



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

Games and Economic Behavior ●●● (●●●●) ●●●●●●●●

---

**GAMES and  
Economic  
Behavior**


---

[www.elsevier.com/locate/geb](http://www.elsevier.com/locate/geb)

# Generalised weakened fictitious play

David S. Leslie<sup>a,1</sup>, E.J. Collins<sup>b,\*</sup>

<sup>a</sup> *Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK*

<sup>b</sup> *Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK*

Received 31 August 2004

---

## Abstract

A general class of adaptive processes in games is developed, which significantly generalises weakened fictitious play [Van der Genugten, B., 2000. A weakened form of fictitious play in two-person zero-sum games. *Int. Game Theory Rev.* 2, 307–328] and includes several interesting fictitious-play-like processes as special cases. The general model is rigorously analysed using the best response differential inclusion, and shown to converge in games with the fictitious play property. Furthermore, a new actor–critic process is introduced, in which the only information given to a player is the reward received as a result of selecting an action—a player need not even know they are playing a game. It is shown that this results in a generalised weakened fictitious play process, and can therefore be considered as a first step towards explaining how players might learn to play Nash equilibrium strategies without having any knowledge of the game, or even that they are playing a game.

© 2005 Elsevier Inc. All rights reserved.

*JEL classification:* C72; D83

*Keywords:* Fictitious play; Best response differential inclusion; Stochastic approximation; Actor–critic process

---

## 1. Introduction

As observed by Brown (1951), a fictitious play process will follow trajectories of the best response differential inclusion (Gilboa and Matsui, 1991; Cowan, 1992), with “mistakes” aris-

---

\* Corresponding author.

*E-mail addresses:* [dleslie@stats.ox.ac.uk](mailto:dleslie@stats.ox.ac.uk) (D.S. Leslie), [e.j.collins@bristol.ac.uk](mailto:e.j.collins@bristol.ac.uk) (E.J. Collins).

<sup>1</sup> Research partly carried out at the University of Bristol, supported by CASE Research Studentship 00317214 from the UK Engineering and Physical Sciences Research Council in cooperation with BAE SYSTEMS.

ing from the discretisation that eventually become negligible due to the decreasing step size. This method has recently been formalised by Benaïm et al. (2005), allowing analysis of the best response differential inclusion to provide convergence proofs for classical fictitious play, and obviating the need for the ingenious but case-specific techniques employed previously (Robinson, 1951; Miyasawa, 1961; Monderer and Shapley, 1996; and see Berger, 2005 for further references).

Weakened fictitious play (Van der Genugten, 2000) is identical to fictitious play, except for the fact that at each step the strategies played need only be  $\epsilon$ -best responses, with  $\epsilon \rightarrow 0$  as time progresses. Intuitively, since these “mistakes” vanish asymptotically, such processes should also follow the best response differential inclusion in the limit (regardless of the rate at which  $\epsilon \rightarrow 0$ , in contrast with Van der Genugten’s analysis). In this work we introduce a significant generalisation of weakened fictitious play and analyse it using the method of Benaïm et al. (2005). This unified approach to the analysis allows us to discuss several interesting variations of fictitious play and show that they will all converge to Nash equilibrium in games known to have the fictitious play property, simply by showing that they are in fact generalised weakened fictitious play processes. One such example is a fictitious play process which places greater weight on recently observed actions than on actions observed in the distant past; this may be seen as compensating for the fact that opponent strategies change over time, as opposed to the implicit assumption in classical fictitious play that opponent strategies are stationary. A second example is stochastic fictitious play (Fudenberg and Kreps, 1993; Benaïm and Hirsch, 1999) for which the payoff perturbations become negligible as time proceeds.

A much more substantial application involves the introduction of a new actor–critic process in which the only information given to a player is the reward received as a result of selecting an action—a player need not even know they are playing a game. This process is also shown to result in a generalised weakened fictitious play process, and therefore to converge in the same games as fictitious play, despite using significantly less information; this can therefore be considered as a first step towards explaining how players might play Nash equilibrium strategies without having any knowledge of the game, or even that they are playing a game.

We note here that other analyses of “simple, payoff-based” learning (Foster and Young, 2003b) tend to focus on the convergence of the average action played, which does not preclude the possibility that strategies are very obviously cycling around equilibrium (Fudenberg and Levine, 1998). Furthermore, they either prove that these averages converge only to correlated equilibria (Hart and Mas-Colell, 2001), or prove convergence only in a very weak sense, essentially due to the fact that if strategies vary randomly then at some point they must get close to a Nash equilibrium (Foster and Young, 2003a, 2003b). In contrast, for our actor–critic process the actual strategies of the players converge for all of the games in which generalised weakened fictitious play is shown to converge.

Observe that Hart and Mas-Colell (2003) prove that no “uncoupled” dynamics can converge to equilibrium in general games. While the processes we consider in this paper are uncoupled (players do not consider any payoffs other than their own) we believe these are still worthy of study. Partly this is for historical interest—several interesting models of learning are combined into a single large class and analysed with a unified approach. In addition, if players are not aware of the fact that they are playing a game, it is less easy to escape the class of uncoupled processes, and so it is important to understand which classes of games may result in convergence to Nash equilibrium strategies under these minimal information requirements.

The paper is structured as follows. In the next section we introduce some notation and preliminary ideas, then in Section 3 the generalised weakened fictitious play model is introduced and analysed. Several interesting examples of the model are discussed in Section 4, and the new actor-critic process is proposed and studied in Section 5.

**2. Preliminaries**

We consider myopic boundedly-rational players in a repeated  $N$ -player normal-form game, in which Player  $i$  has finite pure strategy set  $A^i$ , mixed strategy set  $\Delta^i$  (so that  $\Delta^i$  is the set of probability distributions over  $A^i$ ), and bounded reward function  $r^i : \times_{i=1}^N \Delta^i \rightarrow \mathbb{R}$ . We will write  $r^i(a^i, \pi^{-i})$  (respectively  $r^i(\pi^i, \pi^{-i})$ ) for the expected reward to Player  $i$  if they select pure strategy  $a^i$  (respectively mixed strategy  $\pi^i$ ) and all other players play the mixed strategy profile  $\pi^{-i} = (\pi^1, \dots, \pi^{i-1}, \pi^{i+1}, \dots, \pi^N)$ ; also write

$$b^i(\pi^{-i}) = \operatorname{argmax}_{\pi^i \in \Delta^i} r^i(\pi^i, \pi^{-i}) \quad \text{and} \quad b(\pi) = \times_{i=1}^N b^i(\pi^{-i})$$

for Player  $i$ 's set of best responses to  $\pi^{-i}$ , and for the set of joint best responses to mixed strategy profile  $\pi$  respectively.

In a classical fictitious play process, assuming the players all start with the same prior beliefs about strategies, these beliefs follow the difference inclusion

$$\sigma_{n+1} \in \left(1 - \frac{1}{n+1}\right)\sigma_n + \frac{1}{n+1}b(\sigma_n)$$

(notice this is an ‘inclusion’ since  $b$  is not necessarily single-valued). We can rewrite this as

$$\sigma_{n+1} - \sigma_n \in \alpha_{n+1}(b(\sigma_n) - \sigma_n)$$

with  $\alpha_{n+1} = (n+1)^{-1} \rightarrow 0$ . It is easy to convince oneself that it is reasonable to expect the iterations  $\sigma_n$  to track the differential inclusion

$$\frac{d}{dt}\sigma_t \in b(\sigma_t) - \sigma_t$$

for sufficiently large  $n$ . To capture precisely the way in which this tracking occurs requires the following definition:

**Definition 1.** Let  $\|\cdot\|$  be any norm on  $\mathbb{R}^m$ , and let  $F : \mathbb{R}^m \rightarrow \mathbb{R}^m$  be a closed set-valued map such that  $F(x)$  is a non-empty compact convex subset of  $\mathbb{R}^m$  with  $\sup\{\|z\| : z \in F(x)\} \leq c(1 + \|x\|)$  for all  $x$ . Consider the differential inclusion

$$\frac{dx}{dt} \in F(x). \tag{1}$$

(a) Given a set  $X \subset \mathbb{R}^m$  and points  $x$  and  $y$ , we write  $x \leftrightarrow_X y$  if for every  $\epsilon > 0$  and  $T > 0$  there exists an integer  $n$ , solutions  $\mathbf{x}_1, \dots, \mathbf{x}_n$  to the differential inclusion (1) and real numbers  $t_1, \dots, t_n$  greater than  $T$  such that

- (i)  $\mathbf{x}_i([0, t_i]) \in X$ ,
- (ii)  $\|\mathbf{x}_i(t_i) - \mathbf{x}_{i+1}(0)\| \leq \epsilon$  for all  $i = 1, \dots, n-1$ ,
- (iii)  $\|\mathbf{x}_1(0) - x\| \leq \epsilon$  and  $\|\mathbf{x}_n(t_n) - y\| \leq \epsilon$ .

(b)  $X$  is said to be internally chain-recurrent if  $X$  is compact and  $x \leftrightarrow_X x$  for all  $x \in X$ .

In other words, an internally chain-recurrent set  $X$  of a differential inclusion is a set such that for each  $x \in X$ , there exists a chain of trajectories of the differential inclusion linking  $x$  to itself, all of which are fully contained in  $X$ , and where the start of each trajectory may be a slight perturbation of the end of the previous trajectory. This notion was introduced by Conley (1978), and extends the traditional notion of a recurrent set to the situation where small shocks can be applied to the trajectory of a system, such as those resulting from discretisation. The interested reader is referred to Benaïm and Hirsch (1999), who give a more complete explanation of this concept in a game-theoretical context, while Benaïm et al. (2005) give the full exposition in the case of differential inclusions; note that Benaïm et al. (2005) prefer the phrase “internally chain-transitive” to “internally chain-recurrent”, but we find it useful to emphasise the fact that we are describing an extension of recurrence. Using this concept of chain-recurrence, Benaïm et al. (2005) prove the following general theorem, which holds for any norm  $\|\cdot\|$ :

**Theorem 2.** Assume  $F : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is a closed set-valued map such that  $F(x)$  is a non-empty compact convex subset of  $\mathbb{R}^m$  with  $\sup\{\|z\| : z \in F(x)\} \leq c(1 + \|x\|)$  for all  $x$ . Let  $\{x_n\}_{n \geq 0}$  be a process satisfying

$$x_{n+1} - x_n - \alpha_{n+1} U_{n+1} \in \alpha_{n+1} F(x_n),$$

where  $\{\alpha_n\}_{n \geq 1}$  is a sequence of non-negative numbers such that

$$\sum_{n \geq 1} \alpha_n = \infty, \quad \text{and} \quad \alpha_n \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and  $\{U_n\}_{n \geq 1}$  is a sequence of (deterministic or random) perturbations. If

(1) for all  $T > 0$

$$\lim_{n \rightarrow \infty} \sup_k \left\{ \left\| \sum_{i=n}^{k-1} \alpha_{i+1} U_{i+1} \right\| : \sum_{i=n}^{k-1} \alpha_{i+1} \leq T \right\} = 0,$$

(2)  $\sup_{n \geq 0} \|x_n\| < \infty$ ,

then the set of limit points of  $\{x_n\}_{n \geq 0}$  is a connected internally chain-recurrent set of the differential inclusion

$$\frac{d}{dt} x_t \in F(x_t).$$

Benaïm et al. (2005) show that the conditions are met for a fictitious play process, for which  $U_n = 0$  for all  $n$ , and so prove that the limit points of a fictitious play process are a connected internally chain-recurrent set of the best-response differential inclusion. This result can be used to prove the convergence of fictitious play by showing that all connected internally chain-recurrent sets consist purely of Nash equilibria, which has been shown for two-player zero-sum games (Hofbauer, 1995), potential games (Benaïm et al., 2005), and generic  $2 \times m$  games (Berger, 2005). Recent results (Berger, 2004) show that all trajectories of the best-response differential inclusion converge to Nash equilibrium in further classes of games (including all games

known to have the fictitious play property). Note, however, that this is not a sufficiently strong result for the application of Theorem 2. On the other hand, since the purpose of this paper is to study generalised weakened fictitious play, the strengthening of Berger's results is left to an anticipated paper by Benaïm et al., and so the main result will be in the same form as Theorem 2.

### 3. Generalised weakened fictitious play

Van der Genugten (2000) introduced weakened fictitious play, described briefly in Section 1, as a mechanism for speeding up the convergence of fictitious play in zero-sum games, and therefore proved the result only for two-player zero-sum games and for a rather restricted sequence of  $\epsilon_n$  determining the degree of weakening of the best responses. On the other hand, by linking such a process to the best response differential inclusion, it is intuitively clear that identical results should be achievable for any weakened fictitious play as for classical fictitious play. We proceed to define generalised weakened fictitious play, before proving Theorem 4 which applies Theorem 2 to prove convergence for this wider class of processes.

Define the  $\epsilon$ -best responses of Player  $i$  to opponent mixed strategy profile  $\pi^{-i}$  to be the set

$$b_\epsilon^i(\pi^{-i}) = \{\pi^i \in \Delta^i : r^i(\pi^i, \pi^{-i}) \geq r^i(b^i(\pi^{-i}), \pi^{-i}) - \epsilon\}.$$

That is, the set of Player  $i$ 's strategies that perform not more than  $\epsilon$  worse than Player  $i$ 's best response. The joint  $\epsilon$ -best response to mixed strategy profile  $\pi$  is defined, analogously to the case of best responses, as the set

$$b_\epsilon(\pi) = \prod_{i=1}^N b_\epsilon^i(\pi^{-i}).$$

**Definition 3.** A generalised weakened fictitious play process is any process  $\{\sigma_n\}_{n \geq 0}$ , with  $\sigma_n \in \prod_{i=1}^N \Delta^i$ , such that

$$\sigma_{n+1} \in (1 - \alpha_{n+1})\sigma_n + \alpha_{n+1}(b_{\epsilon_n}(\sigma_n) + M_{n+1}) \tag{2}$$

with  $\alpha_n \rightarrow 0$  and  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ ,

$$\sum_{n \geq 1} \alpha_n = \infty,$$

and  $\{M_n\}_{n \geq 1}$  a sequence of perturbations such that, for any  $T > 0$ ,

$$\lim_{n \rightarrow \infty} \sup_k \left\{ \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\| : \sum_{i=n}^{k-1} \alpha_{i+1} \leq T \right\} = 0.$$

In other words, the current strategies are adapted towards a (possibly perturbed) joint  $\epsilon$ -best response. Clearly, a classical fictitious play process is a generalised weakened fictitious play process with  $\epsilon_n = M_n = 0$  and  $\alpha_n = 1/n$  for all  $n$ . However we will see that both weakening (i.e. allowing non-zero  $\epsilon_n$ ), generalising (letting  $\alpha_n$  be chosen differently), and allowing (certain) perturbations does not affect the convergence result of Benaïm et al. (2005), but allows interesting processes to be considered.

**Theorem 4.** *The set of limit points of a generalised weakened fictitious play process is a connected internally chain-recurrent set of the best response differential inclusion.*

**Proof.** The proof consists of showing that  $\{\sigma_n\}_{n \geq 0}$  satisfies the conditions of Theorem 2. We start by rewriting the definition of a generalised weakened fictitious play process, in a slight abuse of notation, in the form

$$\sigma_{n+1} - \sigma_n - \alpha_{n+1} \{b_{\epsilon_n}(\sigma_n) - b(\sigma_n) + M_{n+1}\} \in \alpha_{n+1} \{b(\sigma_n) - \sigma_n\}.$$

Benaïm et al. (2005) prove that  $F(\sigma) = b(\sigma) - \sigma$  meets the requirements of Theorem 2, and the sequence  $\{\alpha_n\}_{n \geq 1}$  is of the correct form by definition. Therefore it suffices to verify conditions (1) and (2) of Theorem 2. Condition (2) is trivial, since  $\sigma_n \in \times_{i=1}^N \Delta^i$  for all  $n$ . If we take  $k$  such that  $\sum_{i=n}^{k-1} \alpha_{i+1} \leq T$ , then

$$\begin{aligned} & \sup_k \left\| \sum_{i=n}^{k-1} \alpha_{i+1} \{b_{\epsilon_i}(\sigma_i) - b(\sigma_i) - M_{i+1}\} \right\| \\ & \leq \sup_k \left\{ \sum_{i=n}^{k-1} \alpha_{i+1} \|b_{\epsilon_i}(\sigma_i) - b(\sigma_i)\| \right\} + \sup_k \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\| \\ & \leq T \sup_k \|b_{\epsilon_k}(\sigma_k) - b(\sigma_k)\| + \sup_k \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\|. \end{aligned}$$

The second term tends to zero by assumption, and therefore if  $b_{\epsilon}(\sigma) \rightarrow b(\sigma)$  uniformly in  $\sigma$  as  $\epsilon \rightarrow 0$  then the result follows. However, this is immediate from the upper semi-continuity of  $b$ , since the rewards  $r^i$  are bounded.  $\square$

**Corollary 5.** Any generalised weakened fictitious play process will converge to the set of Nash equilibria in two-player zero-sum games, in potential games, and in generic  $2 \times m$  games.

**Proof.** Hofbauer (1995) shows that the set of Nash equilibria is globally attracting for two-player games, and Berger (2005) proves the same for generic  $2 \times m$  games. This is sufficient to prove that any connected internally chain-recurrent set is contained in the set of Nash equilibria. Benaïm et al. (2005) show that any connected internally chain-recurrent set is contained in the set of Nash equilibria for potential games. Combining this with Theorem 4 gives the result.  $\square$

We conclude this section with the observation that both classical fictitious play and Van der Genugten’s weakened fictitious play are both examples of a generalised weakened fictitious play process, obtained by setting  $M_n = 0$ ,  $\alpha_n = 1/n$ , and setting  $\epsilon_n$  identically equal to 0 for classical fictitious play or using a carefully chosen specific sequence of  $\epsilon_n$  values for Van der Genugten’s weakened fictitious play. Hence Theorem 4 and Corollary 5 also hold for these processes. As for these cases, most applications we consider in this paper have perturbation sequences  $\{M_n\}$  that are identically 0. However, for our modification of stochastic fictitious play in Section 4.2 we need to have the full generality of Theorem 4—a suitable form of perturbation sequence  $\{M_n\}_{n \geq 1}$  will be discussed in that section.

#### 4. Examples of GWFP processes

As well as classical fictitious play, and Van der Genugten’s weakened fictitious play, there are several other interesting models of learning that result in generalised weakened fictitious play processes. We discuss three examples here.

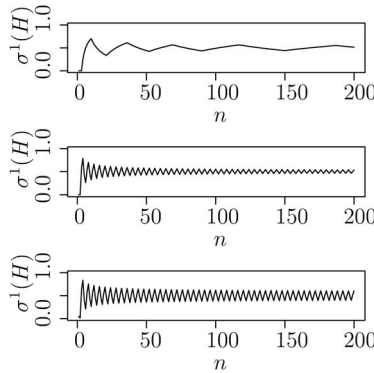


Fig. 1. Belief that Player 1 plays Heads over 200 plays of the two-player matching pennies game under classical fictitious play (top) and under a modified fictitious play with  $\alpha_n = 1/\sqrt{n}$  (middle). The extreme case of  $\alpha_n = 1/\log(n + 2)$  is shown in the bottom plot, where the apparent lack of convergence is due to the fact that the step size is still large ( $\alpha_{200} \approx 0.188$ )—this slowly decreasing step size is, however, likely to be helpful in more complex games.

4.1. Fictitious play with emphasis on recent observations

Recall that in classical fictitious play the adaptation parameters  $\alpha_n$  are simply  $1/n$ , and this means that players make an implicit assumption that all other players have been using the same mixed strategy for all time (since the current estimate of opponent strategy is simply the empirical distribution of actions observed in the past).

However, opponent strategy is not fixed, and an obvious modification is to place greater emphasis on recently observed actions than on the actions played in the early stages of learning. One way in which this can be achieved is to let  $\alpha_n \rightarrow 0$  more slowly than  $1/n$ , for example by setting  $\alpha_n = (C_\alpha + n)^{-\rho_\alpha}$  where  $C_\alpha$  and  $\rho_\alpha \in (0, 1]$  are fixed parameters. Hence a small  $\rho_\alpha$  means that  $\alpha_n \rightarrow 0$  slowly, and so recent observations receive greater weight than under classical fictitious play. An even more extreme example, which still fits the conditions of the theorem, would be  $\alpha_n = 1/\log(C_\alpha + n)$ .

The effect of such a change, in terms of the model, is that beliefs move further on each step of the process, and so should travel more “quickly” along trajectories of the best response differential inclusion. We see this happening in Fig. 1.

4.2. Stochastic fictitious play with vanishing smoothing

One of the major recent modifications of fictitious play is stochastic fictitious play (Fudenberg and Kreps, 1993; Benaïm and Hirsch, 1999), in which players use a smooth best response to their beliefs. An obvious question to ask is, “What happens when the smooth best responses approach best responses as time progresses?”

Player  $i$ 's smooth best response with parameter  $\tau$  to opponent strategy  $\pi^{-i}$  is defined to be the mixed strategy

$$\beta_\tau^i(\pi^{-i}) = \operatorname{argmax}_{\pi^i \in \Delta^i} \{r^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i)\}$$

where  $\tau > 0$  is a temperature parameter, and the function  $v^i$  is a smooth, strictly differentially concave function such that as  $\pi^i$  approaches the boundary of  $\Delta^i$  the slope of  $v^i$  becomes infinite (Fudenberg and Levine, 1998, Chapter 4). The assumption is either that players choose to play

this mixed strategy, or alternatively that they receive payoff perturbations that induce such a mixed strategy (Hofbauer and Sandholm, 2002).

Thus, under vanishing smoothing, if at time  $n$  the players play according to a smooth best response to their current beliefs, instead of (2) we have

$$\sigma_{n+1}^i = (1 - \alpha_{n+1})\sigma_n^i + \alpha_{n+1}(\beta_{\tau_n}^i(\sigma_n^{-i}) + M_{n+1}^i)$$

where  $\tau_n \rightarrow 0$  as  $n \rightarrow \infty$  (vanishing smoothing), and  $M_{n+1}^i$  is the zero mean random variable giving the difference between the actual and expected change in  $\sigma_n^i$  (which exists since the beliefs are not updated towards the mixed strategy  $\beta_{\tau_n}^i(\sigma_n^{-i})$  but instead towards the observed action, as in Benaïm and Hirsch, 1999). These  $M_{n+1}^i$  are bounded martingale differences, so if  $\{\alpha_n\}_{n \geq 1}$  is deterministic and  $o(1/\log(n))$  then the condition on  $\{M_n\}_{n \geq 1}$  in Definition 3 holds with probability 1 (Benaïm et al., 2005).

Furthermore, it is clear that, as  $\tau \rightarrow 0$ ,  $\beta_{\tau}^i(\pi^{-i}) \rightarrow b^i(\pi^{-i})$  for all  $\pi^{-i}$ , and so  $\beta_{\tau_n}^i(\sigma_n^{-i}) \in b_{\epsilon_n}^i(\sigma_n^{-i})$  where  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ . Hence a stochastic fictitious play with decreasing smoothing parameters (or equivalently vanishing payoff perturbations) results almost surely in a generalised weakened fictitious play process, and therefore converges to Nash equilibrium whenever classical fictitious play does.

#### 4.3. Fictitious play in a converging sequence of games

Vrieze and Tijs (1982) considered fictitious play in a converging sequence of games, which proved useful in their study of fictitious play in stochastic games. We note here that such a fictitious play results in a generalised weakened fictitious play for the limit game, since a best response in the converging sequence of games is an  $\epsilon$ -best response in the limit game. However the notational overhead required here to prove this is not justified by the result, which is instead proved in Leslie (2003).

### 5. An actor-critic process

All of the processes discussed so far require that players observe opponent actions, and use their knowledge of the payoff functions to calculate a best response. However, it is of interest to determine whether or not simple adaptive play can converge to equilibrium strategies without this information, and without the players even knowing that they are playing a game. Such learning processes have been described as “simple, payoff-based” learning processes (Foster and Young, 2003b). Recent attempts to examine such processes have focussed on a particular form of reinforcement learning (Roth and Erev, 1995). Beggs (2005) shows that under this model players will learn not to play actions that are removed under iterated removal of dominated actions, and that for two-player constant-sum games the players will learn to play Nash equilibrium. Similarly, Hopkins and Posch (2005) show that play cannot converge to a joint mixed strategy that is not a Nash equilibrium, and that play converges to an equilibrium in two-player partnership games. These results, as well as those of Börgers and Sarin (1997) on a different version of reinforcement learning, are obtained by exploiting a relationship with the replicator dynamics of evolutionary game theory (Taylor and Jonker, 1978), analogously to the relationship between generalised weakened fictitious play processes and the best-response differential inclusion exploited in this paper. Other simple, payoff-based learning processes introduced by the authors (Leslie and Collins, 2003; Leslie and Collins, 2005) have been analysed using their relationship

with the smooth best response dynamics (Hofbauer and Hopkins, 2005). In this section we modify an actor-critic process introduced in Leslie and Collins (2003) to give a simple payoff-based learning process which results in a generalised weakened fictitious play process.

Suppose that an oracle tells each player the current expected reward associated with each action, and that players adjust their strategies towards a reward-maximising action. Then the strategies will follow a process

$$\pi_{n+1} \in (1 - \alpha_{n+1})\pi_n + \alpha_{n+1}b(\pi_n),$$

which is clearly the same as a fictitious play process. However, relying on an oracle is not a feasible learning process in the real world; we will see in this section how players can effectively provide themselves with a fuzzy oracle, which tells them approximately what their current expected rewards are, and this allows strategies to follow a generalised weakened fictitious play process.

In an actor-critic process, each player has both an actor component (the current strategy) and a critic component (estimates of action values) which is used to inform the actor (i.e. update the strategy). We will write  $\pi_n^i \in \Delta^i$  for Player  $i$ 's strategy at time  $n$ , and  $Q_n^i \in \mathbb{R}^{|A^i|}$  for Player  $i$ 's estimates of action values at time  $n$ . At each stage of the process,  $\pi_n^i$  will be adjusted towards a smooth best response based upon the estimates  $Q_n^i$ , while  $Q_n^i$  will be updated based on the observed payoff. As before, we will write  $\pi_n = \times_{i=1}^N \pi_n^i$ ,  $Q_n = \times_{i=1}^N Q_n^i$ , and  $r(\pi) = \times_{i=1}^N r^i(\cdot, \pi^{-i})$  (so that  $Q_n$  will approximate  $r(\pi_n)$ ), and define the following analogues of the  $\epsilon$ -best response correspondences:

$$\tilde{b}_\epsilon^i(Q^i) = \{ \pi^i \in \Delta^i : \pi^i \cdot Q^i \geq \max_{a^i \in A^i} Q^i(a^i) - \epsilon \},$$

$$\tilde{b}_\epsilon(Q) = \times_{i=1}^N \tilde{b}_\epsilon^i(Q^i).$$

**Lemma 6.** *If the strategies  $\pi_n$  follow a process*

$$\pi_{n+1} \in (1 - \alpha_{n+1})\pi_n + \alpha_{n+1}\tilde{b}_{\epsilon_n}(Q_n)$$

*with  $\alpha_n \rightarrow 0$  and  $\epsilon_n \rightarrow 0$ , and if  $\|Q_n - r(\pi_n)\| \rightarrow 0$  as  $n \rightarrow \infty$ , then the  $\pi_n$  follow a generalised weakened fictitious play process.*

**Proof.** Suppose  $\|Q - r(\pi)\| < \eta$ , and  $\tilde{\pi} \in \tilde{b}_\epsilon(Q)$ . Then

$$\tilde{\pi}^i \cdot Q^i \geq \max_{a^i \in A^i} Q^i(a^i) - \epsilon \geq \max_{a^i \in A^i} r^i(a^i, \pi^{-i}) - \eta - \epsilon.$$

Hence

$$\tilde{b}_{\epsilon_n}(Q_n) \subset b_{\epsilon'_n}(\pi_n)$$

for some  $\epsilon'_n \rightarrow 0$ , and the  $\pi_n$  follow a generalised weakened fictitious play process.  $\square$

This leaves only the problem of how to obtain estimates  $Q_n$  that are asymptotically close to  $r(\pi_n)$ . These can be obtained using reinforcement learning (Sutton and Barto, 1998), although we need to be careful that all actions are played often enough that estimates remain close. We start by extending a result of Singh et al. (2000) to give conditions under which the probability of playing any action is bounded below by a suitable decreasing sequence. (In Leslie, 2003 the condition was forced by projecting strategies away from the boundary of strategy space, but the approach here is more readily justified as a model of learning.)

**Lemma 7.** *Suppose that strategies evolve according to*

$$\pi_{n+1}^i(a^i) = (1 - \alpha_{n+1})\pi_n^i(a^i) + \alpha_{n+1}\beta_n^i(a^i) \tag{3}$$

for each  $i$  and each  $a^i \in A^i$ , where  $\beta_n^i$  is Player  $i$ 's Boltzmann smooth best response, with

$$\beta_n^i(a^i) = \frac{\exp(Q_n(a^i)/\tau_n^i)}{\sum_{a \in A^i} \exp(Q_n(a)/\tau_n^i)} \text{ for each } a^i \in A^i. \tag{4}$$

Suppose further that the temperature parameters  $\tau_n^i$  are chosen to be

$$\tau_n^i = \frac{\max_{a \in A^i} Q_n^i(a) - \min_{a \in A^i} Q_n^i(a)}{\rho \log(n)} \tag{5}$$

with fixed constant  $\rho \in (0, 1]$ . If  $\sup_n \|Q_n\| < \infty$  then:

- (1)  $\tau_n^i \rightarrow 0$  for all  $i$ , and
- (2)  $\exists K, \kappa > 0$  such that for  $n \geq K$ , for each  $i$  and for each  $a^i \in A^i$ ,

$$\pi_n^i(a^i) \geq \frac{\kappa}{n^\rho}.$$

**Proof.** We start by noting that, since  $\sup_n \|Q_n\| < \infty$ ,  $\tau_n^i \rightarrow 0$  for all  $i$ .

For the second part, note first that

$$\tau_n^i \geq \frac{\max_{a \in A^i} Q_n^i(a) - Q_n^i(a^i)}{\rho \log n}$$

for all  $a^i \in A^i$ , so that

$$\begin{aligned} n^\rho \exp(Q(a^i)/\tau_n^i) &\geq \exp(\max_{a \in A^i} Q_n^i(a)/\tau_n^i) \\ &= \frac{1}{|A^i|} (|A^i| \exp(\max_{a \in A^i} Q_n^i(a)/\tau_n^i)) \\ &\geq \frac{1}{|A^i|} \sum_{a \in A^i} \exp(Q_n^i(a)/\tau_n^i), \end{aligned}$$

and hence

$$\beta_n^i(a^i) = \frac{\exp(Q_n^i(a^i)/\tau_n^i)}{\sum_{a \in A^i} \exp(Q_n^i(a)/\tau_n^i)} \geq \frac{1}{|A^i|n^\rho}$$

for all  $a^i \in A^i$ .

To show that  $\pi_n^i(a^i) \geq \frac{\kappa}{n^\rho}$  for sufficiently large  $n$ , observe from (3) that if  $\pi_n^i(a^i) \geq \frac{1}{|A^i|n^\rho}$  then  $\pi_{n+1}^i(a^i) > \frac{1}{|A^i|(n+1)^\rho}$ , and if  $\pi_n^i(a^i) < \frac{1}{|A^i|n^\rho}$  then  $\pi_{n+1}^i(a^i) > \pi_n^i(a^i)$ . Hence there exists an  $K_i$  such that  $\pi_{K_i}^i(a^i) \geq \frac{1}{|A^i|K_i^\rho}$  (since otherwise  $\pi_n^i(a^i)$  is monotonically increasing and always less than  $\frac{1}{|A^i|n^\rho}$ ), and then  $\pi_n^i(a^i) \geq \frac{1}{|A^i|n^\rho}$  for all  $n \geq K_i$ . The result follows on taking  $\kappa = \min_i \frac{1}{|A^i|}$  and  $K = \max_i K_i$ .  $\square$

This result enables us to define the following actor-critic process, in which players estimate expected rewards and adapt strategies towards smooth best responses to these rewards in such a way as to ensure that all actions are played sufficiently often to maintain accurate estimates.

**Definition 8.** A Boltzmann actor–critic process is a process  $\{\pi_n, Q_n\}$  such that

$$\pi_{n+1}^i(a^i) = (1 - \alpha_{n+1})\pi_n^i(a^i) + \alpha_{n+1}\beta_n^i(a^i), \quad (6)$$

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}^i \mathbb{I}_{\{a_n^i = a^i\}} (R_n^i - Q_n^i(a^i)), \quad (7)$$

for each  $i$  and each  $a^i \in A^i$ , where:

- $a_n^i$ , the action played by Player  $i$  at time  $n$ , is selected according to strategy  $\pi_n^i$ , and results in reward  $R_n^i$ , and
- $\beta_n^i$  is as defined in (4) and (5).

**Theorem 9.** Suppose that  $\{\pi_n, Q_n\}$  is a Boltzmann actor–critic process for which

- $\alpha_{n+1} = (C_\alpha + n)^{-\rho_\alpha}$  where  $C_\alpha$  and  $\rho_\alpha$  are chosen to satisfy  $C_\alpha > 0$  and  $\rho_\alpha \in (0.5, 1]$ ,
- $\lambda_{n+1}^i = (C_\lambda + c_n^i(a^i))^{-\rho_\lambda}$  where  $c_n^i(a^i) = \sum_{k=1}^n \mathbb{I}_{\{a_k^i = a^i\}}$  is the number of times action  $a^i$  has been selected up to and including game  $n$ , and  $C_\lambda$  and  $\rho_\lambda$  are chosen to satisfy  $C_\lambda > 0$  and  $\rho_\lambda \in (0.5, \rho_\alpha)$ ,
- $\rho = \rho_\pi$ , used to calculate the Boltzmann smooth best responses in (4) and (5), is chosen to satisfy  $\rho_\pi \in (0, \rho_\alpha - \rho_\lambda)$ .

Then, with probability 1, the  $\pi_n$  follow a generalised weakened fictitious play process.

**Proof.** By Lemma 7, if  $\|Q_n - r(\pi_n)\| \rightarrow 0$  then  $\tau_n^i \rightarrow 0$  for each  $i$  and by Lemma 6 we have a generalised weakened fictitious play process. Hence it suffices to prove that  $\|Q_n - r(\pi_n)\| \rightarrow 0$ . We assume in what follows that  $n$  is sufficiently large such that  $\pi_n^i(a^i) \geq \kappa n^{-\rho_\pi}$  for each  $i$  and each  $a^i \in A^i$  (see Lemma 7).

Fix  $i$  and  $a^i$ , and let  $\{v_k\}_{k \geq 1}$  be the sequence of times when action  $a^i$  is played by Player  $i$  (this sequence is well-defined, by Lemma 7 and the Borel–Cantelli lemma), and define the differences

$$D_k = Q_{v_k}^i(a^i) - r^i(a^i, \pi_{v_k}^{-i}).$$

We will show that  $D_k \rightarrow 0$  almost surely, which is sufficient to prove the result, since all actions are played with positive probability at all times.

Note, from (7) and the definition of  $v_k$ , that

$$Q_{v_{(k+1)}}^i(a^i) = Q_{(v_k)+1}^i(a^i) = (1 - \lambda_{(v_k)+1}^i) Q_{v_k}^i(a^i) + \lambda_{(v_k)+1}^i R_{v_k}^i.$$

Hence

$$\begin{aligned} D_{k+1} &= Q_{v_{(k+1)}}^i(a^i) - r^i(a^i, \pi_{v_{(k+1)}}^{-i}) \\ &= (1 - \lambda_{(v_k)+1}^i) (D_k + r^i(a^i, \pi_{v_k}^{-i})) + \lambda_{(v_k)+1}^i R_{v_k}^i - r^i(a^i, \pi_{v_{(k+1)}}^{-i}) \\ &= (1 - \lambda_{(v_k)+1}^i) D_k + \lambda_{(v_k)+1}^i (R_{v_k}^i - r^i(a^i, \pi_{v_k}^{-i})) + r^i(a^i, \pi_{v_k}^{-i}) - r^i(a^i, \pi_{v_{(k+1)}}^{-i}). \end{aligned}$$

Since  $\lambda_{(v_k)+1}^i = (C_\lambda + k)^{-\rho_\lambda}$ , we see that

$$D_{k+1} = (1 - (C_\lambda + k)^{-\rho_\lambda}) D_k + (C_\lambda + k)^{-\rho_\lambda} (M_k - E_k),$$

where  $M_k$  is a bounded martingale difference, and

$$E_k = (C_\lambda + k)^{\rho_\lambda} \{r^i(a^i, \pi_{v_{k+1}}^{-i}) - r^i(a^i, \pi_{v_k}^{-i})\}.$$

If we can show that  $\|E_k\| \rightarrow 0$  almost surely, then Theorem 2 applies to  $D_k$  with  $U_k = M_k - E_k$  and  $F(D) = -D$ . The unique compact invariant set of the associated differential inclusion is  $D = 0$ , and the result follows.

Notice that  $r^i(a^i, \pi^{-i})$  is continuous in  $\pi^{-i}$ , so from (6)

$$\|r^i(a^i, \pi_{n+1}^{-i}) - r^i(a^i, \pi_n^{-i})\| \leq C\alpha_{n+1}$$

for some  $C$  (depending only on the reward function  $r^i$ ). Therefore

$$C^{-1}\|E_k\| \leq (C_\lambda + k)^{\rho_\lambda} \sum_{j=v_k}^{v_{k+1}-1} \alpha_{j+1} \leq (C_\lambda + k)^{\rho_\lambda} (v_{k+1} - v_k)\alpha_{(v_k)+1}.$$

Furthermore, since  $v_k \geq k$ , and  $\alpha_{(v_k)+1} = (C_\alpha + v_k)^{-\rho_\alpha}$ ,

$$\|E_k\| \leq C' \frac{v_{k+1} - v_k}{v_k^{\rho_\alpha - \rho_\lambda}},$$

for some constant  $C'$ . Thus, by the Borel–Cantelli lemma, we see that  $\|E_k\| \rightarrow 0$  almost surely if, for arbitrary  $\delta > 0$ ,

$$\sum_{k \geq 1} \mathbb{P}\left(\frac{v_{k+1} - v_k}{v_k^{\rho_\alpha - \rho_\lambda}} > \delta\right) < \infty. \tag{8}$$

Fix  $v_k$ , and let  $j$  be the greatest integer less than or equal to  $\delta(v_k)^{\rho_\alpha - \rho_\lambda}$ . Then

$$\begin{aligned} & \mathbb{P}\left(\frac{v_{k+1} - v_k}{v_k^{\rho_\alpha - \rho_\lambda}} > \delta\right) \\ & \leq \mathbb{P}(v_{k+1} > v_k + j) \\ & \leq (1 - \kappa(v_k + 1)^{-\rho_\pi})(1 - \kappa(v_k + 2)^{-\rho_\pi}) \cdots (1 - \kappa(v_k + j)^{-\rho_\pi}) \\ & \quad (\text{since } \pi_n^i(a^i) \geq \kappa n^{-\rho_\pi}) \\ & < (1 - \kappa(v_k + j)^{-\rho_\pi})^j \\ & < \exp\left\{-\frac{j\kappa}{(v_k + j)^{\rho_\pi}}\right\} \\ & < \exp\left\{-\kappa \frac{\delta(v_k)^{\rho_\alpha - \rho_\lambda} - 1}{(v_k + \delta(v_k)^{\rho_\alpha - \rho_\lambda})^{\rho_\pi}}\right\} \\ & = \exp\left\{\frac{\kappa}{(v_k + \delta(v_k)^{\rho_\alpha - \rho_\lambda})^{\rho_\pi}}\right\} \exp\left\{-\kappa \delta \frac{(v_k)^{\rho_\alpha - \rho_\lambda - \rho_\pi}}{(1 + \delta(v_k)^{\rho_\alpha - \rho_\lambda - 1})^{\rho_\pi}}\right\}. \end{aligned}$$

Since  $v_k \geq 1$  and  $0 < \rho_\alpha - \rho_\lambda < 1$ ,

$$\begin{aligned} \frac{\kappa}{(v_k + \delta(v_k)^{\rho_\alpha - \rho_\lambda})^{\rho_\pi}} & \leq \frac{\kappa}{(1 + \delta)^{\rho_\pi}}, \quad \text{and} \\ -\kappa \delta \frac{(v_k)^{\rho_\alpha - \rho_\lambda - \rho_\pi}}{(1 + \delta(v_k)^{\rho_\alpha - \rho_\lambda - 1})^{\rho_\pi}} & \leq -\kappa \delta \frac{(v_k)^{\rho_\alpha - \rho_\lambda - \rho_\pi}}{(1 + \delta)^{\rho_\pi}}. \end{aligned}$$

Hence there exist constants  $C_1, C_2 > 0$ , independent of  $k$  and  $v_k$ , such that

$$\mathbb{P}\left(\frac{v_{k+1} - v_k}{v_k^{\rho_\alpha - \rho_\lambda}} > \delta\right) \leq C_1 \exp\{-C_2(v_k)^{\rho_\alpha - \rho_\lambda - \rho_\pi}\}.$$

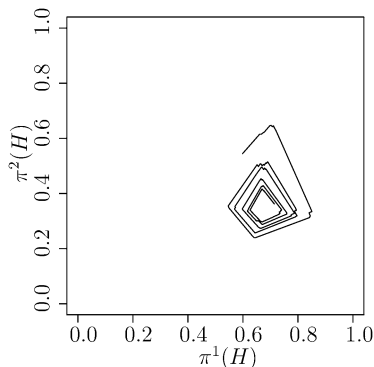


Fig. 2. Strategies of both players over  $10^7$  iterations of the discontinuous actor-critic process in the game (9) (first 5000 iterations omitted). The parameters are  $\rho_\alpha = 1.0$ ,  $\rho_\lambda = 0.6$ , and  $\rho_\pi = 0.1$ , with  $C_\alpha = C_\lambda = 1$ . Strategies spiral clockwise towards the unique equilibrium point.

Now, for  $\eta > 0$ ,

$$\int_0^\infty e^{-C_2 x^\eta} dx \stackrel{y:=x^\eta}{=} \eta^{-1} \int_0^\infty y^{\eta^{-1}-1} e^{-C_2 y} dy = \eta^{-1} \Gamma(\eta^{-1}) C_2^{-\eta^{-1}}$$

where  $\Gamma$  is the Gamma function. Therefore  $\sum_{k \geq 0} C_1 \exp\{-C_2 k^{\rho_\alpha - \rho_\lambda - \rho_\pi}\} < \infty$ , since  $\rho_\pi < \rho_\alpha - \rho_\lambda$  by assumption, and we see that (8) holds.  $\square$

Thus we have shown that this actor-critic process, in which players do not know the reward function, and pay no attention to the other players, is a member of the same class of processes as fictitious play and its variants. Hence it converges in the same games in which we have shown all the other generalised weakened fictitious play processes to converge. This shows that players can learn to play Nash equilibrium strategies in certain classes of games, without having knowledge of the game, or even knowing they are playing a game.

We conclude our analysis of this process by presenting the results of an experiment in the game with reward matrix

$$\begin{pmatrix} (2, 0) & (0, 1) \\ (0, 2) & (1, 0) \end{pmatrix}. \tag{9}$$

This is a rescaled zero-sum game, so strategies evolve exactly as if it is zero-sum (Hofbauer and Sigmund, 1998), and should therefore converge to the unique Nash equilibrium where  $\pi^1 = (2/3, 1/3)$  and  $\pi^2 = (1/3, 2/3)$ . This equilibrium requires the players to assign unequal probabilities to their two actions, despite the fact that both actions receive the same expected reward when equilibrium strategies are played—a situation that could easily cause problems for players that do not observe opponent behaviour or know anything about the game (see Leslie and Collins, 2005 for further discussion of this issue). However, as observed here in Fig. 2, and as predicted by the theoretical results, the actor-critic process is not afflicted by this problem.

**Acknowledgments**

We are indebted to the editor and to two anonymous referees for their helpful comments, and to Dr. Stas Volkov for suggesting the use of the Borel-Cantelli lemma.

## References

- Beggs, A.W., 2005. On the convergence of reinforcement learning. *J. Econ. Theory* 122, 1–36.
- Benaïm, M., Hirsch, M.W., 1999. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games Econ. Behav.* 29, 36–72.
- Benaïm, M., Hofbauer, J., Sorin, S., 2005. Stochastic approximation and differential inclusions. *SIAM J. Control Optim.* 44, 328–348.
- Berger, U., 2004. Two more classes of games with the fictitious play property. Mimeo. Vienna University of Economics.
- Berger, U., 2005. Fictitious play in  $2 \times n$  games. *J. Econ. Theory* 120, 139–154.
- Börgers, T., Sarin, R., 1997. Learning through reinforcement and replicator dynamics. *J. Econ. Theory* 77, 1–14.
- Brown, G.W., 1951. Iterative solution of games by fictitious play. In: Koopmans, T.C. (Ed.), *Activity Analysis of Production and Allocation*. Wiley, New York, pp. 374–376.
- Conley, C.C., 1978. Isolated invariant sets and the Morse index. In: *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence.
- Cowan, S., 1992. Dynamical systems arising from game theory. Ph.D. thesis. University of California, Berkeley.
- Foster, D.P., Young, H.P., 2003a. Learning, hypothesis testing, and Nash equilibrium. *Games Econ. Behav.* 45, 73–96.
- Foster, D.P., Young, H.P., 2003b. Regret testing: A simple payoff-based procedure for learning Nash equilibrium. Available at <http://www.econ.jhu.edu/People/Young/>.
- Fudenberg, D., Kreps, D.M., 1993. Learning mixed equilibria. *Games Econ. Behav.* 5, 320–367.
- Fudenberg, D., Levine, D.K., 1998. *The Theory of Learning in Games*. MIT Press, Cambridge, MA.
- Gilboa, I., Matsui, A., 1991. Social stability and equilibrium. *Econometrica* 59, 859–867.
- Hart, S., Mas-Colell, A., 2001. A reinforcement procedure leading to correlated equilibrium. In: Debreu, G., Neufeind, W., Trockel, W. (Eds.), *Economic Essays: A Festschrift for Werner Hildenbrand*. Springer, New York, pp. 181–200.
- Hart, S., Mas-Colell, A., 2003. Uncoupled dynamics cannot lead to Nash equilibrium. *Amer. Econ. Rev.* 93, 1830–1836.
- Hofbauer, J., 1995. Stability for the best response dynamics. Technical report. Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Vienna, Austria.
- Hofbauer, J., Hopkins, E., 2005. Learning in perturbed asymmetric games. *Games Econ. Behav.* 52, 133–152.
- 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 Univ. Press.
- Hopkins, E., Posch, M., 2005. Attainability of boundary points under reinforcement learning. *Games Econ. Behav.* 44, 459–514.
- Leslie, D.S., 2003. Reinforcement learning in games. Ph.D. thesis. University of Bristol.
- Leslie, D.S., Collins, E.J., 2003. Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Ann. Appl. Probability* 13, 1231–1251.
- Leslie, D.S., Collins, E.J., 2005. Individual  $Q$ -learning in normal form games. *SIAM J. Control Optim.* 44, 459–514.
- Miyasawa, K., 1961. On the convergence of the learning process in a  $2 \times 2$  non-zero-sum two-person game. Research Memorandum 33. Econometric Research Program, Princeton University, Princeton.
- Monderer, D., Shapley, L.S., 1996. Fictitious play property for games with identical interests. *J. Econ. Theory* 68, 258–265.
- Robinson, J., 1951. An iterative method of solving a game. *Ann. Math.* 54, 296–301.
- Roth, A.E., Erev, I., 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav.* 8, 164–212.
- Singh, S., Jaakkola, T., Littman, M.L., Szepesvari, C., 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning* 38, 287–308.
- Sutton, R.S., Barto, A.G., 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Taylor, P.D., Jonker, L.D., 1978. Evolutionarily stable strategies and game dynamics. *Math. Biosci.* 40, 145–146.
- Van der Genugten, B., 2000. A weakened form of fictitious play in two-person zero-sum games. *Int. Game Theory Rev.* 2, 307–328.
- Vrieze, O.J., Tijs, S.H., 1982. Fictitious play applied to sequences of games and discounted stochastic games. *Int. J. Game Theory* 11, 71–85.