A REINFORCEMENT LEARNING PROCESS IN EXTENSIVE FORM GAMES

Jean-François LASLIER (Laboratoire d'Econométrie de l'Ecole Polytechnique,Paris)
Bernard WALLISER (CERAS, Ecole Nationale des Ponts et Chaussées, Paris)

Preliminary version. May 2002

In a preceding paper (Laslier-Topol-Walliser, 2001), we studied the convergence properties, in a repeated finite two-player normal form game, of some learning process where each player uses a CPR (cumulative proportional reinforcement) rule. The CPR rule associates, at each period, a 'valuation rule' stating that the player computes for each action an index equal to its past cumulative utility and a 'decision rule' which 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 some mixed Nash equilibria (which are characterized). By the way, for a 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 two-player extensive form game with perfect information and with generic payoffs (no ties for one player). The CPR rule is now applied by each player, no more to the complete strategy of the player, but to each choice at a given node, along the followed path of the game tree . It is shown that the process converges with probability 1 toward the (unique) subgame perfect equilibrium path, obtained by a backward induction procedure on actions and associated values.

A similar problem was already studied in the literature and leads to a similar result, but with different and less natural learning rules. Jehiel-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 he 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 till the end of the process, the values converge toward the values corresponding to the subgame perfect equilibrium (i.e. the payoffs that the players can reach at each node), 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 occurences 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 both cases, the learning rule reflects a trade-off between an exploration and an exploitation component, which happens in a non stationary context. Exploitation is expressed by the decision rule which is a maximizing one and 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). In the CPR rule, the exploration component is directly integrated in a non maximizing

decision rule, associated with a cumulative valuation rule which favours the exploitation component. Hence, the trade-off is endogenous, leading to much exploration at the beginning of the process and much exploitation at the end (exploration keeping however always active). As shown, the process converges toward the subgame perfect equilibrium actions, but not toward the perfect equilibrium values (even if the last may be recovered by divising the cumulative index by the number of trials of an action).

1. Game and learning assumptions

Consider a finite generic game tree defined 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). Call d the depth of the tree, i.e. the length of the greatest path in the tree. 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 or m (except for r), call $B(n)$ the (unique) action leading to it . For each node m, call $u(m)$ the utility vector for the players, assumed to be positive. For any player, the utility obtained at different terminal nodes differs : if $m \neq m' \in M$, $u_i(m) \neq u_i(m')$ (sometimes, it is assumed, more generally, that if the payoffs are similar for one player, they are similar for the other player). 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 $a^D(n)$ at each node and attributing value $v^D(n)$ at each node.

The stage game is now played an infinite number of times labelled by time t. At each period, a path h_t is described; each player i knows which nodes are successively reached and observes the utility $u_t(i)$ he 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 of each player i 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 obtainined at the end of a trajectory is allocated to all actions in the trajectory); the initial valuation is v_0.

-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 tranformed into a normal-form one by introducing the notion of strategy. 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 strategy in the past;

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

It must be noticed that a generic extensive- form game does not generally lead to a generic normal-form game.

2. Strategy-based vs action-based learning

In this section, only two-player games will be considered. The utility for the first player of the combination of a strategy s_i of the first player and of a strategy s_j of the second player is denoted u_{ij}. The convergence results obtained by LTW for the s-CPR process of players acting on a generic normal-form game can nevertheless 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 other's one). A strict equilibrium is obtained in a reduced extensive-form game only for a very restrictive class of games. These games, like the centipede game, are such that any deviation from the longest path immediately leads to a terminal node.

Lemma 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 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**

Hence, for the specific games where the subgame perfect equilibrium reaches all nodes, the equilibrium is obtained with probability 1 by a s-CPR learning process.

The problem is to know wether the s-CPR process may converge toward a non strict pure strategy Nash equilibrium. Since such equilibria may be weakly dominated, a partial answer would be given if, in a s-CPR process, weakly dominated strategies were eliminated. In fact, it is only possible to prove that strongly dominated srategies are eliminated :

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

Proof : Call, at each period t, x_{ih} the frequency of playing simultaneously s_i (by player 1) and s_h (by player 2) 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:

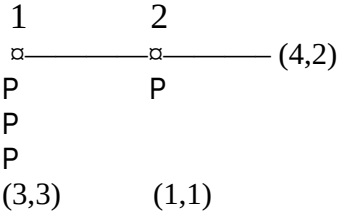
$p_i = \sum_h x_{ih} u_{ih} / \sum_{jh} x_{jh} u_{jh}$

By differenciating 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 $\dot{p}_i / p_i - \dot{p}_j / p_j > S > 0$ implies $p_i/p_j > e^{S(t-t_o)}$, 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**

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



	S	*C*
S	(3,3)	(3,3)
C	(1,1)	(4,2)

In this game, even if actions and strategies structurally coincide, the two Nash equilibria have

different convergence properties for the two learning processes (by anticipating on the further convergence result on a SPE for the a-CPR rule):

-the subgame perfect equilibrium CC is strict, hence it is obtained with positive probability by the s-CPR process and with probability 1 by the a-CPR process

-the equilibrium SS is not strict, hence there is no convergence result available by the s-CPR process and this equilibrium is obtained with probability 0 by the a-CPR process

If the first player continues, for the s-CPR process as well as for the a-CPR process, the second player chooses to stop or to continue according to its index and their indices 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. Hence, the process has more inertia in the first than in the second case since differential utilities have less impact on the indices. Notice that, as concerns convergence of the s-CPR process, applied to extensive-form games, toward a non SPE Nash equilibrium, no result is available, even in specific cases. Conversely, it should come as no surprise that the a-CPR process applied to extensive-form games, which is in fact the analog of the s-CPR process applied to (intrisically) normal- form games, converges to the SPE; this result will be shown now.

3. 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:

Proposition 1: *With the a-CPR rule applied to an extensive-form game, each node is attained an infinite number of times with probability* 1

Proof : Consider any node $n \in N$ where player $I(n)$ has $A(n)$ as her set of available actions. The following statement is first proven: if n is reached an infinite number of time, then each action $a \in A(n)$ is choosen an infinite number of times. For each $a \in A(n)$, the utility that player $I(n)$ obtains after choosing a is in some 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. Therefore, the argument on the proof of proposition 1 in LTW applies. Proposition 1 follows. Since the initial node is obviously reached an infinite number of time, by succesive steps in the tree, such is the case for all nodes. **QED**.

Proposition 1 ensures that each path (including the SPE path) is attained 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 :

Proposition 2: *With the a-CPR rule, the probability of playing the SPE path converges to* 1

Proof : It goes by backward induction on subgames

Let $(\Omega, \mathcal{A}, P)$ be a probability space on which a repeated play of the game is realized; $h(t)$, g is the CPR history at date t and $(h(t), g)_{t \geq 1}$ is the CPR trajectory for draw $g \in \Omega$.

Consider any node n preceding a terminal node (and called a penultimate node). The player $I(n)$ faces an individual choice between actions in $A(n)$ of which $a^D(n)$ is the maximizing one. For any g, the process reaches node n an infinite number of times; one may label these dates by a new index b. Slightly abusing notation, the probability of playing $a^D(n)$ at date b writes $p_b(a^D(n))$. Consider now the event:

$F_n = \{g \in \Omega / \lim_{b \to \infty} p_b(a^D(n)) = 1\}$

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

$P(F_n) = 1$

Especially, for almost all g, there exists T such that, if $t \geq T$, then $p_t(a^D(n)) \geq 1 - \varepsilon$

Consider now any node n' which precedes only penultimate or terminal nodes (and called an antepenultimate node). The player $I(n')$ faces now an individual choice between lotteries in

$A(n')$. Any action $a(n')$ leads to some node n (such that $n' = B(n)$) where another player $I(n)$ plays the actions in $A(n)$ with some probabilities, hence a lottery $L(n)$. However, the probabilities involved in the lotteries vary from one period to the other and proposition 4 in LTW is no more directly applicable. It is necessary to introduce intermediate lotteries with fixed probabilities, conditional to the fact that action $a(n')$ is the SPE action or not

-if $a(n') = a^{D}(n')$ leading to node n, the lottery $\mathbf{L}(n)$ gives utility $u(a^{D}(n))$ with probability $1 - \varepsilon$ and utility 0 with probability ε

-if $a(n') \neq a^{D}(n')$ leading to node n', the lottery $\mathbf{L}(n')$ gives utility $u(a^{D}(n'))$ with probability 1

According to the result for a penultimate node, the lottery $\mathbf{L}(n)$ dominates lottery $L(n)$ as much as $t \geq T$ and lottery $\mathbf{L}(n')$ is always dominated by $L(n)$. According to proposition 4 in LTW, since the game is generic, if e is small enough, the player chooses asymptotically lottery $\mathbf{L}(n)$ over any $\mathbf{L}(n')$. Hence, by the same argument than in the proof of proposition 4, the player chooses asymptotically $L(n)$ over $L(n')$.

The same reasoning can be continued for nodes before antepenultimate ones. **QED**

References

Jehiel, P.- Samet, D. (2000): Learning to play games in extensive form by valuation, mimeo.

Laslier, J.F.- Topol, R.- 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