( (a < c, d > b) \) and no memory effect, \( \mu = 0 \), the origin is an unstable node. When \( \bar{m}(t) = m_* \) then two positive eigenvalues coincide, and all trajectories are straight lines through the origin. In general, however, \( \bar{m}(t) \neq m_* \), if for example they are not equal at time \( t = 0 \). In such cases Eq. (9) can be used to show that \( \lim_{t \to \infty} \bar{m}(t) = m_* \). Let \( \bar{m}(0) > m_* \). Then, according to (9), \( \bar{m}(0) > \bar{m}(t) > m_* \) for all \( t \), and \( \bar{m}(t) \) converges to \( m_* \) provided that

\[
\int_0^\infty \int \frac{s_1 s_2}{(s_1 + s_2)^3} f(x, t) \, dx \, dt
\]

diverges. Notice also, that the derivative of ratio \( s_2/s_1 \) along a flow trajectory equals

\[
\frac{d}{dt} \left( \frac{s_2}{s_1} \right) = (c - a + d - b)(\bar{m}(t) - m_*) \frac{s_1 s_2}{s_1 + s_2} > 0.
\]

Thus, if at \( t = 0 \) the support of \( f_0(\cdot) \) is strictly inside the first quadrant, then there is a constant \( c > 0 \) such that \( s_2/s_1 > c \) for all points in the support of \( f(\cdot, t) \) for all later times. In particular, one can estimate

\[
\int \frac{s_1 s_2}{(s_1 + s_2)^3} f(x, t) \, dx = \int \frac{s_2/s_1}{(1 + s_2/s_1)(s_1 + s_2)} \frac{s_1}{s_1 + s_2} f(x, t) \, dx
\]

\[
> \sup_{(s_1, s_2) \in \text{supp}(\cdot, t)} \frac{c}{(1 + c)(s_1 + s_2)} \bar{m}(t).
\]

Finally, since \( u(x, t) \) is uniformly bounded, i.e., for any \( (x, t) \), \( |u(x, t)| < C \), for some \( C \), then for any \( x \) in the support of \( f(\cdot, t) \) there a constant \( \tilde{C} \) such that \( |x| < \tilde{C}(1 + t) \). From this and (19) it follows that

\[
\int \frac{s_1 s_2}{(s_1 + s_2)^3} f(x, t) \, dx > C m_* (1 + t)^{-1},
\]

for some constant \( C > 0 \), and so, the integral in (18) is infinite. The case \( \bar{m}(0) < m_* \) follows by similar arguments.

Consider now the model with memory decay when \( \mu > 0 \) and residuals \( r_1, r_2 > 0 \). For any value of \( \bar{m}(t) \in [0, 1] \) and any set of positive parameters of the game \( a, b, c, d > 0 \) the right-hand side of (7) has a steady state \( (s_1^0(t), s_2^0(t)) \) in the interior
of the first quadrant, with \( s_1^0 > r_1, s_2^0 > r_2 \), and this point is an asymptotically stable node. The other steady state is the origin, which is an unstable node. The support of \( f(x, t) \) moves in the direction of the stable interior point, contracting in size. When it becomes sufficiently small, the dynamics can be effectively approximated by ODE:

\[
\frac{d(s_1, s_2)}{dt} = (\bar{u}_1, \bar{u}_2)
\]

where the new velocity is given by

\[
\bar{u}_1 = \frac{a(s_1)^2}{(s_1 + s_2)^2} + \frac{ds_1 s_2}{(s_1 + s_2)^2} + \mu(r_1 - s_1),
\]

\[
\bar{u}_2 = \frac{cs_1 s_2}{(s_1 + s_2)^2} + \frac{b s_2^2}{(s_1 + s_2)^2} + \mu(r_2 - s_2).
\]

In the long run the fixed point will settle at the stable, interior state \( x^* = (s_1^*, s_2^*) \), and \( f(x, t) \) will be a delta-function supported at that point. In this process the agents learn to play A with probability \( \bar{m}^* = s_1^*/(s_1^* + s_2^*) \).

A special case of zero residuals \( r_1 = r_2 = 0 \) deserves a discussion. In the limit of zero residuals \( r_1, r_2 \to 0 \), velocity \( u \) (for any fixed \( t > 0 \)) has three fixed points: \( x_0(t) = ((a \bar{m}(t) + d(1 - \bar{m}(t))/\mu, 0), x_1 = (0, (c \bar{m}(t) + b(1 - \bar{m}(t))/\mu) \) and \( (0, 0) \). When \( a > c, d > b \), the first, corresponding to the strategy “always A”, is an asymptotically stable node, the second, corresponding to “always B”, is a saddle, and the origin is an unstable node. One can compute that on any trajectory of the velocity field \( u(x, t) \) inside the first quadrant,

\[
\frac{d}{dt} \left( \frac{s_1}{s_2} \right) > 0.
\]

Thus, the population average probability to play A, \( \bar{m}(t) \), increases to 1, the stable stationary point \( x_0(t) \) converges to \( x_0 = (a/\mu, 0) \), and the support of \( f(x, t) \) moves toward point \( x_0 \). The agents with memory decay and zero residual levels do learn the optimal strategy. Moreover, because learning occurs in the bounded region of the stimuli space, convergence to the equilibrium is faster than the case of learning with perfect memory \( \mu = 0 \). Using Eq. (11) we can also estimate the rate of convergence as a function of \( \mu \). The rate is proportional to \( \mu \) implying that the characteristic time of convergence is \( C/\mu \). Figure 1 shows the simulations of the stochastic process in this regime for different values of \( \mu \). It shows that the prediction based on the model (16) is in good agreement with the stochastic learning process. We conclude that models with low residuals and high \( \mu \) perform better in learning the optimal strategy in \( 2 \times 2 \) games.

Second order effects

The inclusion of the next order approximation into consideration adds a diffusion term into Eq. (16) with the diffusion coefficients proportional to \( h \). In the problems where the drift velocity takes \( f \) to regions with a large values of \( x \), as in the case of the pure or mixed Nash equilibrium, diffusion will have a marginal effect. In the problems with asymptotically stable fixed points in stimuli space (short memory models) diffusion
will create a stationary distribution of \( f \) near the fixed point, preventing all agents to adopt a single strategy.

References

Benaim M, Hirsch MW (1999) Stochastic approximation algorithms with constant step size whose average is cooperative. Ann Appl Probab 9(1):216–241

Borgers T, Sarin R (1997) Learning through reinforcement and replicator dynamics. J Econ Theory 77(1):1–14

Bush RR, Mosteller F (1955) Stochastic models for learning. Wiley, New York

Catania AC (1963) Concurrent performances. A baseline for the study of reinforcement magnitude. J Exp Anal Behav 6:299–300

Chung S-H, Herrstein RJ (1967) Choice and delay of reinforcement. J Exp Anal Behav 10:67–74

Cross JG (1983) A theory of adaptive economic behavior. Cambridge University Press, Cambridge

Domjan M, Burkhard B (1984) The principles of learning and behavior. Brooks/Cole Publishing Co., Monterey

Erev I, Roth AE (1998) Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibrium. Am Econ Rev 88:848–881

Ferster CB, Skinner BF (1957) Schedules of reinforcement. Appleton-Century-Crofts, New York

Fudenberg D, Levine D (1998) The theory of learning in games. MIT Press, Cambridge

Fudenberg D, Takahashi S (2011) Heterogeneous beliefs and local information in stochastic fictitious play. Games Econ Behav 71:100–120

Harley CB (1981) Learning the evolutionarily stable strategy. J Theor Biol 89:611–633

Herrstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4:267–272

Herrstein RJ (1970) On the law of effect. J Exp Anal Behav 13:243–266

Kahneman D, Tversky A (1984) Choices, values and frames. Am Psychol 39:341–350

Luce RD (1959) Individual choice behavior: a theoretical analysis. Wiley, New York

Maynard Smith J, Price GR (1973) The logic of animal conflict. Nat Lond 246:15–18

Mertikopoulos P, Sandholm WH (2016) Learning in games via reinforcement learning and regularization. Math Oper Res INFORMS 41(4):1297–1324

Nax HH, Perc M (2015) Directional learning and the provisioning of public goods. Sci Rep 5:8010

Pareschi L, Toscani G (2014) Interacting multiagent systems: kinetic equations and Monte Carlo methods. Oxford University Press, Oxford

Risken H (1992) The Fokker–Planck equation: methods of solution and applications. Springer, Berlin

Roth AE, Erev I (1995) Learning in extensive-form games: experimental data and simple dynamics models in the intermediate term. Games Econ Behav 8:164–212

Rustichini A (1999) Optimal properties of stimulus-response learning models. Games Econ Behav 29:230–244

Sandholm WH (2010) Population games and evolutionary dynamics. MIT Press, Cambridge

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Taylor PD, Jonker IN (1978) Evolutionarily stable strategies. Math Biosci 40:145–56

Thorndike EL (1898) Animal intelligence: an experimental study of the associative processes in animals. In: Baldwin JM, Cattell JM, with the cooperation of [other] (eds) Psychological review. Series of monograph supplements, vol II, no 4 (whole no 8). The Macmillan Company, New York, London

Traulsen A, Claussen JC, Hauert C (2005) Coevolutionary dynamics: from finite to infinite populations. Phys Rev Lett 95:238701

Traulsen A, Claussen JC, Hauert C (2006) Coevolutionary dynamics in large, but finite populations. Phys Rev E 74:011901

Zeeman EC (1981) Dynamics of the evolution of animal conflicts. J Theor Biol 89:249–70

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.