ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 307-312

Characterization of the absolutely expedient learning algorithms for stochastic automata in a non-discrete space of actions

Carlos Rivero

Departamento de Estadística e Investigación Operativa I, Facultad de Matemáticas, Universidad Complutense de Madrid, 28040, Madrid, Spain, crivero@mat.ucm.es

Abstract

This work presents a learning algorithm to reach the optimum action of an arbitrary set of actions contained in $\mathbb{R}^m$. An initial and arbitrary probability measure on $\mathbb{R}^m$ allow us to select an action and the probability is sequentially updated by a stochastic automaton using the response of the environment to the selected action. We prove that the corresponding random sequence of probability measures converges in law to a probability measure degenerate on the optimum action, with probability as close to one as we desire.

1. Introduction

From a general point of view, learning is defined as a change in behaviour as result of the past experience. From a mathematical point of view, the goal of a learning system is the optimization of a functional not known explicitly (Narendra and Thathachar (1989)).

From the initial works of Bush and Mosteller (1958), Atkinson, Bower and Crothers (1965) and Vorontsova (1965), the learning problem has concentrated the attention of many researchers. The stochastic automaton acting in a stationary random environment, which has received considerable attention in literature (Fu (1966); Lakshmivarahan and Thathachar (1972); Narendra and Thathachar (1989)), is the goal of this work.

Depending on the allowable values of the output of the environment, different models have been defined. If the response of the environment takes on two values the model is called P-model. If the response takes on a finite set of actions, we define the Q-model. Finally, if the response takes on values in a continuum, the model is called S-model. All these learning models deal with a binary space of actions which are the input to the environment. All the learning algorithms in literature are not directly generalized to a finite space of actions. In any case, to my knowledge, there are no previous works concerning a non-discrete space of action. This work presents a stochastic learning automaton when the set of allowable actions is an arbitrary subset of $\mathbb{R}^m$.

The main evaluation criterion for learning algorithm is based on optimality (Lakshmivarahan (1981); Narendra and Thathachar (1989)). When the space of actions is finite, the study of the optimality involves the convergence of a sequence of finite dimensional vectors of probability to a unitary vector (associated with the optimum action). In this work, where the space of actions is non-discrete, the optimality involves the convergence in law of a sequence of probability measures on $\mathbb{R}^m$ to a probability

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 307-312

measure degenerate at the optimum action.

As in the case of a finite account of actions (Lakshmivarahan and Thathachar (1973); Lakshmivarahan and Thathachar (1976a); Lakshmivarahan and Thathachar (1976b)), in the non-discrete case the absolute expediency implies ε-optimality and in this paper a necessary and sufficient condition for the absolute expediency of the learning algorithm is determined.

The generalization of this work to Q- and S-model is straightforward and it could be made as in the case of finite space of actions (Narendra and Thathachar (1989); Najim and Poznyak (1994); Baba (1985)).

2. Reinforcement scheme

Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a set of allowable actions. These actions are performed on an abstract random environment, which responds to the input action by producing an output belonging to the set of allowable output $\mathcal{B} = \{0, 1\}$. An output $\beta = 1$ is identified with an unfavorable response and $\beta = 0$ with a favorable response of the environment. The input $\alpha(n) \in \mathcal{A}$ is applied to the environment at discrete time $t = n$ ($n = 0, 1, 2, 3...$). The output $\beta(n) \in \mathcal{B}$ is the response of the environment at time $t = n$. The set of output $\mathcal{B}$ is probabilistically related to the input actions $\mathcal{A}$ through a set of penalty probabilities $\mathcal{C} = \{c_\alpha | \alpha \in \mathcal{A}\}$, such that c_α represents the unknown probability that the application of an action α to the stationary environment will result in a penalty output $\beta = 1$,

$$c_\alpha = P(\beta(n) = 1 | \alpha(n) = \alpha), \quad \alpha \in \mathcal{A}.$$

Learning involves the performance of experiments on the environment to choose input actions and to use the output data to update the strategy for picking a new action. An automaton is a systematic strategy of choosing the input actions from the outputs of the environment to increase the occurrence of favorable responses, that is, to increase the occurrence of the actions that minimize c_α.

In this work we assume that $(\mathcal{A}, \mathcal{F}, P_n)$ is a probability space and P_n describes the probability of selecting an action of $\mathcal{A}$ at time $t = n$. The stochastic automaton is the mechanism that updates the probability measure P_n to P_{n+1}, using the action $\alpha(n)$ and the response $\beta(n)$ obtained from the environment. Concretely, the automaton that we propose has the following form

$$\begin{aligned} P_{n+1} &= P_n + \Phi(P_n, \alpha(n)), && \text{if } \beta(n) = 0 \\ P_{n+1} &= P_n && \text{if } \beta(n) = 1, \end{aligned} \tag{1}$$

where $\Phi(P_n, \alpha(n))$ is a σ-additive, non null and finite set function, such that for all $A \in \mathcal{A}$ it fulfils

$$\begin{aligned} \Phi(P_n, \alpha)(A) &\geq 0 \text{ if } \alpha \in A \\ \Phi(P_n, \alpha)(A) &\leq 0 \text{ if } \alpha \notin A. \end{aligned}$$

This is a type of conservative stochastic automaton that changes the probability of selecting an action only when the response of the environment is favorable. Note that if P_n is degenerate then $\Phi(P_n, \alpha) \equiv 0$, for every $\alpha \in \mathcal{A}$. From the Jordan decomposition

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 307-312

we can write

$$\Phi(P_n,\alpha) = \varepsilon(P_n,\alpha)(\delta(\alpha) - \Psi(P_n,\alpha)), \tag{2}$$

where $0 < \varepsilon(P_n,\alpha) < 1$, $\Psi(P_n,\alpha)$ is a probability measure on $(\mathcal{A},\mathcal{F})$ and $\delta(\alpha)$ is a probability measure degenerate at $\alpha \in \mathcal{A}$. Note that if P_n is degenerate then $\varepsilon(P_n,\alpha) = 0$, for every $\alpha \in \mathcal{A}$.

Throughout the rest of the paper we assume that $\varepsilon(P_n,\alpha)$ and $\Psi(P_n,\alpha)$ do not depend of α, that is $\varepsilon(P_n,\alpha) = \varepsilon(P_n)$ and $\Psi(P_n,\alpha) = \Psi(P_n)$. Assuming this condition, the stochastic automaton proposed in (1) generalizes the usual definition of the reinforcement scheme of the stochastic automata in a finite space of actions (Narendra and Thathachar (1989); Lakshmivarahan (1973)).

3. Characterization of the absolutely expedient schemes

The average penalty for a given action probability measure is a random value defined as

$$M(n) = E(\beta(n)|P_n) = \int_{\mathcal{A}} c_\alpha P_n(d\alpha).$$

A learning automaton is said to be absolutely expedient if

$$E(M(n+1)|P_n) < M(n),$$

for every $n \in \mathbb{N}$ and penalty probabilities $\{c_\alpha : \alpha \in \mathcal{A}\}$. The absolute expediency assures that the expectation of the average penalty is decreasing, that is,

$$E(M(n+1)) < E(M(n)).$$

In this section a necessary and sufficient condition for absolute expediency is given. The following theorem generalizes the characterization given in Lakshmivarahan and Thathachar (1973) for a finite space of actions.

Theorem 1 *A learning automaton given by the general reinforcement scheme (1) is absolutely expedient, if and only if, $\Psi(P_n) = P_n$. In this case, the reinforcement algorithm is given by*

$$P_{n+1} = [1 - (1-\beta(n))\,\varepsilon(P_n)]\,P_n + [(1-\beta(n))\,\varepsilon(P_n)]\,\delta(\alpha).$$

PROOF: The reinforcement scheme (1) can be written as

$$P_{n+1} = P_n + (1-\beta(n))\,\Phi(P_n,\alpha(n))$$

and, then

$$E(M(n+1) - M(n)|P_n) = E\left(\int_{\mathcal{A}} c_\alpha(1-\beta(n))\Phi(P_n,\alpha(n))(d\alpha)|P_n\right).$$

From the decomposition (2) we can write

$$\begin{aligned} & E(M(n+1) - M(n)|P_n) \\ = & \int_{\mathcal{A}}(1-c_{\alpha_1})\int_{\mathcal{A}} c_\alpha \Phi(P_n,\alpha_1)(d\alpha)P_n(d\alpha_1) \\ = & \varepsilon(P_n)\int_{\mathcal{A}}(1-c_\alpha)c_\alpha P_n(d\alpha) - \varepsilon(P_n)\int_{\mathcal{A}}(1-c_\alpha)P_n(d\alpha)\int_{\mathcal{A}} c_\alpha \Psi(P_n)(d\alpha). \end{aligned}$$

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 307-312

SUFFICIENT CONDITION: If the condition $\Psi(P_n) = P_n$ is fulfilled, the algorithm is absolutely expedient, since

$$E\left(M(n+1) - M(n)|P_n\right) = -\varepsilon(P_n)\left[\int_{\mathcal{A}} c_\alpha^2 P_n(d\alpha) - \left(\int_{\mathcal{A}} c_\alpha P_n(d\alpha)\right)^2\right] \leq 0.$$

The equality is reached if and only if $c_\alpha = c$, P_n-a.s., that is, if and only if there exists a constant $c \in [0,1]$, such that $P_n\{\alpha \in \mathcal{A} : c_\alpha = c\} = 1$.
NECESSARY CONDITION: Assume that $E\left(M(n+1) - M(n)|P_n\right) \leq 0$, where the equality is fulfilled if and only if $c_\alpha = c, P_n$-a.s. Then

$$\int_{\mathcal{A}} (1 - c_\alpha) c_\alpha P_n(d\alpha) - \int_{\mathcal{A}} (1 - c_\alpha) P_n(d\alpha) \int_{\mathcal{A}} c_\alpha \Psi(P_n)(d\alpha) \leq 0,$$

for every set of penalty probabilities $\{c_\alpha : \alpha \in \mathcal{A}\}$. Take an arbitrary $A \in \mathcal{F}$, such that $0 < P_n(A) < 1$, and define $c_\alpha = x$ if $\alpha \in A$ and $c_\alpha = 1 - x$ if $\alpha \notin A$, for a fixed $0 < x < 1$. Then, the quadratic function

$$\begin{aligned} & H(x) \\ = \; & x(1-x) - \left(P_n(A) + x(1 - 2P_n(A))\right)\left(1 - \Psi(P_n)(A) + x(2\Psi(P_n)(A) - 1)\right) \leq 0 \end{aligned}$$

has an unique maximum at $x = 1/2$. Then $\frac{d}{dx} H(x)\,|_{x=1/2} = 0$, produces $\Psi(P_n)(A) = P_n(A)$, for every $A \in \mathcal{F}$.■

4. Convergence and optimality

When the learning algorithm (1) is absolutely expedient the sequence $\{M(n)\}$ is a non-negative, upper bounded submartingale. Then, $M(n) \overset{a.s.}{\to} M(\infty)$ and

$$\varepsilon(P_n)\left[\int_{\mathcal{A}} c_\alpha^2 P_n(d\alpha) - \left(\int_{\mathcal{A}} c_\alpha P_n(d\alpha)\right)^2\right] \overset{a.s.}{\to} 0. \tag{3}$$

Let us assume the following conditions:
(i) $\varepsilon(P)$ is continuous. $\varepsilon(P) = 0$ if and only if P is a degenerate probability measure.
(ii) $c(\alpha) = c_\alpha$ is a continuous function, such that, $c(\mathcal{A})$ is a compact subset contained on $[0,1]$ and for every fixed $b \in [0,1]$ the set $\{\alpha \in \mathcal{A} | c(\alpha) = b\}$ is finite.

The following theorem shows that under the absolutely expedient condition the algorithm (1), almost surely, converges in law to a degenerate probability measure.

Theorem 2 *Under conditions (i) and (ii), if the scheme (1) is absolutely expedient, the random sequence of probability measures $\{P_n\}$ converges in law to P, almost surely, where P is degenerate or there exists $b \in [0,1]$, such that $P\{\alpha \in \mathcal{A} | c(\alpha) = b\} = 1$.*

PROOF: Under absolute expediency, the convergence (3) is fulfilled almost surely. If $\varepsilon(P_n) \to 0$ the theorem is obvious. Otherwise,

$$E_n(c_\alpha^2) - \left(E_n(c_\alpha)\right)^2 = \int_{\mathcal{A}} \left(c_\alpha - E_n(c_\alpha)\right)^2 P_n(d\alpha) \to 0,$$

where $E_n(c_\alpha) = \int_{\mathcal{A}} c_\alpha P_n(d\alpha)$. Using the Tchebychev Inequality, for every fixed $\delta > 0$

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 307-312

and $\varepsilon > 0$, we can obtain

$$P_n\left(\alpha \in \mathcal{A} : |c_\alpha - E_n\left(c_\alpha\right)| \geq \delta\right) < \varepsilon, \quad \forall n \geq n_0.$$

Since $E_n\left(c_\alpha\right) = M(n) \overset{a.s.}{\rightarrow} M(\infty)$, we can write

$$P_n\left(\alpha \in \mathcal{A} : |c_\alpha - M(\infty)| \geq \delta\right) < \varepsilon, \quad \forall n \geq m_0$$

and

$$\lim_{n\rightarrow\infty} P_n(\alpha \in \mathcal{A} : |c_\alpha - M(\infty)| \geq \delta) = 0. \tag{4}$$

Let us denote

$$\{\alpha_1, ..., \alpha_k\} = \{\alpha \in \mathcal{A} : c_\alpha = M(\infty)\}.$$

Let $g\left(\alpha\right)$ be an arbitrary continuous and bounded real function for $\alpha \in \mathcal{A}$. From (4) it is easy to prove that

$$\lim_{n\rightarrow\infty} \int_{\mathcal{A}} g(\alpha) P_n(d\alpha) = \sum_{i=1}^{k} g(\alpha_i) \lim_{n\rightarrow\infty} P_n(B_{\alpha_i}^{\varepsilon}),$$

where $B_{\alpha_i}^{\varepsilon}$ are disjoint neighbourhood of α_i, such that

$$\{\alpha : |c_\alpha - M\left(\infty\right)| < \delta\} = \bigcup_{i=1}^{k} B_{\alpha_i}^{\varepsilon}$$

and

$$|g(\alpha) - g(\alpha_i)| < \varepsilon, \quad \text{for every } \alpha \in B_{\alpha_i}^{\varepsilon}.$$

Since $\lim_{n\rightarrow\infty} P_n\left(\bigcup_{i=1}^{k} B_{\alpha_i}^{\varepsilon}\right) = 1$ for every $\varepsilon > 0$, also $\sum_{i=1}^{k} \lim_{n\rightarrow\infty} P_n(B_{\alpha_i}^{\varepsilon}) = 1$, for every $\varepsilon > 0$. Note that $B_{\alpha_i}^{\varepsilon}$ decrease as $\varepsilon \rightarrow 0$ and then the former property implies that $\lim_{n\rightarrow\infty} P_n(B_{\alpha_i}^{\varepsilon}) = p_i$, without dependency of ε in the limit.

In conclusion, we have proved that

$$\lim_{n\rightarrow\infty} \int_{A} g(\alpha) P_n(d\alpha) = \sum_{i=1}^{k} g(\alpha_i) p_i,$$

for every function $g(\alpha)$. This is equivalent to saying that $P_n \overset{L}{\rightarrow} P$, where P is a probability measure concentrated in $\{\alpha_1, ..., \alpha_k\}$ and $P\left(\{\alpha_i\}\right) = p_i$, for every $i = 1, ..., k$.■

Concerning the optimality of the learning algorithm, we say that a reinforcement algorithm is optimal if $P_n \overset{L}{\rightarrow} P$, where P is a probability measure degenerate at the points $\{\alpha_0 \in \mathcal{A} : c_{\alpha_0} = \inf_{\alpha \in \mathcal{A}} c\left(\alpha\right)\}$.

In the former theorem we have proved that the absolutely espedient algorithm (1) converges to a degenerate probability measure. The next theorem shows that it is always possible to select step sizes $\varepsilon\left(P_n\right)$, such that the limit probability P is degenerate at the optimum actions, with probability as close to one as we desire.

Theorem 3 *Let B_0 be an arbitrary neighbourhood of the optimum actions and fix a value $0 < p < 1$. Then we can select step sizes $\varepsilon\left(P_n\right)$ such that the probability that $P_n \overset{L}{\rightarrow} P$ is greater than p, where P is a degenerate probability measure concentrated*

on B_0 (that is, $P(B_0) = 1$).

References

1. Atkinson, R. C., Bower, G. H. and Crothers, E. J., *An Introduction to Mathematical Learning Theory*, New York: Wiley, 1965.
2. Baba, N., *New Topics in Learning Automata Theory and Applications*, Number 71 in Lecture Notes in Control and Information Sciences, Berlin: Springer-Verlag, 1985.
3. Bush, R. R., and Mosteller, F., *Stochastic Models for Learning*, New York: Wiley, 1958.
4. Fu, K. S., and McMurtry, G. J., A variable structure automaton used as a multimodal searching technique, *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 379-387, July 1966.
5. Lakshmivarahan, S. and Thathachar M. A. L., Optimal non-linear reinforcement scheme for stochastic automaton, Inform. Sci. vol. 4, pp. 121-128, 1972.
6. Lakshmivarahan, S. and Thathachar M. A. L., Absolutely Expedient Learning Algorithms for Stochastic Automata, *IEEE Transactions on Systems, Man and Cybernetics*, pp. 281-286, 1973.
7. Lakshmivarahan, S., *Learning Algorithms: Theory and Applications*, New York: Springer-Verlag, 1981.
8. Lakshmivarahan, S. and Thathachar M. A. L., Bounds on the convergence probabilities of Learning Automata, *IEEE Transactions on Systems, Man and Cybernetics*, pp. 756-763, 1976.
9. Lakshmivarahan, S. and Thathachar M. A. L., Absolute Expediency of Q- and S-model Learning Algorithms, *IEEE Transactions on Systems, Man and Cybernetics*, pp. 222-226, 1976.
10. Najim, K. and Poznyak, A. S., *Learning Automata: Theory and Application*, Tarrytown, NY: Elsevier Science Ltd., 1994.
11. Narendra, K. and Thathachar M. A. L., *Learning Automata: An Introduction*, Prentice Hall, Englewood Cliffs, 1989.
12. Vorontsova, I. P., Algorithms for Changing Automaton Transition Probabilities, *Problemi Peredachii Informarsii*, vol. 1, pp. 122-126, 1965.