

A reinforcement learning process in extensive form games

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

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

June, 27, 2002

Abstract

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

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

1 Introduction

In a preceding paper (Laslier, Topol and Walliser, 2001, henceforth LTW), we studied the convergence properties, in a repeated finite two-player normal form game, of some learning process where each player uses the CPR (cumulative proportional reinforcement) rule. Like any reinforcement learning rule, the CPR rule associates, at each period, a 'valuation rule' and a 'decision rule'. For the CPR rule, the first states that the player computes for each action an index equal to its past cumulative utility and the second states that the player plays an action with a probability proportional to the preceding index. It is shown that the process converges with positive probability toward any strict pure Nash equilibrium and with zero probability toward any non Nash state as well as toward some mixed Nash equilibria (which are characterized). By the way, for a

*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

single decision-maker under risk, it is shown that the process converges toward the expected utility maximizing action(s).

The present paper considers a repeated finite extensive form game with perfect information, moreover assumed to have generic payoffs (no ties for any player). The CPR rule is now applied by each player, not to her “strategies” (in the usual sens of set of intended conditional actions), but simply to each action at a each node in the game tree when reached. It is shown that the CPR process converges with probability 1 toward the unique subgame perfect equilibrium path (obtained by backward induction).

For any learning process, it is convenient to distinguish between ‘convergence in actions’ of the moves which are selected and ‘convergence in values’ of the indices which are computed. The CPR process converges in actions, even if it does not converge in values. However, the perfect equilibrium values (i.e. the payoffs that the players reach at each node 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 and leads to a similar result, but with more complex learning rules. Jehiel and Samet (2000) consider a valuation rule where the player computes for each action an index equal to its past average utility, and a decision rule where 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 the end of the process, it converges in values toward the subgame perfect equilibrium values, but the actions only approach the subgame perfect equilibrium actions (they reach the equilibrium actions for their maximizing part). Pak (2001) considers a valuation rule where each action has a stochastic index equal either to its past utilities (with a probability proportional to their frequency) or to some random values (with a probability decreasing with the number of occurrences of that action), and a decision rule where he 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 the last 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 in both cases by the decision rule which is a maximizing one and by the valuation rule which is an averaging one. Exploration is expressed by a random perturbation either on the decision rule (first case) or on the valuation rule (second case); moreover, such a perturbation is constant (first case) or decreasing (second case). Conversely, as concerns the CPR rule, the exploration component is directly integrated in a non maximizing decision rule (allowing for mutations), while the exploitation component is associated with a cumulative valuation rule (since it creates a feedback effect on best actions). Hence, the trade-off is endogenous, leading to much exploration at the beginning of the process (since the initial indices are uniform) and much exploitation at the end if convergence occurs (exploration keeping however

active till the end).

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, $A(n)$ the set of actions at his disposal, $G(n)$ 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 utility vector for the players, assumed to be strictly positive. The game is assumed to be “generic” in the following sense: for any player, the utility obtained at different terminal nodes differs: if $m \neq m' \in M$, $u_i(m) \neq u_i(m')$ ¹. Call $\underline{u}_i = \min\{u_i(m) : m \in M\}$ the smallest utility player i can get from any terminal node (in fact, one can take $\underline{u}_i = 0$) and $\bar{u}_i = \max\{u_i(m) : m \in M\}$ the largest one. A strategy s specifies an action played at each node; a mixed strategy specifies a probability distribution on strategies; a behavioral strategy specifies a probability distribution on the actions available at each node. The game has a unique subgame perfect equilibrium (SPE); it is obtained by a backward induction procedure selecting maximizing action $a^*(n)$ for player $I(n)$ at each node n and providing utility $u_i^*(n)$ to any player i .

The stage game is now played an infinite number of times, labelled by t . At each period, a path h_t is followed; each player i knows which nodes she successively reached and observes the utility $u_t(i)$ she gets at its end. After t periods, call $N_t(a)$ the number of times that action a was used. The a-CPR (“action cumulative proportional reinforcement”) rule is not defined 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 utility 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); 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 (such as $a \in A(n)$) with a probability $p_t(a)$ proportional to $v_t(a)$.

Of course, the extensive form stage game can be transformed into a normal form one by introducing the notion of strategy. It must be noticed that a generic extensive form game does not generally lead to a generic normal form game. Using the CPR rule on that normal form defines the s-CPR (“strategy 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 utility obtained by that

¹A weaker assumption states that if the payoffs are similar for one player, they are similar for the others; this assumption would lead to the same results, at the cost of lengthier exposition.

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 Strategy-based vs action-based learning

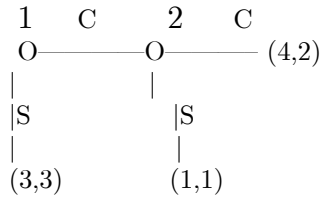
Even if the paper focuses on action-based learning, this section is devoted to the comparison with strategy-based learning. The convergence results obtained by LTW for the s-CPR process of players acting on a generic normal form stage game can be applied to a normal form game obtained from an extensive-form one. However, the relevant results only concern convergence toward a strict pure-strategy Nash equilibrium (i.e. each player's equilibrium strategy is a strict best response to the others' ones). A strict Nash equilibrium is obtained in a reduced extensive form game only under drastic conditions. These conditions can only be met in a restricted class of games (containing the centipede game) where, at each node, the player can only continue or stop (and stop the game).

Claim 1 *For a generic extensive-form game, a Nash equilibrium is strict iff it reaches all non-terminal nodes.*

Proof. A strict Nash equilibrium reaches all nodes. If some node were not reached, by modifying the action of the player playing at that node, the equilibrium would be kept, hence this player would obtain the same (equilibrium) utility with a different strategy. Conversely, if a Nash equilibrium reaches all nodes, it is strict. If a Nash equilibrium reaches all nodes, it must be the unique subgame perfect equilibrium since the perfect equilibrium is obtained by a backward induction procedure; moreover, a subgame perfect equilibrium is strict due to genericity of the extensive-form game. **QED**

For an extensive form game, one may now compare the respective effect of the s-CPR process and the a-CPR process. For the specific games where the subgame perfect equilibrium reaches all nodes, the equilibrium is obtained with positive probability with the s-CPR learning process (as was shown) and with probability 1 with the a-CPR process (as will be shown). But, the game may have other non strict Nash equilibria, which are obtained with probability 0 with the a-CPR process and with an unknown probability with the s-CPR process (since no result is available). For all other games, the subgame perfect equilibrium is a non strict Nash equilibrium, hence is obtained with probability 1 with the a-CPR process and with an unknown probability with the s-CPR process. Some Nash equilibrium strategies may well be weakly dominated, but it cannot be proved that they are eliminated with the s-CPR process. The only general result is that strongly dominated strategies are eliminated with the s-CPR process (since this last statement has not been formally proved in previous papers, a proof is provided in the appendix).

Consider for instance the following game (similar to the chain-store paradox) in extensive and normal form:



	<i>S</i>	<i>C</i>
<i>S</i>	$(3,3)^N$	$(3,3)$
<i>C</i>	$(1,1)$	$(4,2)^N$

In this game, even if actions and strategies structurally coincide, the subgame perfect equilibrium CC and the other Nash equilibrium SS (where S is weakly dominated for player 2) have different convergence properties for the two learning processes :

- CC is strict, hence it is obtained with positive probability by the s-CPR process and with probability 1 by the a-CPR process
- SS is not strict, hence there is no convergence result available by the s-CPR process and it is obtained with probability 0 by the a-CPR process.

If the first player continues, for both processes, the second player chooses to stop or to continue according to its index and the indices associated to the s-CPR rule and the a-CPR rule are likewise increased. If the first player stops, for the s-CPR process, the second player chooses to stop or to continue with a probability proportional to its index and, since each strategy gets the same result, their indices grow on average proportionally to their initial value; but for the a-CPR process, the second player has not to act and the indices of his strategies are unchanged. In other words, the process has more inertia in the first than in the second case since differential utilities have less impact on the indices.

>From that example, it should come as no surprise that, for extensive form games, the a-CPR process converges more easily than the s-CPR process. In (intrinsically) normal form games, the s-CPR rule is the natural expression of the reinforcement behavior. Dealing with extensive form games, the a-CPR rule appears not only as the most natural expression of that behavior but also as the easiest to implement for the players (since it does not require to consider strategies).

4 Convergence results

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

Lemma 1 *With the a-CPR rule, 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 utility that player $I(n)$ obtains after choosing a is in some positive interval $[u_{\min}(a), u_{\max}(a)]$. The cumulative utility 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. The lemma follows. **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. Let (Ω, π) be a probability space on which the repeated play of the game following the a-CPR rule is realized. A draw $\omega \in \Omega$, defines the path $h(t, \omega)$ at date t and the history $H(t, \omega) = (h(\tau, \omega))_{1 \leq \tau \leq t}$ 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 perfect equilibrium path at date t from a non terminal node n is denoted by $q_t(n, \omega)$. We prove that, ω -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 particular, for a non-terminal node n , $q_t(n, \omega) = p_t(a^*(n), \omega) q_t(n', \omega)$, where n' is the node resulting from $a^*(n)$. Hence, the proof goes by backward induction on subgames.

For the initial induction step, consider any node $n = B(m)$ preceding a terminal node m . Here $q_t(n, \omega) = p_t(a^*(n), \omega)$. The player $I(n)$ faces an individual choice between actions in $A(n)$ of which $a^*(n)$ is the maximizing one. According to the lemma, ω -almost surely, the process reaches node n an infinite number of times; one may label these dates by a new index θ . Slightly abusing notation, the probability of playing $a^*(n)$ at date θ writes $p_\theta(a^*(n))$. Consider now the event:

$$F(n) = \{\omega \in \Omega / \lim_{\theta \rightarrow \infty} p_\theta(a^*(n), \omega) = 1\}$$

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

$$\pi(F(n)) = 1$$

Especially, for almost all ω , for all $\varepsilon > 0$, there exists Θ such that, if $\theta \geq \Theta$, then $p_\theta(a^*(n), \omega) \geq 1 - \varepsilon$. By definition of the a-CPR rule, $p_t(a^*(n), \omega)$ is only modified at dates θ (when node n is reached); hence there exists T such that, if $t \geq T$, then $p_t(a^*(n), \omega) \geq 1 - \varepsilon$. This proves that, ω -almost surely:

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

For the general induction step, consider any non-terminal node \tilde{n} . The player $i = I(\tilde{n})$ faces now an individual choice between lotteries in $A(\tilde{n})$. Label $a_0, a_1, \dots, a_k, \dots$ the actions in $A(\tilde{n})$, any action a_k leading to node n_k , with $a_0 = a^*(\tilde{n})$ 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 . 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, depending 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 utility $u_i(n_0) = u_i^*(\tilde{n})$ with probability $1 - \varepsilon_0$ and utility \underline{u}_i with probability ε_0 ;
- if $k \neq 0$, the lottery $\mathbf{L}(n_k)$ gives utility $u_i(n_k) = u_i^*(n_k)$ with probability $1 - \varepsilon_k$ and utility \bar{u}_i 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 utility 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 , that we shall denote by $\mathbf{p}_t(a_0)$ tends almost surely to 1.

Now compare, starting at \tilde{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 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, ω -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 \bar{u}_i) is larger than the payoff in the true one.

It follows that, ω -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 \tilde{n} . It states that the probability of choosing action a_k is: $\frac{v_t(a_k)}{\sum_{a_k \in A(\tilde{n})} v_t(a_k)}$. 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 \tilde{n} is $q_t(\tilde{n}) = p_t(a_0)q_t(n_0)$ and

$q_t(n_0)$ tends to 1 by the induction hypothesis. It follows that, almost surely, $q_t(n)$ tends to 1 when t tends to infinity. **QED**

A Appendix

Theorem 2 *For a normal-form game, the s-CPR process eliminates strongly dominated strategies*

Proof. The utility for the first player of the combination of a strategy s_i of the first player and of a strategy s_h of the second player is denoted u_{ih} . Call, at each period t , x_{ih} the frequency of playing simultaneously s_i and s_h in the past. In continuous time, the evolution of the deterministic associated process is given by (equation 8 in LTW):

$$\dot{x}_{ih} = -x_{ih} + p_i q_h$$

The probability of playing strategy s_i is given by:

$$p_i = \sum_h x_{ih} u_{ih} / \sum_{jh} x_{jh} u_{jh}$$

By differentiating the second equation and replacing along the first, one gets the differential evolution of two strategies s_i and s_j of the first player :

$$\dot{p}_i / p_i - \dot{p}_j / p_j = \sum_h q_h (u_{ih} - u_{jh}) / \sum_{lh} x_{lh} u_{lh}$$

If strategy s_i is strictly dominated by strategy s_j , the numerator is greater than some positive lower bound (and the denominator is strictly positive). The differential inequation $\dot{p}_i / p_i - \dot{p}_j / p_j > \theta > 0$ implies $p_i/p_j > e^{\theta(t-t_0)}$, hence (since p_i is upper bounded), p_j goes to 0. By a usual proof, the stochastic process converges too to the elimination of s_j in continuous time, hence in discrete time.

QED

References

- [1] Jehiel, P, and Samet, D. (2000). "Learning to play games in extensive form by valuation," mimeo.
- [2] Laslier, J.-F., Topol, R., and Walliser, B. (2001). "A behavioral learning process in games," *Games and Economic Behavior*, **37**, 340-366.
- [3] Pak, M. (2001). "Reinforcement learning in perfect-information games," mimeo, University of California at Berkeley.