An Analysis of a Simple Reinforcing Dynamics: Learning to Play an “Egalitarian” Equilibrium*

Alexandre Possajennikov
CentER for Economic Research
Tilburg University
P.O.Box 90153
5000 LE Tilburg
The Netherlands

e-mail: A.Possajennikov@kub.nl

January 1997

Abstract

The paper analyses a simple reinforcing dynamics. The dynamics can be interpreted as a learning dynamics with fixed aspiration level. All payoffs are assumed to be above this aspiration level, therefore all strategies are reinforcing. Different versions of the dynamics exhibit different convergence properties. The analysis starts with one-agent decision problems and proceeds to games. Some results are available for decision problems and simple games. For complex games computer simulations are performed. The hypothesis is that the dynamics favors an "egalitarian" equilibrium even if it does not satisfy other refinements.

Keywords: Equilibrium selection; Stochastic learning; Bounded rationality.
JEL Codes: C72, D81

* I am grateful to Eric van Damme, Aldo Rustichini, Martin Dufwenberg, Bas Werker, Jos Jansen, Tone Dieckmann for help and suggestions on the paper. I thank also participants of the seminar at Tilburg University and the conference at Stony Brook for discussion and CentER for Economic Research for providing conditions for the research.

1 Introduction

Evolutionary and learning processes in games attract much attention at the present time. In these processes the dynamics of the game, or players' choice of strategies over time, is modeled explicitly. One can distinguish two main applications of such modeling. First, explicit dynamics usually gives the answer to the question which equilibrium will be selected starting from certain initial conditions. Stochastic dynamic processes might serve as a selecting device among equilibria. Therefore, the evolutionary processes can be used to refine Nash equilibria, that is for the very theoretical problems of game theory. Seminal papers on this question are Kandori *et al*. (1993) and Young (1993).

The second application of evolutionary and learning processes is a more positive one. By means of experiments the dynamics approximating real human behavior can be obtained and analyzed. Usually people playing a game do not perform sophisticated calculations to find a Nash equilibrium but learn to play the game by a simple learning dynamics. When more than one equilibrium is present, sometimes the players fail to arrive at the "plausible" (e.g. subgame perfect) Nash equilibrium or even at a Nash equilibrium at all under simple dynamics. Even in case of one-player decision problems the optimal behavior is not guaranteed. Known phenomena are "probability matching" (see, for example, Börgers and Sarin (1996)) and "melioration" (see Herrnstein and Prelec (1991)), where examples are given of dynamics and human behavior which show non-optimality. Arthur (1993) also gives some experimental data on non-optimal behavior in decision problems. A case of non-equilibrium play in games was reported in Roth and Erev (1995).

Those papers assert that a simple dynamics can explain to a large extent the behavior of players in decision problems and simple games. The dynamics, as the human behavior in their experiments, does not necessarily converge to the optimal action or to a (subgame perfect) equilibrium in some of the examples they consider. In this paper we investigate this dynamics analytically whenever possible and by means of computer simulations in more complicated cases. Although the dynamics is formulated quite simply, the analysis of it requires sophisticated tools of stochastic optimization.

We compare also different versions of the dynamics, since some modifications of it lead to different convergence results. However, we shall focus not only on the long run optimality but also on the medium run since it can be of economic importance. The environment in an economy is not likely to be constant indefinitely. Thus the speed of convergence will play a role in our analysis. There is a trade-off between long-run convergence and speed thereof and we shall consider this issue.

The remainder of the paper is structured as follows: Section 2 describes the model, Section 3 gives the analysis for the case of one-player decision problem, Section 4 reports results of simulations for several games and Section 5 concludes.

2 The model

We formulate the model for games though it can be easily simplified to decision problems. There are n players. Denote the set of players by I. The stage game is a game in normal form. Let $S_1,...,S_n$ be the sets of pure strategies. The payoff functions $\pi_m: S_1 \times ... \times S_n \to R$ are assumed to be non-negative for every player and for every profile of pure strategies, $\pi_m(s_1,...s_n) \geq 0\ \forall m \in \mathrm{I}, \forall s_1 \in S_1,...,s_n \in S_n$.

Let player m have k pure strategies, $|S_m| = k$. The state of player m at time t is described by the vector $q^t_m = (q^t_{m1},...,q^t_{mk}) \in R^k$. q^t_{mj} denotes the propensity of player m to play the strategy $j \in S_m$ at time t. The propensities are assumed to be strictly positive, $q^t_{mj} > 0$ $\forall t, \forall m, \forall j$. Denote the sum of the propensities $\Sigma^k_{i=1}\, q^t_{mi}$ by Q^t_m. Given the vector of propensities the probability to play strategy j is defined then as

$$p^t_{mj} = \frac{q^t_{mj}}{Q^t_m}$$

The vector $p^t_m = (p^t_{m1},...,p^t_{mk}) \in \Delta S_m$, where ΔS_m is the set of mixed strategies of player m. Working with propensities rather than probabilities is easier since with probabilities we have to change them in such a way that they were between 0 and 1 while the only restriction on propensities that they are non-negative.

An interpretation of this is that every player possesses at every moment of time a mixed strategy, characterized by the vector of probabilities p^t_m. Another interpretation can be that for every player $1,...,n$ there is a large population of agents, the members of the population possess a pure strategy and in each population the distribution of pure strategies is given by the vectors p^t.

The state of the whole process at time t is determined by the vectors $q^t_1,...q^t_n$. According to the vector of probabilities derived from q's, each player chooses a pure strategy to play in the present period. The precise mechanism of choice is not modeled. In the first interpretation above it could be a random device used by a player. In the second interpretation above such a choice can be understood as a random draw of an agent with a pure strategy from the population. The stage game is played with the chosen strategies. We are interested in the dynamics of p^t_m. Since it is determined through q^t_m, we must specify the dynamics of the propensities.

We pose two main requirements on the dynamics. Firstly, it should be reinforcing: if a strategy is played, its probability increases. Secondly, it should be in a sense "simple": no complex functions must be involved. One class of dynamics satisfying these requirements is following. The player triggers a strategy, observes the payoff and increases the propensity of playing this strategy by this payoff. Then he renormalizes the propensities by multiplying them by a certain variable, since this does not change the probabilities. Therefore the expected motion of the dynamics does not change. However, the normalization plays an important role, since it change the variance of the stochastic process which can influence the convergence results and the speed of convergence. We try to answer the question for which forms of normalization the dynamics converges to an equilibrium and for which forms the speed of the dynamics is fast enough.

Formally, if player m chooses strategy j while the other players choose strategies $s^{-m}\in S^{-m}$, then the state of player m at time $t + 1$ is defined as follows:

$$q^{t+1}_{mj} = (q^t_{mj} + B_{mj})A^t_m$$
$$q^{t+1}_{mk} = q^t_{mk}A^t_m, k \neq j$$

where $B_{mj} = \pi(j,s^{-m})$ is the payoff, A^t_m is the normalizing multiplier. To keep the propensities positive $\pi(j,s^{-m})$ were assumed non-negative. The normalizing multiplier can be chosen as to keep the sum of the propensities equal to a predetermined variable C^t_m, in which case $A^t_m = \frac{C^{t+1}_m}{Q^t_m + B_{mj}}$. We consider this type of normalization since it is mathematically tractable while gives already rich variety of results.

The variable C^t_m can be deterministic, for example, $C^t_m = Ct^\nu$, where C,ν are constant. In case of no normalization $A^t_m = 1$ or $C^{t+1}_m = Q^t_m + B_{mj}$, a random variable. Another interesting case is $A^t_m = \delta < 1$. The parameter δ can be understood as a forgetting parameter since payoff got τ periods before will enter the sum multiplied by δ^τ. C^t_m will determine the step size of the algorithm as it would be clear later. In fact, the inverse to C^t_m has the same order as the step size of the dynamics. At time $t = 0$ the vectors of initial propensities q^0_m are given, $q^0_{mj} > 0$ $\forall m \in \mathrm{I}, \forall j \in S_m$.

The model is fairly simple. However, it captures some important aspects of human behavior. The updating of the strategies depends only on player's own payoff. Notice that it is exactly what is being done in one-agent decision problems with unknown distributions of payoffs. The justification of carrying the dynamics to games is that people may not know that they are participating in a game, or who their opponents are, or the preferences of the opponents. Of course, in economic reality it is not the case, people usually have some vague idea what is going on, or form expectations, but we shall consider this extreme case. The model can also be applied to extensive form games where not all information sets are reached during the course of the game, because the players do not need to know the other players' exact strategy.

The application of the model to one-agent decision problems and calibration of the normalization version of it with $C^t_m = Ct^\nu$ against human subjects was considered in Arthur (1993). Roth and Erev (1995) also applied the non-normalized version with some other extensions in some specific games to approximate human behavior. Posch (1996) considered the normalized ($C^t_m = Ct^\nu$) dynamics in *2x2* games. Laslier *et al*. (1996) analyzed a more general version for both decision problems with 2 actions and *2x2* games.

In their paper the new propensities were determined by a general function of the past play, not necessarily the normalized sum. They considered the long-run convergence of the dynamics. However, not much attention was paid to the speed of convergence and medium term results, which certainly has an economic relevance.

An extended analysis of similar machine learning is provided in Narendra and Thathachar (1989). However, they consider mostly the schemes with binary payoffs or with reward-penalty nature, which require knowledge of maximal and minimal payoff. If a payoff to an action is close to the maximal one, the probability of playing this action increases (reward) while if a payoff is close to the minimal one, the probability decreases (penalty). Our scheme can be considered as reward-reward scheme since independently of the outcome the probability of playing a strategy increases. Some of the results for reward-penalty schemes carry over to our scheme.

Roth and Erev (1995) argue that the dynamics captures two important aspects of learning. The first one, the “Law of Effect”, states that the choices that have led to good outcomes should be repeated more often in the future. In the model, each strategy gives a positive payoff and the probability of playing it at the next round increases, hence this law is fulfilled. The second aspect, the “Power Law of Practice”, says that learning tends to be fast in the beginning and then slows down. In the no normalization case the propensities only can increase not more than by a fixed amount (the maximal payoff to a strategy), so it is indeed the case in the model. If $\nu > 0$ in normalization then C^t_m grows over time and the learning slows down. The payoff at a late stage in the game changes the probability less than at an early stage, when the aggregate propensity is not yet very high.

Another interpretation of the dynamics can be in the spirit of learning dynamics with an aspiration level. In such dynamics a strategy is regarded as successful and its weight is increased if it gives a payoff that is greater than the aspiration level. In our model the aspiration level is set to 0, hence every strategy is successful or at least, not unsuccessful. The aspiration level does not change throughout the game. From the “aspiration level” literature side, a similar model for decision problems was considered by Börgers and Sarin(1996), where the probabilities change directly, not through propensities. A general model with aspirations can be found in Bendor *et al* (1994).

3 The one-player decision problem

The long run convergence

Though the ultimate goal is to analyze the dynamics in games, it is interesting to consider its behavior in one-player decision problems. Binmore *et al* (1996) expressed view that it might be wiser to consider first the human behavior in decision problems and then apply it to games. We shall proceed in this spirit.

The main question for every dynamics is whether it converges and if so, to what point. For games it is desirable that a dynamics converges to a Nash equilibrium and even better to a certain refined Nash equilibrium, for example, to a (subgame) perfect equilibrium. In one-agent decision problems Nash equilibrium corresponds to the choice of the action which gives highest expected payoff. Since all strategies are successful, the problem of path-dependency may arise and it is not clear that the optimal action will be chosen in the long run.

From the model described in Section 2 we have now $n = 1$ and we can omit subscript m in the formulas. The payoff to strategy j $\pi(s_j)$ is now a random variable, given by the environment. The environment is assumed to have a finite number of states which occur with fixed probabilities independent of time. Denote the realization of $\pi(s_j)$ at time t by B^t. The agent still has k strategies or actions.

Let e_j be the unit k-vector with 1 on j-th place and let $b^t = B^t e_j$. Then we can rewrite the formulas for the dynamics of the propensities in vector form with the multiplier expressed in the form of given sum of propensities:

$$q^{t+1} = (q^t + b^t)\frac{C^{t+1}}{Q^t + B^t}\,.$$

Notice that after normalization $Q^{t+1} = C^{t+1}$. Rearranging terms,

$$\frac{q^{t+1}}{Q^{t+1}} = \frac{q^t}{Q^t + B^t} + \frac{b^t}{Q^t + B^t} = \frac{q^t}{Q^t} - \frac{B^t Q^t}{(Q^t + B^t)Q^t} + \frac{b^t}{Q^t + B^t}\,.$$

Since $q^t/Q^t = p^t$, it can be rewritten as

$$p^{t+1} = p^t + a^t(b^t - B^t p^t),$$

where $a^t = (Q^t + B^t)^{-1}$ determines the step size of the process. If payoffs are bounded a^t has the order of $(Q^t)^{-1}$. In the normalization case $Q^t = C^t = Ct^{\nu}$, hence $a^t = O(t^{-\nu})$. In the no normalization case $Q^0 + t \cdot m \leq Q^t \leq Q^0 + t \cdot M$ (where m is the minimal payoff, M is the maximal payoff), hence $a^t = O(t^{-1})$.

Let k be the optimal action, that is the expected payoff to it is highest. To see the expected motion of the process, we calculate $E[p^{t+1}{}_k|p^t]$.

From the expression for probabilities above $E[p^{t+1}{}_k|p^t] = p^t{}_k + E[(a^t(b^t - B^t p^t))_k]$.

$E[\frac{b_k^t - B^t p_k^t}{Q^t + B^t}] = p_k^t \frac{\Phi_k - \Phi_k p_k^t}{Q^t + \Phi_k} + (1 - p_k^t)\frac{-\Phi_{-k} p_k^t}{Q^t + \Phi_{-k}}$, where Φ_k is the expected payoff of action k and Φ_{-k} is the expected payoff if action k is not chosen. Rearranging terms,

$E[\frac{b_k^t - B^t p_k^t}{Q^t + B^t}] = p_k^t (1 - p_k^t)(\frac{\Phi_k}{Q^t + \Phi_k} - \frac{-\Phi_{-k}}{Q^t + \Phi_{-k}})$. Since k is the optimal action,$\Phi_k > \Phi_{-k}$ and $(\frac{\Phi_k}{Q^t + \Phi_k} - \frac{-\Phi_{-k}}{Q^t + \Phi_{-k}}) > 0$. It means that $E[p^{t+1}{}_k|p^t] > p^t{}_k$, that is the process is absolutely expedient (Narendra and Thathachar (1989)) for the optimal action. However, it does not necessarily mean that the optimal action is played in the limit $t \to \infty$ with probability 1 as the following results show.

Proposition 3.1 (Arthur (1993), Posch (1996)) In the normalization case ($C^t = Ct^{\nu}$) if $\nu < 1$ or $\nu = 1$ and $C < m$ the non-optimal action is played in the limit with non-zero probability.

The proof of $\nu < 1$ case is in Arthur (1993), of $\nu = 1$ and $C < m$ case in Posch (1996).

The result states that the process does not necessarily converge to the optimal action for some cases; however for other cases it does.

Proposition 3.2 (Arthur (1993), Posch (1996)) In normalization case if $\nu = 1$ and $C \geq m$ then the probability of the optimal action converges to 1 almost surely[1].

The proof is in the papers referred to.

Proposition 3.3 In the no normalization case the probability of the optimal action converges to 1 almost surely.

We see that in the no normalization case the optimality is guaranteed; however by changing it a little bit this result disappears.

Proposition 3.4 In the no normalization model with forgetting the probability of playing a non-optimal action is not zero in the limit.

The proofs of propositions 3.3 and 3.4 are in Appendix 1.

We can see that the long run results are very different for different specifications of the model. However, for economic relevance we should also look at the speed of convergence and medium term results. It may well be that the optimal learning algorithm is too slow to achieve good results in the medium run and may be inferior in that respect to a non-optimal learning. The next section presents some analytical and simulation results comparing different variations of the model.

The speed of convergence and the probability of convergence to the optimal action
There is a trade-off between the two, i.e. if the speed is high then the scheme is not optimal and if the speed is sufficiently low then it is. In the normalization case, if $\nu < 1$ then the learning is quick and non-optimal, while if $\nu = 1$ the learning is slow enough to achieve optimality. However, even in the non-optimal normalization case $\nu < 1$ by

[1] A sequence of random variables x_t is said to converge almost surely to a random variable x if $Prob[lim_{t\to\infty}|x_t-x|<\varepsilon]=1\ \forall\varepsilon>0$.

changing other parameters one can achieve probability of the optimal action as close to 1 as one desires, that is the scheme is ε-optimal[2] (Narendra and Thathachar (1989)).

We can calculate expected motion of the processes and derive from it the expected rate of convergence. However, to illustrate both issues of non-optimal convergence and the speed thereof, we run simulations for two decision problems:

1) s_1: 4 with probability 1/3, 1 with probability 2/3.
   s_2: 1 with probability 1/3, 4 with probability 2/3.

2) s_1: 2 with probability 1
   s_2: 3 with probability 1
   s_3: 2.5 with probability 1

For the first problem Mean(s_1) = 2, Mean(s_2) = 3, hence the second problem is the certain case of the first one plus an additional action. The second problem is devised to show how convergence slows down with more actions. The sum of initial propensities is chosen quite arbitrarily as 30, which is simply equal $10 \cdot M$ for the second problem and close to estimated in Arthur (1993) from human behavior. Initial propensities q^0 are equal. The simulations were run for the normalization case with $\nu = 0$, for the no normalization case, and for the forgetting case with $\delta = 0.999$.

The question remains what should be taken as the medium run. We have chosen, again quite arbitrarily, 100000 periods as the long run, hence all what is before it is the medium run. A justification for 100000 could be that if you have to chose an action every hour then 100000 hours is approximately 11.5 years, a period that could be roughly considered as with constant environment.

The results of the simulations are reported in Tables 1,2 in Appendix 2 for the problems 1),2) correspondingly. From Table 1 it is seen that if $\nu = 0$ then the dynamics learns very quickly: already at period 300 the probability of playing the optimal action 0.998. Both others modification of the dynamics are slower. However, the dynamics with forgetting

[2] A scheme is said to be ε-optimal if $\forall \varepsilon\ \exists T$ such that for $t > T$ the probability of optimal action $p^t > 1$-ε.

parameter accelerates and at period 100000 it almost catches up the model with $\nu = 0$. All three variations converges to the optimal action in this case; the '$\nu = 0$' case collects greater average payoff due to higher speed.

The second problem shows the non-optimality of the normalization approach. As Table 2 shows, not all of the simulation goes to the optimal action; some of them lock in another one. Again it collects larger average payoff in the beginning. However, two other models find out the optimal action and regain the payoff. The no-normalization model did not catch up with the '$\nu = 0$' case but the forgetting variation of it did.

The main conclusion from the analysis of the decision problems is that in the medium run a model which gives us sufficient speed of convergence while it does not lock in an inferior action is in a sense "optimal". The model with the forgetting parameter seems to satisfy the criterion since it learns slowly in the beginning while the normalized model with $\nu = 0$ gets locked in, and it accelerates after, while the no-normalization model explore its optimality too slowly to catch up. The model with forgetting has a non-zero probability of getting trapped into an inferior action, but the probability is very small. The formal criterion could be the average payoff at period 100000. In the first problem the model with forgetting has slightly less average payoff than the model with normalization but in the second problem the payoff of the model with forgetting much higher. The payoff of the no-normalization model is less in both problems. It should be noticed, however, that the results may depend on the forgetting parameter and the choice of magnitude of initial propensities. Nevertheless, we shall use the model with forgetting later in games.

4 Games

Games provide additional insight to the behavior of the dynamics. The payoffs now depend not only on player’s own action but on actions of the opponents. Since the opponents do not always choose their optimal strategy, it is more difficult for a player to learn his optimal strategy. Normally, if the opponents play the same strategy all the time, the player will eventually learn the best response to this strategy. However, a Nash equilibrium is most likely outcome of the dynamics, since in an equilibrium all players play mutual best responses. We shall start the analysis with *2x2* games and proceed to more complex ones.

2x2 games

The main result about the convergence in games is stated in Posch (1996) and Laslier et al. (1996):

Theorem

(I) If the game has strict Nash equilibria and the dynamics uses normalization with $\nu = 1$ and $C > m$ or no normalization then the algorithm converges to the set of strict Nash equilibria almost surely and all equilibria are attained in the limit with positive probability.

(II) If the game has no strict Nash equilibria then cycling is possible.

The proof for normalization case is in Posch (1996). Since the no-normalization case is essentially equivalent to the normalization case with conditions given in the theorem, the theorem carries over to it.

The theorem does not say anything about selection among strict Nash equilibria. To get some insight we have run a number of simulations for two games with two Nash equilibria. The first game is of pure coordination type, the second is of 'stag-hunt' type.

Game 1	s_1	s_2
s_1	3, 3	1, 1
s_2	1, 1	2, 2

Game 2	s_1	s_2
s_1	3, 3	0.5, 2
s_2	2, 0.5	2, 2

Both games have two strict equilibria $(s_1,s_1),(s_2,s_2)$. In the first game the efficient equilibrium (s_1,s_1) is risk-dominant, in the second one the inefficient equilibrium (s_2,s_2) is risk-dominant. Since the model with forgetting proved to be the most plausible one in the medium run according to the average payoff criterion, we present results only for this model. The results of the simulations are reported in Tables 3,4 in Appendix 3.

Table 3 shows that the efficient risk-dominant equilibrium is almost exclusively chosen by the dynamics. Only one simulation converges to the inferior equilibrium. In Table 4 the inefficient risk-dominant equilibrium is chosen much more often than the efficient one. The results suggest that the risk-dominance, rather than efficiency, is the most important criteria in choosing among equilibria by the dynamics. This result is in line with results in Kandori *et al* (1993) and Young (1993) where different dynamics also favor risk-dominant equilibrium.

Other games

Our main hypothesis is that the more "central" Nash outcome is more likely to observe under the dynamics. 2x2 games are of no use in this respect since they have at most two pure Nash equilibria, hence both are in a sense extreme.

We shall show that if the game possesses several equilibria then the “egalitarian” one, that is the one with more or less equal payoffs for the players, has high chances of being chosen in the medium run. The intuition for this is since the learning is simultaneous for both players they are pressing each other and more likely to end up in a compromise.

To illustrate this point, we show the results of simulations of the dynamics for several games. The first two games, the ultimatum game and the best shot game were analyzed in Roth and Erev (1995). The third game, the oligopoly leadership game, has the same structure as the best shot game but a different set of equilibria and through comparison of these games one can see that the “egalitarian” equilibrium indeed has almost the same probability of being chosen in the long run as the subgame perfect equilibrium. The fourth game considered is a kind of "property" game. This game was analyzed in Young (1996).

The ultimatum game

Two players are to divide 10$. The first player can demand x∈{1,...,9} for himself and, accordingly, leave 10-x to the second player. The second player can then accept or reject this demand. We assume that each player has 9 strategies {1,...,9}. The strategy j for player 1 means that he demands j$ for himself and leaves (10-j)$ for the second player. The strategy k for player 2 means that she accepts any demand ≤ k$, which leaves ≥ (10-k)$ for her, while she rejects any demand >k$. Thus the strategy set for player 2 is restricted to monotone strategies, that is the strategies where the player accepts a demand m > k but rejects a demand m ≤ k are ruled out. The game proceeds as follows. Each player randomly chooses a strategy according to the vector of propensities. Let player 1 choose strategy j and player 2 choose strategy k. If j ≤ k, the demand of player 1 is accepted and the players get j$ and (10-j)$ correspondingly. If j > k the demand is rejected and both get 0$. Then the propensities of played strategies are updated according to the dynamics. Note that the game is essentially a game in extensive form, though we analyzed it in normal form. The first player knows whether his current demand is accepted or rejected but does not know what would happen with greater or smaller demands. Our model allows us to analyze the game since the updating depends only on obtained payoff and does not depend on the payoffs that might be obtained. A pair of strategies (j,j) is an equilibrium ∀ j ∈ {1,...,9}. (9,9) is the subgame perfect equilibrium.

Table 5 in Appendix 3 shows the results of the simulations for 100000th period. It reports the numbers of simulations which have the probability of playing a particular pair of strategies larger than 0.5. Not all the simulation converge to the subgame perfect equilibrium (9,9). At period 100000 the equilibrium (8,8) managed to attract the largest number of simulations, while one simulation converges even to (5,5). The average payoff of about 7 for Player 1 also shows that there are some money left on the table.

Roth and Erev (1995) use experiments for ultimatum game. Their data also favor more “egalitarian” equilibria more than the subgame perfect one. The mean demand in the experiments was between 5 and 6. From the point of view of the dynamics, a high demand can yield 0 if rejected. A modest demand has smaller probability to be rejected because of monotonicity of the strategies of Player 2. More often it yields a positive amount, thus reinforcing itself. Though a high demand reinforces itself better, it happens less often, hence it is not clear a priori whether the subgame perfect equilibrium will be chosen. A model of noisy replicator dynamic and extended discussion about the convergence to a not subgame perfect equilibria in the Ultimatum Game can be found in Binmore *et al* (1995).

The best-shot game

There are 2 players in the game, each has 3 strategies {s_1,s_2,s_3}. The first player plays first, the second player observes first player's move and plays her strategy. Thus the second player has 3^3 = 27 strategies in normal form corresponding to the game. However, the payoffs can be described by the 3×3 bimatrix since the payoffs depend only on strategy of the first player and the answer on this strategy by the second player irrespective to what she would do in response to other strategies of Player 1.

	s_1	s_2	s_3
s_1	0, 0	1.95, 0.31	3.7, 0.42
s_2	0.31, 1.95	0.31, 0.31	2.06, 0.42
s_3	0.42, 3.7	0.42, 2.06	0.42, 0.42

The underlying story for the game is that the players choose their level of provision of a public good. The cost of provision is an increasing function of the quantity provided. The benefit from the good is an increasing function of the maximum between two players' levels of provision. The strategies s_1,s_2,s_3 correspond to the low, medium and high level of provision. The numbers are taken from Roth and Erev (1995) who report about an experiment on an extended version of this game and about simulations of various versions of the dynamics.

Since the first player chooses first, the subgame perfect equilibrium strategy for him is to choose s_1, and for the second player to choose s_3 if the first player played s_1, and to choose s_1 if the first player played s_2 or s_3. Denote this strategy of Player 2 as $s_3s_1s_1$. Then the subgame perfect equilibrium can be denoted by (s_1, $s_3s_1s_1$). The set of all Nash equilibria in the game consists of (s_1, s_3*xx*), where *x* stands for any strategy s_1,s_2,s_3 (the above mentioned subgame perfect one belongs to this subset) and (s_3, $s_1s_1s_1$), (s_3, $s_1s_2s_1$). Note that in the first subset the payoffs are 3.7 for Player 1 and 0.42 for Player 2, while in the second one the payoffs are inverse, 0.42 for Player 1 and 3.7 for Player 2. Thus there is no "egalitarian" equilibrium in the game.

The averages of probabilities of strategies over a hundred simulations are reported in Table 6 in Appendix 3. It is clearly seen that the set containing the subgame perfect equilibrium (s_1, s_3*xx*) is chosen with probability indistinguishably close to 1. The results do not differ much from those reported in Roth and Erev (1995, Table II). Player 1 learns to choose strategy 1 rather quickly. Learning of Player 2 is slower in the beginning since she has much more strategies to choose from but it catches up towards period 100000. The game quickly converges to the set of equilibria (s_1, s_3*xx*). The distribution among *xx* is such that (s_1, $s_3s_1s_1$) is likely outcome though others are also present since the second player does not have much opportunity to learn what she should play in respond to s_2 and s_3.

A possible explanation for finding the subgame perfect equilibrium is that the difference in payoffs between two equilibria is rather high for Player 1 and he learns not to play the other equilibrium quickly. This differs from the ultimatum game where the difference in payoffs from a demand of, say, 6 and 7 is not big if they are both accepted and the difference is high if 7 is rejected. The best-shot game does not possess such "egalitarian" equilibria, hence the convergence is to the subgame perfect one. In the next subsection we consider a game with the same structure as the best-shot game but with an "egalitarian" equilibrium.

The oligopoly leadership game

The structure of this game is as in the best-shot game. There are 2 players, one of which moves first. However, the interpretation of the strategies is different and the payoffs are different too. The players are firms; they choose levels of production. Firm 1 chooses first, Firm 2 follows. The price for the good produced by the firms and therefore the profit received by the firms depend on the aggregate level of production. However, the demand function is not linear as it is usually assumed, hence the price does not depend linearly on the quantity. By choosing the appropriate demand and cost functions (quadratic in the total output and in firm's own output correspondingly) the following payoffs can be obtained.

	s_1	s_2	s_3
s_1	1,1	1,2.3	1,4
s_2	2.3, 1	2,2	0.6, 0.3
s_3	4, 1	0.3, 0.6	0, 0

(the magnitude of the payoffs has the same order as in the best-shot game)

Interpreting strategies s_1,s_2,s_3 as low, medium and high level of production correspondingly, one can see that the subgame perfect equilibrium strategy for Firm 1 is to capture the market by choosing high level of production s_3. Firm 2 is then left with a small remaining fraction of the market. The payoffs are 4 for Firm 1 and 1 for Firm 2. These strategies correspond to the subset of the Nash equilibria of the game (s_3, xxs_1), where again x denotes any of the strategies s_1,s_2,s_3. The equilibrium (s_3, $s_3s_2s_1$), belonging to this subset, is subgame perfect. The game, however, possesses two other types of Nash equilibria. One of them, as in the best-shot game, is the inversion of the subgame perfect equilibrium, namely (s_1,$s_3s_3s_2$) and (s_1,$s_3s_3s_3$) with payoffs 1 for Firm 1 and 4 for Firm 2. The new type of Nash equilibria is the "egalitarian" one: (s_2, xs_2s_2) and (s_2, xs_2s_3), where payoffs are 2 for both firms. We argue that this equilibrium does not have much fewer chances of being chosen in the medium run than the subgame perfect equilibrium. Table 7 of Appendix 3 shows the averages of probabilities of strategies over a hundred simulations.

From the table it can be seen that the "egalitarian" strategy (s_2,s_2) is learnt faster than the subgame perfect one. At period 1000 the equilibrium (s_2,s_2) has larger probability of being played while with time the subgame perfect one regain its strength. At period 10000 the probabilities of playing both strategies s_2 and s_3 for Player 1 are equal and at period 100000 the subgame perfect equilibrium strategy s_3 finally gets larger probability. If the "egalitarian" one fails to gain the lion's share in probability distribution in the beginning then the subgame perfect equilibrium can regain the probability in later periods. However, the “egalitarian” equilibrium succeeds in being selected in about one third of the simulations. 23 simulations have the probability of playing this equilibrium after 100000 periods larger than 0.5. For the set of equilibria containing subgame perfect one the number of such simulations are 45. This shows that the "egalitarian" equilibrium has not much fewer chances to be selected in the medium run. And since the subgame perfect equilibrium (s_3, $s_3s_2s_1$) also has s_2 as response to s_2, the early recognition of the "egalitarian" equilibrium helps to discriminate the subgame perfect equilibrium among the set (s_3, xxs_1). The dynamics converges to the subgame perfect equilibrium in 2/3 cases when it converges to this set.

The oligopoly leadership game, as well as the ultimatum game, possesses an equilibrium that lies between two extreme equilibria where almost all payoff goes to one player. In this “intermediate” equilibrium the payoffs are divided more or less equally. This is why we call it the “egalitarian” equilibrium. In distinction to these two games, the best-shot game does not possess such an equilibrium and the subgame perfect one gains dominance very easily. The implication of this observation is that the selection of equilibria by the dynamics presented in the medium run may depend mainly on the structure of the set of Nash equilibria of the game. If the game has only “extreme” equilibria, the players (or at least one of them) quickly learn to play the subgame perfect one. A possible explanation might be that it is too risky for the other player to insist on the other extreme, therefore she has to allow the unfavorable for her subgame perfect outcome. However, in the presence of an equilibrium between those extremes, it is not easy for them to find out the subgame perfect one. My conjecture is that this is an inherent feature of the dynamics that an equilibrium between extreme equilibria has rather big chances to attract the process in the medium run, though in the long run a subgame perfect equilibrium prevails. However, we have no proof of that at hand. More work has to be done to derive such a proof.

The "property" game

In analysis of above games the "egalitarian" equilibrium was not subgame perfect while there was a subgame perfect equilibrium which regains the probability though the "egalitarian" equilibrium did not perform badly. In this subsection we consider a game where all three equilibria seems equally plausible and we cannot give preference to one of them from conventional refinement criteria. The game has following payoff matrix

	s_1	s_2	s_3
s_1	1, 1	1, 1	4, 2
s_2	1, 1	3, 3	1, 1
s_3	2, 4	1, 1	1, 1

The story is that suppose two people divide a property of 6 (in a case of divorce, for example). They can agree on three outcomes (4,2),(3,3)(2,4). In the case of disagreement they both get payoff of 1. The equilibria in the game are (s_1,s_3),(s_2,s_2),(s_3,s_1). The second one is "egalitarian" one in the sense that it has equal distribution of payoffs while two others has one player getting more than the other.

The results of the simulations are reported in Table 8 of Appendix 3. The simulations show that starting from equal initial propensities for all three strategies, the middle equilibrium is chosen more often than the other two though all three equilibria are chosen in the long run. The probability of playing the pair of strategies (s_2,s_2) is larger than the probability of playing the other two equilibria and in larger number of simulation the dynamics converges to the middle equilibrium rather than to the other two as it is seen from "Modes" column. The explanation for this is that if we calculate the expected probabilities of playing the equilibria for period 2 (for period 1 they are all equal since the initial propensities are equal) then the expected probability of playing the middle equilibrium is slightly higher than for other two equilibria. Hence in expectations the dynamics should go to the "egalitarian" equilibrium. Due to the noise other equilibria also have a chance to be selected and it is confirmed from simulations. Our hypothesis about the likeliness of the "egalitarian" equilibrium in some games is supported too. Another supporting paper on this subject is Young (1996), where the finding for pure coordination games is similar under a different dynamics.

5 Conclusion

The simple dynamics captures certain aspects of human learning such as the “Law of Effect” and the “Power Law of Practice”. Hence it may describe the behavior of humans in decision problems and games. The analysis for the case of one-player decision problem shows that the dynamics selects the optimal strategy in the long run despite the fact that the non-optimal strategy is also reinforcing. However, as it was shown, the speed of convergence is slow. The speed of convergence for the dynamics seems to be dependent on the difference in payoffs between the optimal and non-optimal strategies and on the number of strategies. Since the speed is too slow to produce satisfactory results for the medium run, we modified the dynamics such as there is a small probability of locking in an inferior action but the speed of convergence is improved considerably. The speed of convergence play a role for most real games. In principle chess can be solved explicitly but we do not have all the time in the world just to play chess. Therefore we must sometimes admit having non-optimality in moves. According to the average payoff criterion the best model given the tradeoff between speed and convergence seems to be the model with forgetting parameter.

Application of the dynamics for games yields some interesting observations. Though it seems that in the long run the dynamics will eventually converge to the subgame perfect equilibrium, in the medium run it sometimes fails to find an equilibrium at all (as in the ultimatum game) or it converges to an equilibrium that is not subgame perfect. This equilibrium often has the feature to be “in the middle” of the set of Nash equilibria for the game (as in the oligopoly leadership game and in the property game) and gives more or less equal payoffs for both players. Such an equilibrium can be called “egalitarian”. In the absence of a suitable “egalitarian” equilibrium the dynamics finds the subgame perfect equilibrium and rather quickly.

The paper analyzes the dynamics formally only for the case of one-player decision problem and gives some examples how it can perform in games. The direction for further research are establishing the analytical results for games since simulation studies can give only partial insight onto the problem. More games with various properties should be analyzed to establish certain features of the dynamics. Some extensions of the basic model are possible. An interesting extension is to games with negative payoffs or, equivalently, a dynamics where not all strategy profiles are reinforcing. If the players know the game or the decision problem, the assumption that all strategies are reinforcing seems not very natural since the players may hope to obtain the highest payoffs. Also the assumption of the fixed aspiration level can be relaxed though it may bring about the phenomenon of “probability matching”, which is absent in the present model.

Appendix 1

Proposition 3.3 In the no-normalization case the probability of playing the optimal strategy converges to 1 almost surely.

Proof Notice that $Q^0 + t \cdot m \le Q^t \le Q^0 + t \cdot M$ in this case. Then $c_1 t \le Q^t \le c_2 t$ with $m < c_1$. Then the dynamic is equivalent to the normalization case with $\nu = 1$. According to Proposition 3.2 the process then converges to the optimal action.

Proposition 3.4 In the no-normalization case with forgetting the probability of playing non-optimal action in positive in the limit $t \to \infty$.

Proof The proof is along the lines of similar proof for normalization case given in Arthur (1993). Notice that now we have $Q^0 + m\, \Sigma_{\tau=1}^{t}\, \delta^\tau \le Q^t \le Q^0 + M\, \Sigma_{\tau=1}^{t}\, \delta^\tau$. Calculating the sum of geometric series we get $Q^0 + m\dfrac{1-\delta^t}{1-\delta} \le Q^t \le Q^0 + M\dfrac{1-\delta^t}{1-\delta}$, that is Q^t has the order of $(1\text{-}\delta^t)$.
To proof the non-optimality consider the event that an inferior action j is triggered from time t. Denote this event as D_t. We need to show that $Prob\{D_t\} = \Pi_t^\infty p_j^t > 0$. Let $a_j^t = 1 - p_j^t$. Since $0 < a_j^t < 1$ the convergence of $\Pi_t^\infty(1 - a_j^t)$ is necessary from the convergence of Σa_j^t. From our dynamic equation for probabilities $p^{t+1} = p^t + a^t(b^t - B^t p^t)$, the dynamics for action j when it is played is written as

$p_j^{t+1} = p_j^t + \dfrac{B^t(1-p_j^t)}{Q^t + B^t}$. Then $a_j^{t+1} = a_j^t(1 - \dfrac{B^t}{Q^t + B^t})$. The ratio of successive terms is then $\dfrac{a_j^{t+1}}{a_j^t} = (1 - \dfrac{B^t}{Q^t + B^t})$. Given that Q^t has the order of $(1\text{-}\delta^t)$ we can rewrite it as $\dfrac{a_j^{t+1}}{a_j^t} = (1 - \dfrac{B^t}{c_1 + c_2(1-\delta^t) + B^t}) < (1 - \dfrac{B^t}{c_1 + c_2 + B^t})$. It means that a_j^t decreases faster than a geometric series and therefore converges. Therefore the probability of playing the inferior action j from time t is positive and the proposition is proven.

Appendix 2

In the row labeled 'Probability' the probability of the optimal action is given; in the row labeled 'Mode' the number of simulations where the probability of the optimal action is highest and larger than 0.5; the row labeled 'Av.Payoff' gives the average payoff up to time t. The column '$\nu = 0$' is for normalization case with $\nu = 0$, the column 'no-norm' for no normalization case, and the column 'forgetting' for the case with $\delta = 0.999$.

Table 1. Averages of the probability of playing the optimal strategy and the numbers of simulations where the mode is the optimal strategy and the average payoff for problem 1)

Time		ν=0	no-norm	forgetting
300	Probability	0.998	0.764	0.768
	Mode	100	99	97
	Av. Payoff	2.896	2.706	2.701
10000	Probability	1	0.914	0.992
	Mode	100	100	100
	Av. Payoff	2.997	2.879	2.937
100000	Probability	1	0.959	1
	Mode	100	100	100
	Av.Payoffs	2.999	2.939	2.994

100 simulations, q^0=(15,15)

Table 2. Averages of the probability of playing the optimal strategy and the numbers of simulations where the mode is the optimal strategy and the average payoff for problem 2)

Time		ν=0	no-norm	forgetting
300	Probability	0.841	0.511	0.532
	Mode	85	75	76
	Av. Payoff	2.812	2.630	2.634
10000	Probability	0.860	0.687	0.904
	Mode	86	90	100
	Av. Payoff	2.926	2.777	2.863
100000	Probability	0.860	0.774	1
	Mode	86	97	100
	Av.Payoffs	2.930	2.846	2.983

100 simulations, q^0=(10,10,10)

Appendix 3

The column 'Probabilities' reports the probabilities of equilibria, (s_1,s_1),(s_2,s_2), respectively. The column 'Modes' reports the number of simulations where the probability of the equilibrium is larger than 0.5. The column 'Average Payoffs' gives average payoffs up to time t for players 1,2 respectively.

Table 3. Averages of the probability of playing the equilibrium, the numbers of simulations where the mode is on the eqilibrium and the average payoffs for Game 1.

Time	Probabilities	Modes	Average Payoffs
300	0.638, 0.062	79, 1	2.136, 2.136
1000	0.816, 0.023	93, 1	2.410, 2.410
10000	0.990, 0.010	99, 1	2.882, 2.882
100000	0.990, 0.010	99, 1	2.979, 2.979

100 simulations, q^0=(15,15)

Table 4. Averages of the probability of playing the equilibrium, the numbers of simulations where the mode is on the equilibrium and the average payoffs for Game 2.

Time	Probabilities	Modes	Average Payoffs
300	0.150, 0.482	7, 52	1.866, 1.862
1000	0.120, 0.620	9, 69	1.893, 1.887
10000	0.106, 0.877	11, 88	2.028, 2.027
100000	0.120, 0.880	12, 88	2.110, 2.109

100 simulations, q^0=(15,15)

For the ultimatum game we report the results only for 100000th period. The number in cell (i,j) represents the number of simulations where the average probability of playing pair (i,j) is highest and larger than 0.5.

Table 5. Numbers of simulations with mode on pairs of strategies and the average payoffs for Ultimatum game.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	4	1	2	0
7	0	0	0	0	0	0	10	6	12
8	0	0	0	0	0	0	0	18	15
9	0	0	0	0	0	0	0	0	11

100 simulations, q^0=(10,10,10,10,10,10,10,10,10)

Average payoffs: Player 1: 7.141
Player 2: 2.789

For the best shot and oligopoly leadership games the results are reported in following manner. For Player 1 the probabilities of playing each of the three strategies are reported. For Player 2 the probabilities of answers to given strategy of Player 1 are reported.

Table 6. Averages of probabilities for the best-shot game

Time		s_1	s_2	s_3	Av.Payoffs
1000	Player 1	0.951	0.036	0.009	2.151
	Player 2 on s_1	0.127	0.331	0.538	0.506
10000	Player 1	0.999	0.000	0.000	3.195
	Player 2 on s_1	0.000	0.058	0.941	0.411
100000	Player 1	1	0	0	3.645
	Player 2 on s_1	0	0	1	0.419

100 simulations; q^0=(10,10,10) for both players

Table 7. Averages of probabilities for the oligopoly leadership game

Time		s_1	s_2	s_3	Av.Payoffs
1000	Player 1	0.052	0.567	0.380	1.975
	Player 2 on s_2	0.164	0.743	0.092	1.445
	on s_3	0.467	0.306	0.226	
10000	Player 1	0.000	0.500	0.500	2.621
	Player 2 on s_2	0.068	0.893	0.037	1.489
	on s_3	0.567	0.220	0.175	
100000	Player 1	0	0.420	0.580	3.079
	Player 2 on s_2	0.069	0.897	0.034	1.436
	on s_3	0.623	0.176	0.167	

100 simulations; q^0=(10,10,10) for both players

For the Property game the set of equilibria is (s_1,s_3),(s_2,s_2),(s_3,s_1). The numbers in the table are given correspondingly in this order of equilibria. 'Probabilities' are the probabilities of the equilibria, 'Modes' are the numbers of simulations where the probability of playing an equilibrium more than 0.5. 'Average Payoffs' gives average payoffs up to time t for players 1,2 correspondingly.

Table 8. Averages of the probability of playing the equilibrium, the numbers of simulations where the mode is on the equilibrium and the average payoffs for the "property" game.

Time	Probabilities	Modes	Average Payoffs
300	0.113, 0.191, 0.143	2, 12, 4	1.777, 1.831
1000	0.133, 0.260, 0.175	8, 29, 18	1.920, 1.993
10000	0.289, 0.406, 0.284	30, 41, 29	2.612, 2.651
100000	0.300, 0.410, 0.290	30, 41, 29	2.969, 2.955

100 simulation, q^0=(10,10,10)

References

Arthur W.B. (1993) "On Designing Economic Agents that Behave Like Human Agents", *Journal of Evolutionary Economics*, 3, 1-22.

Binmore K., Gale J., Samuelson L. (1995) "Learning to Be Imperfect: The Ultimatum Game", *Games and Economic Behavior*, 8, 56-90.

Binmore K., Swierzbinski J., Proulx C. (1996) "Does Minimax Work? An Experimental Study", mimeo, University College London.

Bendor J., Mookherjee D., Ray D. (1994) "Aspirations, Adaptive Learning and Cooperation in Repeated Games", mimeo, Boston University.

Börgers T. and Sarin R. (1996), "Naive Reinforcement Learning with Endogenous Aspiration", mimeo, University College London.

Herrnstein R.J., Prelec D. (1991) "Melioration: A Theory of Distributed Choice", *Journal of Economic Perspectives*, 5, No.3.

Kandori M, Mailath G.J., Rob R. (1993), "Learning, Mutation, and Long Run Equilibria in Games", *Econometrica*, 61, 29-56

Laslier J.F., Topol R., Walliser B. (1996) "A Behavioral Learning Process in Games", CERAS, Document de travail, 96-03.

Narendra K., Thathachar M.A.L. (1989) *Learning Automata: An Introduction*, Prentice Hall International.

Posch M. (1996) "Cycling in a Stochastic Learning Algorithm for Normal Form Games", mimeo, Institute for Mathematics of Vienna University.

Roth A. and Erev I. (1995) "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Run", *Games and Economic Behavior*, 8, 164-212.

Young P.H. (1993) "The Evolution of Conventions", *Econometrica*, 61, 57-84

Young P.H. (1996) "Social Coordination and Social Change", IIASA Working Paper, WP-96-32