Bush–Mosteller learning for a zero-sum repeated game with random pay-offs

ALEXANDER S. POZNYAK† and KADDOUR NAJIM‡

This paper deals with the design and analysis of a modified version of the Bush–Mosteller reinforcement scheme applied by partners in a zero-sum repeated game with random pay-offs. The suggested study is based on the learning automata paradigm and a limiting average reward criterion is tackled to analyse the arising Nash equilibrium. No information concerning the distribution of the pay-off is a priori available. The novelty of the suggested adaptive strategy is related to the incorporation of a 'normalization procedure' into the standard Bush–Mosteller scheme to provide a possibility to operate not only with binary but also with any bounded rewards of a stochastic nature. The analysis of the convergence (adaptation) as well as the convergence rate (rate of adaptation) are presented and the optimal design parameters of this adaptive procedure are derived. The obtained adaptation rate turns out to be of $o(n^{-1/3})$.

1. Introduction

Several studies have been dedicated to repeated games and learning in games (Brown 1951, Robinson 1951, Aumann and Maschler 1967, Kohlberg 1975, Aumann 1985, Fudenberg et al. 1990, Zamir 1992, Mertens et al. 1994, Gardner 1995, Friedman 1998). Repeated games have the effects of an enforcement mechanism and of a signalling mechanism at the same time (Forges 1992). In parallel, many theoretical as well as practical results related to the Learning automata paradigm (Lakshmivarahan and Narendra 1982, Najim and Poznyak 1994, Sastry et al. 1994, Poznyak and Najim 1997, Poznyak et al. 2000) have been carried out in the last two decades. There 'learning process' are treated as the ability of systems to improve their response (performance in some criterion sense) based on past experience (Cross 1973, 1983).

In general, learning systems are information processing systems whose architecture and behaviour are inspired by the structure of biological systems (the organism is born with relatively little initial knowledge and learns actions that are appropriate through trial and error). The vocabulary and the concepts associated with learning automata concept are borrowed from biology and psychology. Learning models stem from diverse approaches, frequently grounded on heuristic intuitions and experiments. As for learning automata, they have been used to solve engineering problems which are characterized by nonlinearity and a high level of uncertainty (note that many economic decision problems may be reduced to the control of stochastic process). Frequently, the uncertainties are of a higher order, and even the probabilistic characteristics such as the conditional distribution functions of a random reward may not be completely known. It is then necessary to learn (acquire) additional information. It has been shown that the concept of a mixed strategy behaviour rule developed by Borgers et al. (1998) is similar to the concept of a stochastic automaton of variable structure. Erev and Roth (1998) considered learning models based on reward–inaction learning automata to model the interaction between players. They show that these learning models can be used to predict as well as to explain observed behaviour of several games (see also Roth and Erev (1995)). Nazin and Poznyak (1977) presented an algorithm for a two-person zero-sum repeated
game with incomplete information. This algorithm uses a projectional reinforcement scheme. The projection operator guarantees that the probability measure that remains satisfied at each time is time consuming. Unfortunately, there does not exist any biological system with such behaviour justifying this numerical procedure. Sastry et al. (1994) used a linear reward-inaction algorithm with binary input to solve the problem related to decentralized learning Nash equilibria in multiplayer stochastic games with incomplete information. Using the ordinary differential equation technique, they have only stated the convergence (but not the rate) of this algorithm. Borgers and Sarin (1997) considered several agents playing in discrete time repeatedly the same normal-form game. They also used a learning model based on the Bush–Mosteller reinforcement scheme. They assumed that the pay-offs are normalized between zero and one and showed that in the continuous time limit the learning model converges to the replicator dynamics of evolutionary game theory. To state this result, the authors used Norman’s theorem which is concerned with the continuous time limit of discrete-time Markov processes with infinite state spaces.

This paper deals with the development and analysis of a new adaptive algorithm for a two-person zero-sum repeated game with a random pay-off. Each player is modelled with a stochastic learning automaton. The profit of one player equals the loss of the other player. Each player has finitely many actions one of which is played at each stage (time). After each stage, the pay-off (loss) to each player is a random variable. No information concerning the distribution of the pay-off is a priori available. In other words, the players have no complete information on the parameters of the game that is being played. In particular, a player does not know the pay-off as a function of its actions and does not know the actions of his or her opponent. Each player obtains, however, signals called environment response (the necessary information is obtained during the course of the game). The development of the suggested 'adaptive algorithm' is based on the use of 'learning stochastic automata with variable structure' (Nazin and Poznyak 1977, Najim and Poznyak 1994, Sastry et al. 1994). A modified version of the Bush–Mosteller reinforcement (updating) scheme complemented by a normalization procedure, designed by Poznyak and Najim (1977), has been used for this purpose. The reinforcement scheme operates with a continuous input (not obligatory binary) and uses a time-varying correction (updating) factor. Based on Lyapunov-like approach and martingale theory, the convergence with probability one as well as in the mean square sense are stated. The convergence rate of this adaptive procedure within a wide class of parameters (Poznyak and Najim 1977) is estimated too. The optimal values of the design parameters are also derived. They provide an 'adaptation rate' equal to $o(n^{1/3})$.

Since any theoretical result, based on some set of a priori accepted assumptions, has the restricted area of its application, here we present the basic features of the approach suggested in this paper and briefly discuss the arising restrictions to its applicability. The most important assumptions (and, as the result, restrictions) concerning the considered game are as follows.

**Assumption 1:** This is the two-person repeated game where the behaviour of each player is the static model given by a finite-state automaton with 'time-variable structure' (at each stage of the game the distribution of the selected control actions can vary according to a reinforcement procedure).

**Assumption 2:** If some controls are realized by players at a current stage, then each participant immediately obtains the information on the realized values of his or her pay-off and constrained junctions which are individual for each player. The necessary information (realizations of both pay-offs and constraints) is obtained during the course of the game. These realized pay-offs (called 'realizations' below) are random variables uniformly bounded (non obligatory binary) with probability one and having constant (conditional) moments which are assumed to be a priori unknown. In other words, the game is one of incomplete information.

**Assumption 3:** The participants accept some sort of agreement to use the same reinforcement learning procedure (in this case the Bush–Mosteller scheme) to maximize their average pay-off and not to change it to another one during all long-range time of the game. Any other 'co-operations' during the game are prohibited. Only the parameters of this procedure can be modified by player during the game, but not the fixed structure of the reinforcement procedure.

All these assumptions have a simple physical sense and may serve for wide enough class of multiparticle constrained repeated games with a finite set of player actions. So, the results of the study of such game models can be expected to make them of high interest and importance.

Briefly, the main contributions of this paper can be summarized as follows.

(i) A new affine transformation ('normalization procedure') of current information (pay-offs realizations) is suggested, leading to the formation of a scalar $[0, 1]$ random variable, which is shown to be sufficient for use in the selected learning reinforcement scheme; this transformation preserves the probability measure for the corresponding mixed strategies over the set of game actions.
(ii) For each player the considered reinforcement procedure, based on Lagrange multipliers and an appropriate regularization, is shown to be oriented to obtain the optimal response (in an average sense).

(iii) This learning tactic is proven to lead to the unique Nash equilibrium point.

(iv) Based on the stochastic approximation technique, the convergence (with probability one and in mean square sense) of the considered learning procedure to the Nash equilibrium is stated and the rate of learning is also estimated.

The remainder of this paper is organized as follows. The problem related to the repeated game, involving two learning stochastic automata models as players, is formulated in §2. In §3 we show that the Nash equilibrium is obtained at a 'saddle point' of a corresponding bilinear functions given for simplex sets of mixed strategies. The reinforcement learning procedure is presented in §4. In §5 the asymptotic properties (convergence and rate of adaptation) of the algorithm are dealt with. Some conclusions end this paper.

2. Problem setting

All the random variables considered in this paper are assumed to be defined on the probability space \((\Omega, \mathcal{F}, P)\). Before we formulate the problem, let us introduce some definitions extensively used throughout this paper.

**Definition 1:** A two-person zero-sum stochastic game is a quadruple

\[ \Gamma := (U^1, U^2, \{\xi_n\}, P), \quad (1) \]

where \(U^1 = (u^1(1), u^1(2), \ldots, u^1(N_1))\) is a finite action set of the first player, \(U^2 = (u^2(1), u^2(2), \ldots, u^2(N_2))\) is a finite action set of the second player, \(\{\xi_n\}\) is a sequence of absolutely integrable Borel functions (random variables or, below, pay-offs) \(\xi_n = \xi_n(\omega)\) defined at each time \(n = 1, 2, \ldots\) by the random event \(w \in \Omega\) and whose distribution function is dependent on the control actions \(u^1_n\) and \(u^2_n\) chosen by the players. \(P\) is a probability measure given on the sigma algebra \(\mathcal{F}\).

Below we give a detailed description of the pay-offs.

**Definition 2:** A \(N_1 \times N_2\) matrix \(V = [v_{ij}]\) is said to be the conditionally expected pay-off matrix of the stationary game \(\Gamma\), if each element \(v_{ij}\) corresponds to the expected pay-off induced at any stage of the play by the actions \(u^1(i)\) and \(u^2(j)\) chosen by the players, that is

\[ v_{ij} = \int_{\omega \in \Omega} \xi_n(\omega) \, dP(\omega | u^1_n = u^1(i), u^2_n = u^2(j)) \]

\[ = E \{ \xi_n | u^1_n = u^1(i), u^2_n = u^2(j) \}. \quad (2) \]

The next definition concerns the randomized or mixed strategies of each player, and the expected pay-off.

**Definition 3:** The probabilistic vectors \(p^1_n\) and \(p^2_n\) defined as

\[ p^k_n := \left( p^k_n(1), p^k_n(2), \ldots, p^k_n(N_k) \right)^T, \quad k = 1, 2, \quad (3) \]

are called respectively randomized or mixed strategies of players 1 and 2 if

\[ p^k_n(i_k) := P \{ \omega : u^k_n = u^k(i_k) \wedge \mathcal{F}_{n-1} \} \quad (4) \]

The function

\[ V(p^1_n, p^2_n) := (p^1_n)^T V p^2_n \quad (5) \]

is said to be the expected pay-off (at time \(n\)) of the first player in the stationary game \(\Gamma\) with the given conditionally expected pay-off matrix \(V \in R^{N_1 \times N_2}\) when the current mixed strategy (3) is used.

In this paper we deal with the case of incomplete information, that is the exact values of the matrix \(V\) are assumed to be a priori unknown, and only the ‘realization’ of the applied actions \(u^1_n = u^1(i), u^2_n = u^2(i)\) as well as the corresponding (realized) pay-off \(\xi_n\) at the current stage of the play are available.

Define the averaged pay-off function \(\Phi_n\) by

\[ \Phi_n := \frac{1}{n} \sum_{i=1}^{n} \xi_i. \quad (6) \]

It depends on the realized actions \(u^1_n\) and \(u^2_n\) selected by both players up to time \(n\) (\(n\)th stage), this function corresponds to the gain of the first player (the loss of the second player), that is we deal with zero-sum repeated game.

The following assumptions will be in force throughout this paper.

**Assumption 4:** The conditionally expected payoff \(V\) is independent of the history of the game, that is

\[ v_{ij} = E \{ \xi_n | u^1_n = u^1(i), u^2_n = u^2(j) \wedge \mathcal{F}_{n-1} \} = \]

\[ = E \{ \xi_n | u^1_n = u^1(i), u^2_n = u^2(j) \}. \quad (7) \]

where \(\mathcal{F}_{n-1} = \sigma(\xi_i, u^1_i, u^2_i | i = 1, 2, \ldots, n-1)\) is the sigma algebra generated by the corresponding data.
Assumption 5: For any action sequences \( \{u^1_n\} \) and \( \{u^2_n\} \)
the absolute value of the pay-off \( \xi_n \) at the current stage \( n \)
of the play is uniformly bounded with probability one (on \( w \) and \( n \)), that is
\[
|\xi_n| \leq \sigma^+ \leq \infty.
\]

Now we are ready to formulate a two-person zero-sum matrix stationary game problem with incomplete information: to generate sequences (may be random) actions \( \{u^1_n\} \) and \( \{u^2_n\} \) in such a way to achieve maximal pay-off for the first player and minimal loss for the second one, that is
\[
\limsup_{n \to \infty} \Phi_n \overset{\text{as}}{\to} \inf \sup_{u^1_n \in u^1_n} \sup_{u^2_n \in u^2_n}
\]
\[
\text{(8)}
\]

The actions \( \{u^1_n\} \) and \( \{u^2_n\} \) of both players are randomly selected according to the random sequences \( \{P^1_n\} \) and \( \{P^2_n\} \) of the conditional distribution vectors defined by (4).

In the next section we shall show that, in the case of complete information on the matrix \( V \), this problem is equivalent to a minimax problem for bilinear functions given on the corresponding simplex sets.

3. Equivalent minimax

3.1. Stationary strategies as a subclass of minimax strategies

Consider now the subclass of stationary mixed strategies \( D = \{p^k, k = 1, N\} \).

Definition 4: The point \( (p^1, p^2) \) is said to be an equilibrium point of a zero-sum matrix game \( \Gamma \) within the class \( D \) of stationary strategies if
\[
V(p^1, p^2) = \max_{p'} |V(p^1, e^x)| = \min_{p'} [V(p^1, p^2)].
\]
\[
\text{(9)}
\]

At this point no player can improve his or her pay-off by a unilateral change in his or her strategy.

Remark: Evidently, the set of all equilibrium strategies contains the subset of stationary strategies \( \{p^k\} \). This follows from:

(i) the continuity property of the function \( V \) (5),

(ii) the compactness of the simplexes \( S_{N_1}^{N_2} \) where
\[
S_{N_1}^{N_2} := \left\{ \bar{p}^k : \bar{p}^k(i_k) \geq \varepsilon \geq 0, \sum_{i_k=1}^{N_1} \bar{p}^k(i_k) \right\}
\]
\[
\text{(10)}
\]

and

(iii) the Nash theorem (Aubin 1979, p. 268, theorem 1).

3.2. Saddle-point (Nash equilibrium) as a minimax solution

Consider the following minimax problem:
\[
V(p^1, p^2) := (p^1)^T V p^2
\]
\[
= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} v_{ij} p^1(i) p^2(j) \rightarrow \min \max_{p^1 \in S_{N_1}, p^2 \in S_{N_2}}
\]
\[
\text{(11)}
\]

According to the theorem of von Neumann and Morgenstern (1944) this problem has a solution (not obligatorily unique) \( ((p^1)^*, (p^2)^*) \) called a saddle point of the function \( V(p^1, p^2) \) which for any \( p^1 \in S_{N_1} \) and \( p^2 \in S_{N_2} \) verifies that
\[
V(p^1, (p^2)^*) \leq V((p^1)^*, (p^2)^*) \leq V((p^1)^*, p^2)
\]
\[
\text{(12)}
\]

As follows from (9), this saddle point is one of Nash equilibrium of the zero-sum game with matrix \( V \) (Filar and Vrieze 1979).

3.3. Asymptotically minimax strategies

We shall show that, if any mixed strategies \( \{p^1_n\} \) and \( \{p^2_n\} \) (3) converge as \( n^{-1} \) to a saddle point \( ((p^1)^*, (p^2)^*) \), in other words, it is asymptotically minimax strategy, then the corresponding averaged pay-off function \( \Phi_n \) (6) also converges (with the same rate) to the game value \( V^* = V((p^1)^*, (p^2)^*) \).

Theorem 1: If under assumptions (4) and (5) there exists a positive parameter \( \tau \) such that
\[
\limsup_{n \to \infty} \left\{ n^{-1} E[||p^1_n - (p^1)^*||^2 + ||p^2_n - (p^2)^*||^2] \right\} < \infty,
\]
then
\[
\limsup_{n \to \infty} \left\{ n^{-1} E[||\Phi_n - V^*||^2] \right\} < \infty.
\]
\[
\text{(13)}
\]

Proof: Evidently
\[
E[||\Phi_n - \hat{V}_n||^2] \overset{n \to \infty}{\to} 0(n^{-1})
\]
\[
\hat{V}_n := \frac{1}{n} \sum_{i=1}^{n} V(p^1_n, p^2_n).
\]

So, to prove (13) it is sufficient to state that
\[
\limsup_{n \to \infty} \left\{ n^{-1} E[||\hat{V}_n - V^*||^2] \right\} < \infty
\]
since, for \( \tau \leq 1 \)
\[
\limsup_{n \to \infty} \left\{ n^{-1} E[||\Phi_n - V^*||^2] \right\}
\]
\[
\leq 2 \limsup_{n \to \infty} \left\{ n^{-1} E[||\Phi_n - \hat{V}_n|| + ||\hat{V}_n - V^*||^2] \right\}
\]
\[
\leq \text{constant} + 2 \limsup_{n \to \infty} \left\{ n^{-1} E[||\hat{V}_n - V^*||^2] \right\}.\]
Consider the following function
\[ d_n := |\tilde{V}_n - v^*|. \]
Taking into account that the function \( V(p^1, p^2) \) is smooth enough and satisfies the Lipschitz condition, and using the inequality
\[ 2ab \leq a^2 + b^2, \]
we derive
\[
E\{d_n|\mathcal{F}_{n-1}\} \\
\leq \left(1 - \frac{1}{n}\right)d_{n-1} + \frac{C_1}{n^2} + \frac{C_1}{n^{1+\tau}},
\]
where
\[ \mathcal{F}_{n-1} := \sigma(p^1_k, \ldots, p^2_k; k = 1, 2). \]

In view of lemma A.3.1 (appendix A) of Poznyak and Najim (1997), we conclude that
\[ d_n \to 0. \]
Using lemma A.3.22 (appendix A) of Poznyak and Najim (1997) and considering the mathematical expectation of both sides of the previous inequality, we obtain the desired result (13).

So, if we construct a mixed strategies \( \{p^1_n, p^2_n\} \) which converges to the saddle point \((i(p^*_1, (p^*_2)^*) \) of the function \( V(p^1, p^2) \), and use it in a given stochastic game, then the Nash equilibrium will exist and the game will have a value.

### 3.4. Regularization of the game

As already mentioned above, the saddle point may not be unique and, to avoid this indeterminacy, we use the regularization approach presented by Nazin and Poznyak (1977) and Poznyak and Najim (1997). Consider the following regularized pay-off function:
\[
V^\varepsilon(p^1, p^2) := V(p^1, p^2) - \frac{\varepsilon}{2} \left(\|p^1\|^2 - \|p^2\|^2\right), \quad \varepsilon > 0,
\]
where the arguments belong to the \( \varepsilon \) simplexes (10).

This regularized function is strictly concave in \( p^1 \) and strictly convex in \( p^2 \). As a consequence it has a unique saddle point (Aubin 1979) which will be denoted by \(((p^1(\varepsilon, \delta))^*, (p^2(\varepsilon, \delta))^*)\).

**Lemma 1:** Let \( \{\varepsilon_n\} \) and \( \{\delta_n\} \) be the sequences verifying \( \varepsilon_n \in (0, \min(N_i^{-1}, N_j^{-1})) \), \( \delta_n > 0 \), \( n = 1, 2, \ldots \),
\[
\lim_{n \to \infty} \varepsilon_n = 0, \lim_{n \to \infty} \delta_n = 0, \quad n \geq 1, \quad n \geq 1
\]

Then all the possible saddle points \( V(p^1, p^2) \) can be parametrized as \(((p^1(\eta))^*, (p^2(\eta))^*)\) and the sequences \(((p^1(\varepsilon_n, \delta_n))^*, (p^2(\varepsilon_n, \delta_n))^*)\) converge respectively to the unique point \(((p^1(\eta))^*, (p^2(\eta))^*)\), that is
\[
\|(p^1(\varepsilon_n, \delta_n))^* - (p^1(\eta))^*\|
\]
and there exist constants \( C_i(i = 1, 2, 3) \) such that
\[
\|(p^1(\varepsilon_{n+1}, \delta_{n+1}))^* - (p^1(\varepsilon_i, \delta_i))^*\|
\]

Finally, we conclude that
\[
\varepsilon_n \to 0, \delta_n \to 0, \quad n \to \infty
\]

**Proof:** The proof is similar to the proof of theorem 2 of §4.3 of the book by Poznyak and Najim (1997) and can be found in the paper by Nazin and Poznyak (1977).

The next section deals with the learning (adaptive) algorithm which, being used by both players, provides asymptotically the Nash equilibrium for the given repeated game.

### 4. Adaptive Algorithm

Learning automata have been used to solve many engineering problems (Najim and Poznyak 1994, Poznyak and Najim 1997). The Bush–Mosteller reinforcement scheme and the normalization procedure designed by Poznyak and Najim (1977) will be used for the design of a new adaptive algorithm for the two-person zero-sum repeated game with incomplete information. The necessary information is obtained during the course of the game.

**For the first player,**
\[
p_{1}^{i}_{n+1} = p_{1}^{i}_{n} - \gamma_1 \left( e_{N_i}(u_{1}^{i}_{n}) - p_{1}^{i}_{n} + \frac{e_{N_i}[e_{N_i} - N_i e_{N_i}(u_{1}^{i}_{n})]}{N_i - 1} \right),
\]
if \( u_{1}^{i} = u_{1}^{j} \), then \( e_{N_i}(u_{1}^{i}) \)
\[
= \left( 0, \ldots, 0, 1, 0, \ldots, 0 \right)^{T} \in R^{N_i},
\]
\[
e_{N_i} = (1, \ldots, 1)^{T} \in R^{N_i}, \quad p_{1}^{i} = \frac{1}{N_i},
\]
A. S. Poznyak and K. Najim.

Theorem 2 — Convergence with probability one: Consider the learning reinforcement procedure under assumptions (4) and (5), and, in addition, assume the following.

(i) There exist three non-negative sequences \( \{\gamma_n\} \), \( \{\delta_n\} \) and \( \{\tau_n\} \) such that

\[
\gamma_n \in (0, \gamma^+), \quad \gamma_n \uparrow 0,
\]

\[
\delta_n \rightarrow 0, \quad \min\{N_1, N_2\} < \tau_n \downarrow 0.
\]

(ii) The correction factor \( \{\gamma_k\} \) is selected as

\[
\gamma_k = \frac{\gamma_n(N_k - 1)}{\alpha_n N_k}, \quad k = 1, 2,
\]

where \( \alpha_n \) is defined by (15) and \( \gamma_n \) satisfies

\[
\sum_{n=1}^{\infty} \gamma_n \delta_n = \infty.
\]

(iii)\[
\sum_{n=1}^{\infty} \left( \frac{\gamma_n}{\tau_n} \right)^2 < \infty,
\]

\[
\sum_{n=1}^{\infty} \left( \tau_n - \tau_{n-1} \right)^2 + \left| \delta_{n+1} - \delta_n \right|^2 + \left| \tau_n \delta_{n+1} - \tau_{n-1} \delta_n \right|^2
\]

\[
[1 + \gamma_n(\tau_n)^{-2}(\delta_n)^{-1}] < \infty.
\]

Then the mixed strategies of the players converges with probability 1 to one of the saddle points of the game with matrix \( V \), that is

\[
\|p_{n+1} - (p^1(\eta))\|_2^2 + \|p_{n+1} - (p^2(\eta))\|_2^2 \rightarrow 0.
\]

where

\[
\eta = \lim_{\tau_n \downarrow 0} \frac{\tau_n}{\delta_n}.
\]

Proof: Introduce the following Lyapunov function:

\[
W_{n+1}(p_{n+1}, p_{n+1}^2) = \|p_{n+1} - (p^1(\varepsilon_n, \delta_{n+1}))\|_2^2
\]

\[
+ \|p_{n+1}^2 - (p^2(\varepsilon_n, \delta_{n+1}))\|_2^2.
\]

Taking into account the algorithm and, adding and subtracting \( (p^1(\varepsilon_n, \delta_n))^* \) and \( (p^2(\varepsilon_n, \delta_n))^* \) respectively to the first term and to the second term of the right-hand side of the previous equality, we obtain

\[
W_{n+1}(p_{n+1}, p_{n+1}^2) = \|p_{n+1} - (p^1(\varepsilon_n, \delta_{n+1}))\|_2^2 - \|p_{n+1} - (p^1(\varepsilon_n, \delta_{n+1}))\|_2^2
\]

\[
- \|p_{n+1} - (p^1(\varepsilon_n, \delta_{n+1}))\|_2^2 - \|p_{n+1}^2 - (p^2(\varepsilon_n, \delta_{n+1}))\|_2^2
\]

\[
+ \|p_{n+1}^2 - (p^2(\varepsilon_n, \delta_{n+1}))\|_2^2.
\]

5. Convergence analysis

The objective of this section is to state the properties of the learning reinforcement procedure described in the previous section.

For the second player

\[
p_{n+1}^2 = p_{n+1}^2 + \gamma_n \left( e_{N_1}(u_n^2) - p_{n+1}^2 + \frac{\gamma_n^2 - N_2 e_{N_1}(u_n^2)}{N_2 - 1} \right),
\]

if \( u_n^2 = u^2(j) \), then \( e_{N_1}(u_n^2) = \begin{pmatrix} 0, \ldots, 0, 1, 0, \ldots, 0 \end{pmatrix}^T \in \mathbb{R}^{N_2} \).

\[
e_{N_2} = (1, \ldots, 1)^T \in \mathbb{R}^{N_2}, \quad p_{n+1}^2 = \frac{1}{N_2},
\]

\[
\varepsilon_n^2 := \frac{\alpha_n^2}{p_{n+1}^2(i)}, \quad y_n := \varepsilon_n + \delta_n p_{n+1}^2(i).
\]

The positive deterministic sequences \( \{\alpha_n^2\} \) and \( \{\beta_n^2\} \) are constructed as follows:

\[
\alpha_n^k = \frac{\alpha_n^k}{\delta_n + \sigma_n}, \quad \beta_n^k = \frac{\alpha_n^k + (N_k - 1)(\tau_n)^2}{1 + (N_k - 2)(\tau_n)}, \quad 0 < \tau_n \downarrow 0,
\]

\[
\beta_n^0 = \frac{(1 - \tau_n)\tau_n}{2(1 + (N_k - 2)(\tau_n)), \quad k = 1, 2}
\]

(15)

This adaptive algorithm is constructed using the Bush–Mosteller reinforcement scheme and a normalization procedure, that is

(i) the Bush–Mosteller construction provides at each step of the procedure the increase in the component \( p_{n+1}^k(i) \) with the value \( \gamma_n^k \left[ p_{n+1}^k(i) - 1 + \varepsilon_{n+1}^k \right] \) (more \( \varepsilon_{n+1}^k \) and more \( p_{n+1}^k(i) \)) and, analogously, for \( p_{n+1}^k(j) \) and

(ii) it is easy to verify (see lemma 2 of §4.4 of the paper by Poznyak and Najim (1997)) that, for any time \( n \), \( \varepsilon_n^k \) and \( \varepsilon_{n+1}^k \) are \( (0, 1) \) and

\[
p_{n+1}^k(i) \geq \tau_{n+1} := \varepsilon_n.
\]

that is

\[
p_{n+1}^1 \in S_{\tau_{n+1}}, \quad p_{n+1}^2 \in S_{\tau_{n+1}}^2.
\]

Now we shall show below that, under certain conditions, this adaptive algorithm has nice asymptotic properties.
where

$$A_n^k := e_N(u_n^k) - p_n^k + \frac{\xi_n^k[e_N - N_k e_N(u_n^k)]}{N_k - 1}.$$  

(16)

Observe that

$$\|A_n^k\| \leq C_A = \text{constant} < \infty.$$

From the last expression and in view of lemma 1, it follows that

$$W_{n+1}(p_{n+1}^1, p_{n+1}^2) \leq W_n(p_n^1, p_n^2) + 2\sum_{k=1}^2 \gamma_n^k[W_n(p_n^1, p_n^2)]^{1/2} \kappa_n$$

$$+ \kappa_n^2 - 2\sum_{k=1}^2 \gamma_n^k[p_n^k - (p^k(\varepsilon_n, \delta_n))]<A_n^k$$

$$+ C_A \sum_{k=1}^2 (\gamma_n^k)^2.$$  

(17)

Let us calculate the following conditional mathematical expectation:

$$E\{A_n^k | F_{n-1}\}$$

$$= \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} E\{A_n^k | u_n^i = u^i(j) \wedge F_{n-1}\} p_n^1(i)p_n^2(j)$$

$$= \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \left( e_N(u_n^i) - p_n^k + \frac{\xi_n^k[e_N - N_k e_N(u_n^i)]}{N_k - 1} \right).$$  

(18)

$$p_n^1(i)p_n^2(j)$$

$$= \frac{1}{N_k - 1} \sum_{u=1}^{N_k} \sum_{i=1}^{N_k} \tau_n^k[e_N - N_k e_N(u^k(i))] p_n^1(i)p_n^2(j)$$

where

$$i_k = \begin{cases} i & \text{if } k = 1, \\ j & \text{if } k = 2, \end{cases}$$

and

$$\tau_n^k = E\{\xi_n^k | u_n^i = u^k(i), u_n^j = u^k(j) \wedge F_{n-1}\}$$

$$= \frac{\alpha_n^k \tau_n^k + \beta_n^k}{p_n^k(j)}$$

$$= \frac{\alpha_n^k \tau_n^k + \beta_n^k}{p_n^k(j)}.$$
hence, in view of the identity
\[ [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top \sum_{k=1}^{N_k} \frac{\partial}{\partial p_k^k(i_k)} V_\delta(p_\alpha, p_\beta) = 0, \]
we obtain
\[ [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top E\{ A^k_n | \mathcal{F}_{n-1} \} \]
\[ = -C_n^k [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top \nabla_{p_k^k} V_\delta(p_\alpha, p_\beta), \]
\[ a_n^k \gamma_n = \gamma_n. \]
Substituting this relation into (19) leads to
\[ E\{ W_{n+1}(p_{n+1}^1, p_{n+1}^2) | \mathcal{F}_{n-1} \} \]
\[ \leq \frac{a_n}{2} W_n(p_\alpha, p_\beta) + \sum_{k=1}^{2} \gamma_n^k W_n(p_\alpha, p_\beta)^{1/2} \]
\[ - 2 \sum_{k=1}^{2} (-1)^k \frac{\gamma_n^k}{\gamma_n} [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top \nabla_{p_k^k} V_\delta(p_\alpha, p_\beta) \]
\[ + C A^2 \sum_{k=1}^{2} (\gamma_n^k)^2 + \kappa_n^2. \]
Now, let us recall that the regularized pay-off function is strictly concave with respect to \( p_1 \) and is strictly convex with respect to \( p_2 \); so, in view of the relation
\[ a_n^k \gamma_n = \gamma_n, \]
and lemma 1, §4.3, of the book by Poznyak and Najim (1997), it follows that
\[ \sum_{k=1}^{2} (-1)^k \frac{\gamma_n^k}{\gamma_n} [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top \nabla_{p_k^k} V_\delta(p_\alpha, p_\beta) \]
\[ = \gamma_n \sum_{k=1}^{2} (-1)^k [p_n^k - (p_k^k(x_n^k, b_n^k))^*]^\top \nabla_{p_k^k} V_\delta(p_\alpha, p_\beta) \]
\[ \leq \frac{C \gamma_n^k}{2 W_n(p_\alpha, p_\beta)}, \]
Substituting the last inequality (21) into (20) and in view of the assumptions of this theorem, we derive
\[ E\{ W_{n+1}(p_{n+1}^1, p_{n+1}^2) | \mathcal{F}_{n-1} \} \]
\[ \leq \frac{a_n}{2} W_n(p_\alpha, p_\beta) + \sum_{k=1}^{2} \gamma_n^k W_n(p_\alpha, p_\beta)^{1/2} \]
\[ - \gamma_n \delta_n W_n(p_\alpha, p_\beta)^2 \]
\[ + C A^2 \sum_{k=1}^{2} (\gamma_n^k)^2 + \kappa_n^2. \]
Now let us use the following inequality:
\[ x' \leq (1 - r) x_0 + (r x_0^{-1}) x, \]
which is valid for any \( x, x_0 > 0 \) and any \( r \in [0, 1] \). For \( r = \frac{1}{2} \) we obtain
\[ C \gamma_n^k \gamma_n^{-1} W_n(p_\alpha, p_\beta)^{1/2} \]
\[ \leq C \gamma_n^k \gamma_n^{-1} x_0^{1/2} \]
\[ + C \gamma_n^k \gamma_n^{-1} x_0^{-1/2} W_n(p_\alpha, p_\beta)^{1/2}, \]
by selecting \( x_0^{1/2} \) such that
\[ C \gamma_n^k \gamma_n^{-1} x_0^{1/2} = \delta_n, \]
it follows that
\[ C \gamma_n^k \gamma_n^{-1} W_n(p_\alpha, p_\beta)^{1/2} \]
\[ \leq \frac{C \gamma_n^k \gamma_n^{-1} \kappa_0^2}{2} + \frac{\delta_n}{2} W_n(p_\alpha, p_\beta)^{1/2} \]
\[ + \frac{C (\gamma_n^k)^2}{\gamma_n^2} \]
Using this estimation in (22), we finally obtain
\[ W_{n+1}(p_{n+1}^1, p_{n+1}^2) \leq W_n(p_\alpha, p_\beta)^2 \]
\[ + \frac{C (\gamma_n^k)^2}{\gamma_n^2} \]
\[ + \frac{C (\gamma_n^k)^2}{\gamma_n^2}. \]
The statement of this theorem follows directly from the Robbins-Siegmund (1971) theorem (Poznyak et al. 2000). □

The next theorem concerns the convergence in the mean square sense.

**Theorem 3—Convergence in the mean square sense:** For the adaptive algorithm, assume that the conditions of the previous theorem except for condition (iii) are fulfilled. In addition, suppose that
\[ (|\tau_n - \tau_{n-1}|^2 + |\delta_n - \delta_{n-1}|^2 + |\tau_n \tau_{n-1} - \tau_n \delta_{n-1}|^2) \]
\[ \times \frac{1 + \gamma_n \tau_n^2 \delta_n^{-1}}{\gamma_n \tau_n^2 \delta_n^{-1}} \to 0; \]
then the mixed strategies of the players converge in the mean square sense, that is
\[ E\{ ||p_{n+1}^1 - (p_1(\eta))^*||^2 + ||p_{n+1}^2 - (p_2(\eta))^*||^2 \} \to 0. \]

**Corollary 1:** For the class of the design parameters defined as
\[ \gamma_n = \frac{\gamma_0}{n^2}, \quad \tau_n = \frac{\tau_0}{n^2}, \quad \delta_n = \frac{\delta_0}{n^2}, \]

\[ x' \leq (1 - r) x_0 + (r x_0^{-1}) x, \]
conditions (i)-(iii) of theorem 2 and the additional condition of theorem 3 will be verified if
\[ \gamma + \delta < 1 \quad (\gamma > 0, \quad \tau \geq \delta > 0) \]
and
(i) the convergence with probability one is ensured if
\[ \tau < \gamma - \frac{1}{2}, \quad 1 + 2\tau + \delta < 2 \min \{1 + \delta, 1 + \tau - \delta\} \quad (25) \]
and
(ii) the convergence in the mean square sense is guaranteed if
\[ 2 - \gamma - 2\tau > 0, \quad 2 - \gamma - 4\delta > 0. \quad (26) \]

**Proof:** The proof follows directly from the conditions of the previous theorem and in view of the fact
\[ \sum_{n=1}^{\infty} n^{-\alpha} = \begin{cases} \infty & \text{if } \alpha < 1, \\ \infty & \text{if } \alpha > 1. \end{cases} \]

However, not only is the convergence important but also the speed is essential. For a specific class of the design parameters, the next theorem states the convergence rate of the adaptive game algorithm described above.

**Theorem 4:** Under the conditions of the previous theorems and for the class of design parameters (24), there exists \( \nu \) such that
\[ W_n^* = \sum_{k=1}^{2} \| p_k^* - p_k^*(\zeta) \|^2 \approx o(n^{-\nu}), \]
where the order of the adaptation rate satisfies the constraint
\[ \nu < \nu^*(\gamma, \delta, \tau) \leq \nu^* = \frac{1}{3} \]
\[ \nu^*(\gamma, \delta, \tau) = \min \{2(\gamma - \tau) - 1; 1 - 2\tau + \delta; 1 - 3\delta; 2\tau\}, \]
and the maximum adaptation rate \( \nu^* = \nu^*(\gamma^*, \delta^*, \tau^*) \) is reached for
\[ \tau = \tau^* = \delta = \delta^* = \frac{1}{6}, \quad \gamma = \gamma^* = \frac{5}{6}. \]

**Proof:** The proof follows from lemma 1, and the following inequality
\[ W_n^* \leq 2W_n + 2 \sum_{k=1}^{2} \| p_k^* - p_k^*(\tau_{n-1}, \delta_n) \|^2 \leq 2W_n + 2C\epsilon_n^2. \]
The expression for \( \nu^*(\gamma, \delta, \tau) \) follows from lemma A.3-2 (appendix A) of Poznyak and Najim (1997). First, we show that the case where \( \tau > \delta \) leads to a contradiction. So, if \( \tau = 6 \), the optimal design parameters are the solution of the following constrained optimization problem
\[ \nu^*(\gamma, \delta, \tau) \rightarrow \max \]
over all the parameters \( \gamma, \delta, \tau \) satisfying (29). The solution of this problem is achieved when all terms within the operator \( \min \{\cdot\} \) turn out to be equal, that is
\[ \nu^*(\gamma, \delta, \tau) = \max \min \{2(\gamma - 6) - 1; 1 - 6; 1 - 36; 26\} \]
\[ = \max \min \{1 - 4\delta; 2\delta\} = \frac{1}{6} \text{ under } \delta = \frac{1}{6}. \]

6. Conclusion

This paper was concerned with the development of a new adaptive algorithm for a two-person zero sum repeated game with incomplete information. The objective was the optimization of the limiting average pay-off. The modified Bush–Mosteller reinforcement scheme with a normalization procedure has been used to design of this algorithm. Based on the Lyapunov approach and martingale theory, the convergence with probability one as well as the convergence in the mean square sense have been stated. The convergence rate has been estimated and the optimal values of the design parameters of this learning procedure have also been determined.

**Acknowledgements**

The authors are grateful to the anonymous reviewers for their helpful comments and advice.

**References**


Bush-Mosteller learning for a zero-sum repeated game with random pay-offs

Filari, J., and Vrieze, K., 1979, Competitive Markov Decision Processes (Berlin: Springer-Verlag).


