A reinforcement learning process in extensive form games

Jean-François Laslier
CNRS and Laboratoire d'Econométrie
de l'Ecole Polytechnique, Paris.

Bernard Walliser*
CERAS, Ecole Nationale des Ponts et Chaussées, Paris.

October 11, 2004

Abstract

The CPR ("cumulative proportional reinforcement") learning rule stipulates that an agent chooses a move with a probability proportional to the cumulative payoff she obtained in the past with that move. Previously considered for strategies in normal form games (Laslier, Topol and Walliser, *Games and Econ. Behav.*, 2001), the CPR rule is here adapted for actions in perfect information extensive form games. The paper shows that the action-based CPR process converges with probability one to the (unique) subgame perfect equilibrium.

Key Words: learning, Polya process, reinforcement, subgame perfect equilibrium.

1 Introduction

Reinforcement learning has a long history spanning from Animal Psychology to Artificial Intelligence (see Sutton and Barto, 1998 for more details). In Games and Economics, several reinforcement rules have been introduced (see Fudenberg and Levine, 1998). Some rules were considered in order to explain the choices made by individuals in laboratory experiments when observed at intermediate stages in interactive situations (Roth and Erev, 1995; Camerer, 2003). The same and other rules were more extensively

*ENPC, 28 rue des Saints Pères, 75007 Paris, France. Phone: (33) 1 44 58 28 72. Fax: (33) 1 44 58 28 80. walliser@mail.enpc.fr

studied in the literature from a theoretical point of view, especially with respect to their asymptotic behavior (see Sarin and Vahid, 1999, 2001). The *CPR* (Cumulative Proportional Reinforcement) *rule*, also called the 'basic' reinforcement rule, is perhaps the simplest mathematical model of reinforcement learning. It associates a 'valuation rule' and a 'decision rule' with each period. The former states that the player computes for each action an index equal to its past cumulative payoff. The latter states that the player plays an action with a probability proportional to that index.

A preceding paper (Laslier, Topol and Walliser, 2001, henceforth LTW) studied the convergence properties, in repeated finite two-player normal form games, of the learning process where each player uses the CPR rule. On the one hand, LTW proved that the process converges with positive probability toward any strict pure Nash equilibrium. On the other hand, LTW proved that the process converges with zero probability toward any non Nash state as well as toward some mixed Nash equilibria (duly characterized). Lastly, for some non strict Nash equilibria, convergence could not be elucidated. Note that, for a single decision-maker under risk, LTW showed that the process converges toward the expected payoff maximizing action(s). Related theoretical results are given in Hopkins (2002), Beggs (2002) and Ianni (2002).

The present paper considers repeated finite extensive form games with perfect information, assumed to have generic payoffs (no ties for any player). Passing from normal to extensive form games, the CPR principle can be adapted in two ways. With the *s-CPR* (strategy-based CPR) *rule*, each player applies the CPR rule to its "strategies", a strategy being defined, as usual in game theory, as a set of intended conditional actions at each node of the game tree. With the *a-CPR* (action-based CPR) *rule*, each player applies the CPR rule to each action at each node in the game tree when reached (although the player only receives the payoff at the end of the path). The main result of the present paper is that the *a-CPR* process converges with probability 1 toward the unique subgame perfect equilibrium path (obtained by backward induction). Moreover, for any learning process, one may distinguish between 'convergence in actions' of the moves which are selected and 'convergence in values' of the indices which are computed. The *a-CPR* process converges in actions, even if it may not converge in values. However, the perfect equilibrium values (i.e. the payoffs that the players reach at each node, when at the subgame perfect equilibrium) may be asymptotically recovered by dividing the cumulative index by the number of trials of an action.

A similar problem was already studied in the literature, with a simi-

lar result, but with different reinforcement learning rules. Jehiel and Samet (2000) considered the ε-greedy rule: the valuation rule asserts that the player computes for each action an index equal to its past average payoff; the decision rule asserts that she plays, with some given probability, the action maximizing the index and, with the complementary probability, a random -uniformly distributed- action. Since some randomness is present until its end, the process converges in values toward the subgame perfect equilibrium values, but the actions only approach the subgame perfect equilibrium actions. More precisely, the sub-sequences of index-maximizing actions converges toward the subgame perfect equilibrium, but random actions continue to be played with a fixed positive probability. Pak (2001) considered a more sophisticated rule: the valuation rule associates to each action a stochastic index equal either to its past payoffs (with a probability proportional to their frequency) or to some random values (with a probability decreasing with the number of occurences of that action); the decision rule states that she chooses the maximizing action. Here, the process converges (for even a larger class of rules containing the preceding one) toward the subgame perfect equilibrium actions, but not toward the equilibrium values, even if they are recovered by taking the expected value of the random variable.

In both cases, the learning rule reflects a trade-off faced by each player between exploration and exploitation, which takes place in a non stationary context. Exploitation is expressed by the decision rule, which is close to a maximizing rule, and by the valuation rule, which is a nearly averaging rule. Exploration is expressed by a random perturbation, either on the decision rule (first case) or on the valuation rule (second case). In addition, the perturbation is constant in the first case, and decreasing in the second case. Conversely, the exploration component of the CPR rule is directly integrated in a non maximizing decision rule (allowing for mutations). Its exploitation component is associated with a cumulative valuation rule (since it creates a feedback effect on the best actions). Hence, the trade-off is endogenous, leading to more exploration at the beginning of the process (since the initial indices are uniform) and to more exploitation in the latter stages if convergence occurs (exploration decreases to zero but remains active till the end).

The paper first presents the game assumptions and the two variants of the CPR learning process. Then the main convergence result concerning the action-based CPR rule is proven. Finally, the convergence properties of the action-based and the strategy-based processes are compared using an example.

2 Game and learning assumptions

Consider a perfect information stage game defined by a finite tree formed by a set I of players, a set N of non terminal nodes (including the root node r), a set M of terminal nodes, a set A of edges (actions). For each node n, call $I(n)$ the player who has the move at that node, $A(n)$ the set of actions at her disposal, and $G(n)$ the subgame starting at the node. For each node n, except for r, call $B(n)$ the unique node leading to it. For each terminal node m, call $u(m)$ the payoff vector, assumed to be strictly positive: $\forall i \in I, \forall m \in M, u_i(m) > 0$. Denote by $\underline{u_i} = \min\{u_i(m) : m \in M\} > 0$ the smallest payoff player i can get from any terminal node and by $\overline{u}_i = \max \{u_i(m) : m \in M\}$ the largest one.

The game is said to be "generic" if, for any player, her payoffs at different terminal nodes differ: *if* $m \neq m' \in M$, then $\forall i \in I,\ u_i(m) \neq u_i(m')$. The game is said to be "weakly generic" under the condition that, if for one player, her payoffs at two different terminal nodes are identical, then this occurs for all players : *if* $\exists\ i \in I$ *and* $m, m' \in M$ *such that* $u_i(m) = u_i(m')$, *then* $\forall j \in I,\ u_j(m) = u_j(m')$. In this paper, we only consider generic games, although some results can easily be extended to weakly generic games.

A pure strategy s_i of player i specifies an action played at each node of player i (i. e. each node $n \in N$ such that $I(n) = i$). A player's mixed strategy specifies a probability distribution over all her pure strategies. A player's behavioral strategy specifies, for each node n such that $I(n) = i$, a probability distribution on the actions available to player i at this node. The combination of strategies $(s_i)_{i \in I}$ played by all players is denoted s. A generic game has a unique subgame perfect equilibrium (SPE) s^*, obtained by a backward induction procedure. To each node n, the equilibrium strategy vector s^* associates an action $a^*(n)$ for player $I(n)$ and a unique terminal node $m^*(n)$. The payoff obtained by player i in the subgame $G(n)$ is denoted $u_i^*(n) = u_i(m^*(n))$.

The stage game is now played an infinite number of times, labelled by t. At each period t, a path of play is followed; denote $\delta_t(a) = 1$ when the path reached action a and $\delta_t(a) = 0$ otherwise. Each player i knows which nodes she successively reached and observes the payoff $u_t(i)$ she gets at the end. After t periods, call $N_t(a)$ the number of times that action a was used. The a-CPR ("action-based cumulative proportional reinforcement") rule is defined not on mixed strategies, but on behavioral strategies. It is composed of two parts:

- the valuation rule states that, at the end of each period t, for each node n such that $i = I(n)$, each action a such as $a \in A(n)$ is associated with an

index $v_t(a)$ which is the cumulative payoff obtained by that action in the past (each payoff obtained at the end of a path is allocated simultaneously to all actions in the path): $v_t(a) = \Sigma_{\tau \in [0,t-1]} u_\tau(i)\ \delta_\tau(a)$; the initial valuation is $v_0(a)$.

-the decision rule states that, at each period t, if node n is attained, the player chooses an action $a \in A(n)$ according to a probability distribution p_t proportional to the index vector v_t: $p_t(a) = v_t(a)/\Sigma_{b \in A(n)} v_t(b)$.

Of course, the extensive form stage game can be tranformed into a normal form stage game by introducing the notion of a strategy. Notice that a generic extensive form game does not generally lead to a generic normal form game (i.e. a game in which, for each player, all payoffs are different) but to a weakly generic normal form game (i.e. a game in which, if for some player, two issues yield the same payoff, then, for all players the payoffs at these two issues are equal too). Using the CPR rule on that normal form defines the s-CPR ("strategy-based cumulative proportional reinforcement") rule:

-the valuation rule states that, at the end of period t, each strategy s is associated with an index $v_t(s)$ which is the cumulative payoff obtained by that strategy in the past;

-the decision rule states that, at each period, each player chooses a strategy s among the available strategies, with a probability $p_t(s)$ proportional to its index $v_t(s)$.

3 Convergence results

Considering the a-CPR process, a necessary condition for sufficient exploration is that the process visits each node an infinite number of times. This condition is ensured by the first result:

Lemma 1 *With the a-CPR rule applied to a generic perfect information extensive form game, each node is almost surely reached an infinite number of times.*

Proof: First, the following statement is proven: for any node n, if n is reached an infinite number of times, then each action $a \in A(n)$ is chosen an infinite number of times. For each $a \in A(n)$, the payoff that player $i = I(n)$ obtains after choosing a is in some positive interval $[\underline{u}_i(a), \overline{u}_i(a)]$. The cumulative payoff associated to an action other than a is thus bounded above by an affine function of time, and the probability of playing action a is bounded below by the inverse of an affine function of time. Therefore, the

argument of the proof of Proposition 1 in LTW applies. Second, since the initial node is obviously reached an infinite number of times, by successive steps in the finite tree, such is the case for all nodes. **QED**.

Lemma 1 ensures that each path (including the SPE path) is played with probability 1 an infinite number of times. The second result shows that the SPE path is played infinitely more often than any other path :

Theorem 1 *With the a-CPR rule applied to a generic perfect information extensive form game, the probability of playing the SPE path at time t converges almost surely to* 1.

Proof:
(a) Notation and argument. Let (Ω, π) be a probability space on which the repeated play of the game following the a-CPR rule is realized: Ω is the set of all possible complete histories of the repeated game; π is the probability distribution induced by the stochastic CPR rule on this set. An event happens "almost surely" if the π-probability that this event does not happen is equal to zero. A draw $\omega \in \Omega$ defines the path $h(t, \omega)$ at date t and the history $H(t, \omega) = (h(\tau, \omega))_{1 \leq \tau \leq t-1}$ up to date t. The probability of playing any path at date t is a function of $H(t, \omega)$ which we simply see as a function of t and ω. The probability of playing the subgame perfect equilibrium path at date t from a non-terminal node $n \in N$ is denoted by $q_t(n, \omega)$. What is to be proved is that, for all n, π-almost surely, $q_t(n)$ tends to 1 when t tends to infinity.

By definition of the a-CPR process, for any draw ω, $q_t(n, \omega)$ is the product of the probabilities $p_t(a^*(n'), \omega)$ of choosing the perfect equilibrium action at all the non-terminal nodes n' (including n) on the equilibrium path in the subgame $G(n)$. In other terms, $q_t(n, \omega) = p_t(a^*(n), \omega)\, q_t(n', \omega)$, where n' is the node resulting from $a^*(n)$. Hence, the proof proceeds by induction on subgames.
(b) Initial step. For the initial induction step, consider any node $\widetilde{n}$ which is followed only by terminal nodes. Here, $q_t(\widetilde{n}, \omega) = p_t(a^*(\widetilde{n}), \omega)$. Player $I(\widetilde{n})$ faces a choice between actions in $A(\widetilde{n})$ among which $a^*(\widetilde{n})$ is the maximizing one. According to the lemma, π-almost surely, the process reaches node $\widetilde{n}$ an infinite number of times t_θ, $\theta = 1, 2, ...$; one may number these (random) dates by the new index θ. By definition of the a-CPR rule, $p_t(a^*(\widetilde{n}), \omega)$ is only modified when node $\widetilde{n}$ is reached. Thus, slightly abusing notation, the probability of playing $a^*(\widetilde{n})$ can be written $p_\theta(a^*(\widetilde{n}), \omega)$. Consider now the

event:

$$F(\widetilde{n}) = \left\{ \omega \in \Omega \ / \ \lim_{\theta \to \infty} \ p_\theta(a^*(\widetilde{n}), \omega) = 1 \right\}.$$

According to Proposition 4 in LTW applied to time scale θ, the process converges almost surely towards the maximizing action:

$$\pi(F(\widetilde{n})) = 1.$$

In particular, for almost all $\omega \in \Omega$, and for any $\varepsilon > 0$, there exists Θ such that, if $\theta \geq \Theta$, then $p_\theta(a^*(\widetilde{n}), \omega) \geq 1 - \varepsilon$. We already noted that $p_t(a^*(\widetilde{n}), \omega)$ is only modified at dates θ (when node n is reached). Hence, there exists T such that, if $t \geq T$, then $p_t(a^*(\widetilde{n}), \omega) \geq 1 - \varepsilon$. This proves that, π-almost surely:

$$\lim_{t \to \infty} p_t(a^*(\widetilde{n}), \omega) = \lim_{t \to \infty} q_t(\widetilde{n}, \omega) = 1.$$

(c) Induction. For the general induction step, consider any non-terminal node n. Player $i = I(n)$ faces a choice between actions in $A(n)$, but the payoff to such an action is now random. Label $a_0, a_1, ..., a_k, ...$ the actions in $A(n)$, any action a_k leading to node n_k, with $a_0 = a^*(n)$ the perfect equilibrium action. Each n_k is the root of a subgame $G(n_k)$, hence defines, given the history, a lottery $L_t(n_k)$ at time t for player i. However, the probabilities involved in $L_t(n_k)$ are not fixed and Proposition 4 in LTW is no longer directly applicable. It is necessary to introduce auxiliary lotteries with fixed probabilities. These lotteries are denoted by $\mathbf{L}(\cdot)$. They depend on action a_k being the SPE action or not:

- if $k = 0$, the lottery $\mathbf{L}(n_0)$ gives to player i the equilibrium payoff $u_i(n_0) = u_i^*(\widetilde{n})$ with probability $1 - \varepsilon_0$ and payoff $\underline{u_i}$ (the smallest payoff player i can get) with probability ε_0;

- if $k \neq 0$, the lottery $\mathbf{L}(n_k)$ gives payoff $u_i(n_k) = u_i^*(n_k)$ with probability $1 - \varepsilon_k$ and payoff $\overline{u}_i$ (the largest payoff player i can get) with probability ε_k.

By definition of a subgame perfect equilibrium, $u_i(n_0) \geq u_i(n_k)$ for all k, and by the genericity hypothesis, each inequality is strict for $k \neq 0$. Consider the auxiliary 1-player CPR process defined by the lotteries $\mathbf{L}(n_k)$. For ε_0 and ε_k small enough, the expected payoff in $\mathbf{L}(n_k)$ is smaller than the one in $\mathbf{L}(n_0)$, thus Proposition 4 in LTW applies, and the player chooses asymptotically lottery $\mathbf{L}(n_0)$. The probability of choosing action a_0, which is denoted by $\mathbf{p}_t(a_0)$, tends almost surely to 1.

Now compare, starting at n, the auxiliary fixed-lottery a-CPR process with the true a-CPR process. By the induction hypothesis, in the true process, there exists T_k such that for $t > T_k$, the probability $q_t(n_k)$ is almost

surely greater than $1-\varepsilon_k$. Given that action a_k is played, the probability of receiving $u_i(n_k)$ is larger in the true process than in the auxiliary one. Thus, one can define the auxiliary and true processes on the same space (Ω, π) in such a way that, for all $\omega \in \Omega$ such that a_k is played, $u_i(n_k)$ is obtained in the true process whenever it is obtained in the auxiliary one. The comparative payoffs, π-almost surely, are the following:

- if a_0 is played , then the payoff in the auxiliary process ($u_i(n_0)$ or $\underline{u}_i$) is lower than the payoff in the true one;

- if $a_k \neq a_0$ is played, then the payoff in the auxiliary process ($u_i(n_k)$ or $\overline{u}_i$) is larger than the payoff in the true one.

It follows that, π-almost surely, the cumulative payoff $v_t(a_0)$ is larger in the true process than in the auxiliary one while $v_t(a_k)$ is lower (for $k \neq 0$). Consider now the decision rule at node n. It states that the probability of choosing action a_k is proportional to $v_t(a_k)$. It follows that, almost surely, a_0 is played more often in the true process: $p_t(a_0) \geq \mathbf{p}_t(a_0)$. Since $\mathbf{p}_t(a_0)$ tends to 1, so does $p_t(a_0)$. The probability of playing the equilibrium path from n is $q_t(n) = p_t(a_0)q_t(n_0)$ and since $q_t(n_0)$ tends to 1 by the induction hypothesis, it is the same for $q_t(n)$. **QED.**

4 Concluding remarks

For a generic extensive form game, one wishes to compare the respective effects of the s-CPR and the a-CPR processes. To highlight the differences, consider an example, similar to the chain-store paradox, and depicted in extensive and normal form (Figure 1 and matrix 1). In such a 2×2 game, the strategies of a player coincide with his actions. In this game, CC is the subgame perfect equilibrium and it is a strict pure Nash equilibrium. SS is another pure Nash equilibrium, but it is not strict. Action S for the second player is moreover weakly dominated.

	C	S
C	$(3,3)^N$	(1,1)
S	(2,5)	$(2,5)^N$

(1)

Normal form

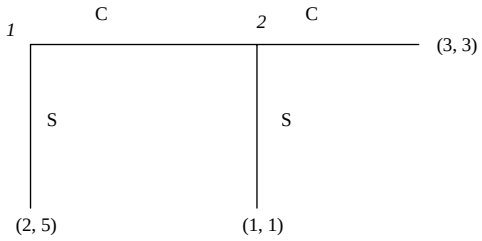

Figure 1: Extensive form

Simulations of the game are achieved for the a-CPR and s-CPR processes (see Figure 2). In each diagram, a point represents a probability distribution on actions (or strategies) for each player. The upper left corner of each diagram corresponds to the upper left corner of the payoff matrix, that is the subgame perfect equilibrium CC, and similarly for the other issues. The two left diagrams show typical paths starting from an initial situation where each action for each player is played with equal probability (initial valuations $v_0(a)$ all set to 10) and for 5,000 iterations. The two right diagrams show typical paths starting from an initial situation close to the SS equilibrium (intial valuations $v_0(a)$ set to 90 and 10) and for 50,000 iterations. On all these diagrams, one can see that both processes are moving faster (and thus more randomly) at the beginning, and are moving slower and more smoothly with time.

Looking at the left side diagrams, one can see the convergence of both processes toward the subgame perfect equilibrium CC. For the a-CPR process, this is in accordance with the theorem proved in this paper. For the s-CPR process, the results in LTW do not apply since the normal form game is not generic; however one may conjecture that the s-CPR process converges with a positive probability toward CC. On these same diagrams, one can also notice that the a-CPR process is moving faster than the s-CPR process. This can be explained by the fact that the s-CPR process has more inertia than the a-CPR one. If the first player plays C, then for both processes, the second player plays S or C according to her index; hence the indices associated to the s-CPR rule and to the a-CPR rule are increased by the same amount. If the first player plays S, the s-CPR and a-CPR processes lead to different revisions. For the a-CPR process, the second player does not act and the indices of his strategies remain unchanged. For the s-CPR process, the second player plays S with a probability proportional to her index, but since each strategy gets the same result, their indices grow on

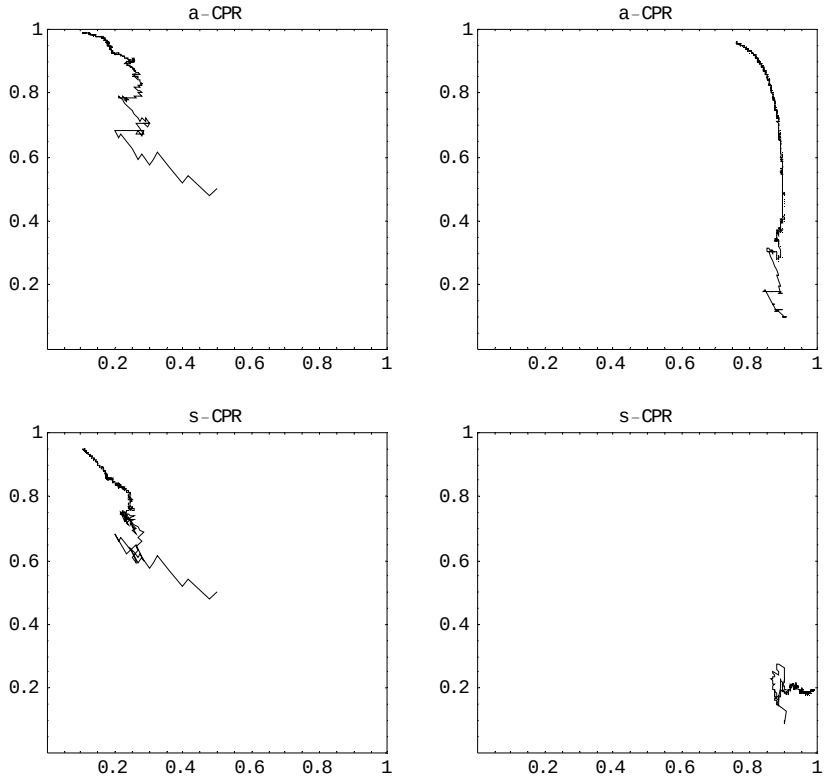

Figure 2: Simulations of a-CPR and s-CPR processes

average proportionally to their initial values.

Looking at the right side diagrams, one can see that the two processes behave differently. The upper right diagram shows the typical behavior of the a-CPR process. Even starting from a point where he seldom uses action S, the first player uses it more and more (path going upward), which, progressively, causes the second player to reinforce her own action S (path turning left). According to the theorem proved in this paper, this path will finally converge to SS. The lower right diagram is much less clear, but notice that the path is typically going toward the right. The weakly dominated strategy S for the second player is played more and more often, but convergence is hypothetical. No theoretical result is available in this case.

5 References

Beggs, A.W. (2002) : On the convergence of reinforcement learning, *mimeo,* University of Oxford.

Camerer, C. (2003) *Behavioral Game Theory Princeton* University Press.

Fudenberg, D., Levine, D.(1998) : *The Theory of Learning in Games*, MIT Press

Hopkins, E. (2002) : Two competing models of how people learn in games, *Econometrica,* 70: 2141-2166.

Ianni, A. (2002) : Reinforcement learning and the power law of practice: some analytical results, *mimeo*, University of Southampton.

Jehiel, P. and Samet, D. (2000) : Learning to play games in extensive form by valuation, *mimeo*, Ecole Nationale des Ponts et Chaussées.

Laslier, J.F., Topol, R. and Walliser, B. (2001) : A behavioral learning process in games, *Games and Economic Behavior*, 37: 340-366.

Pak, M. (2001): Reinforcement learning in perfect-information games, *mimeo*, University of California at Berkeley.

Roth, A. and Erev, I. (1995) : Learning in extensive-form games : experimental data and simple dynamic models in the intermediate term, *Games and Economic Behavior*, 29: 244-73.

Sarin, R. and Vahid, F. (1999): Payoff assessments without probabilities: a simple dynamic model of choice, *Games and Economic Behavior*, 28: 294-309.

Sarin, R. and Vahid, F. (2001) : Predicting how people play games: a simple dynamic model of choice, *Games and Economic Behavior*, 34: 104-122.

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