Lecture Notes on Stochastic Learning Automata
Lecture #2

Aly El-Osery

1 Reinforcement Schemes

In the previous lecture the basic operation of an SLA has been discussed. In this lecture some of the reinforcement schemes are presented.

1.1 Reward-Inaction Reinforcement Scheme (L_{R-I})

Assume that $u(n) = u_i$.
If $y(n) = 0$,

$$P_i(n+1) = (1-\alpha)P_i(n) + \alpha, \tag{1}$$

$$P_j(n+1) = (1-\alpha)P_i(n), \qquad (j \neq i) \tag{2}$$

If $y(n) = 1$,

$$P_i(n+1) = P_i(n), \qquad (i = 1, \ldots, m) \tag{3}$$

where

$$P_1(0) = \ldots = P_m(0) = \frac{1}{m} \tag{4}$$

The above reinforcement scheme is ϵ-optimal in the general stationary random environment. The L_{R-I} has the drawback in the point that the state probability vector $P(n)$ is not altered when the environment response at time n is penalty $y(n) = 1$. In the next subsection the general class of absolutely expedient learning algorithms which take penalty inputs from the random environment into account.

1.2 Absolutely Expedient Algorithm

Assume that $u(n) = u_i$.
If $y(n) = 0$,

$$P_i(n+1) = P_i(n) + \sum_{j \neq i} \xi_i(P(n)), \tag{5}$$

$$P_j(n+1) = P_j(n) - \xi_j(P(n)), \qquad (j \neq i) \tag{6}$$

If $y(n) = 1$,

$$P_i(n+1) = P_i(n) - \sum_{j \neq i} \zeta_i(P(n)), \tag{7}$$

$$P_j(n+1) = P_j(n) + \zeta_j(P(n)), \qquad (j \neq i) \tag{8}$$

Theorem 1
A necessary and sufficient condition for the stochastic automaton with the above reinforcement scheme to be absolutely expedient is

$$\frac{\xi_1(P(n))}{P_1(n)} = \ldots = \frac{\xi_m(P(n))}{P_m(n)} = \phi(P) \tag{9}$$

$$\frac{\zeta_1(P(n))}{P_1(n)} = \ldots = \frac{\zeta_m(P(n))}{P_m(n)} = \psi(P) \tag{10}$$

where $\phi(P)$ and $\psi(P)$ are arbitrary continuous functions satisfying

$$0 < \phi(P) < 1 \tag{11}$$

and

$$0 < \psi(P) < \min\left(\frac{P_j}{1-P_j}\right), \quad \textit{for all } j = 1, \ldots, m. \tag{12}$$

The L_{R-I} algorithm is included in this class of algorithms, i.e., let $\xi_j(P(n)) \triangleq \alpha P_j(n)$ and $\zeta_j(P(n)) \triangleq 0$. As an example of the absolutely expedient algorithm is the following nonlinear reinforcement scheme.

1.2.1 Nonlinear Reinforcement Scheme

Assume that $u(n) = u_i$.
If $y(n) = 0$,

$$P_i(n+1) = (1-\alpha)P_i(n) + \alpha, \tag{13}$$

$$P_j(n+1) = (1-\alpha)P_j(n) \qquad (j \neq i) \tag{14}$$

If $y(n) = 1$,

$$P_i(n+1) = P_i(n) - k\alpha(1 - P_i(n))\left(\frac{H}{1-H}\right), \tag{15}$$

$$P_j(n+1) = P_j(n) + k\alpha P_j(n)\left(\frac{H}{1-H}\right) \qquad (j \neq i) \tag{16}$$

where

$$H = \min[P_1(n), \ldots, P_m(n)], \tag{17}$$

$$0 < \alpha < 1, \tag{18}$$

$$0 < k\alpha < 1, \tag{19}$$

$$P_1(0) = \ldots = P_m(0) = \frac{1}{m} \tag{20}$$

2 Multi-Teacher Environment

– Up to this point we have discussed only a single-teacher environment.

– However, learning behaviors of stochastic automata under a single teacher environment cannot be applied to problems where one input elicits multi-responses from the environment having multi-criteria. Many practical problems exhibits this behavior. For these cases multi-teacher environment needs to be considered.

– A basic model of a multi-teacher environment is shown in Figure 1.

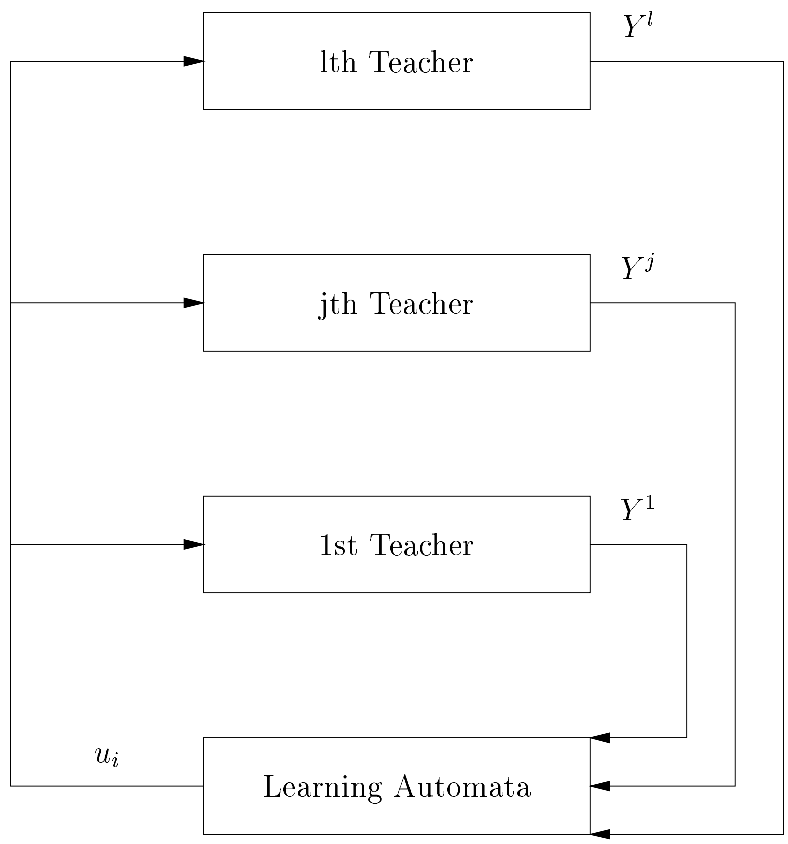

Figure 1: Stochastic automaton operating in l-teacher environment

– This case is a little different than the single-teacher environment in that the different responses can be given by the different environments.

– In multi-teacher environment, an action of the automaton should receive a greater reward from the l-teacher environment the more teachers agree with it. This leads to the definition of *average weighted* reward.

Definition 1
The average weighted reward in the l-teacher environment $W(n)$ is defined as follows:

$$W(n) = \sum_{i=1}^{m} \left[P_i(n) \left(\sum_{j=1}^{l} j D_i^{lj} \right) \right] \tag{21}$$

where D_i^{lj} is the probability that j teachers approve of the ith action u_i of the stochastic automaton.

2.1 Absolutely Expedient Nonlinear Reinforcement Schemes in the General l-Teacher Environment (GAE)

When the output of the stochastic automaton at time step n is u_i and the responses from the multi-teacher environment are r rewards and $l-r$ penalties, the state probability vector $P(n)$ is transformed as follows:

$$P_i(n+1) = P_i(n) + (1 - \frac{r}{l}) \sum_{j \neq i}^{m} \kappa_i(P(n)) - \frac{r}{l} \sum_{j \neq i}^{m} \eta_i(P(n)), \tag{22}$$

$$P_j(n+1) = P_j(n) - (1 - \frac{r}{l}) \kappa_j(P(n)) + \frac{r}{l} \eta_j(P(n)), \qquad (j \neq i) \tag{23}$$

where

$$\frac{\kappa_1(P(n))}{P_1(n)} = \ldots = \frac{\kappa_m(P(n))}{P_m(n)} = \phi(P(n)) \tag{24}$$

$$\frac{\eta_1(P(n))}{P_1(n)} = \ldots = \frac{\eta_m(P(n))}{P_m(n)} = \psi(P(n)), \tag{25}$$

$$P_j(n) + \eta(P_j(n)) > 0, \tag{26}$$

$$P_i(n) + \sum_{j \neq i}^{m} \kappa(P(n)) > 0, \tag{27}$$

and

$$P_j(n) - \eta(P_j(n)) < 1, \tag{28}$$

for $j = 1, \ldots, m$ and $i = 1, \ldots, m$.

Theorem 2
If

$$\phi(P(n)) \leq 0, \tag{29}$$

$$\psi(P(n)) \leq 0, \tag{30}$$

and

$$\phi(P(n)) + \psi(P(n)) < 0 \tag{31}$$

Then, the stochastic automaton with the reinforcement scheme defined by the GAE algorithm is absolutely expedient in the general l-teacher environment.

2.1.1 Algorithm 1- (GL_{R-I})

Let $\kappa_j \triangleq 0$ and $\eta_j \triangleq -l\alpha P_j(n)$ for $(j = 1, \ldots, m)$, then

$$P_i(n+1) = (1 - r\alpha)P_i(n) + r\alpha, \tag{32}$$

$$P_j(n+1) = (1 - r\alpha)P_j(n) \qquad (j \neq i) \tag{33}$$

where $0 < l\alpha < 1$.

2.1.2 Algorithm 2- (GNA)

Let $\eta_j \triangleq -\alpha P_j(n)$ and $\kappa_j \triangleq -k\alpha P_j(n)\{H/(1-H)\}$ for $(j = 1, \ldots, m)$, then the reinforcement scheme becomes

$$P_i(n+1) = P_i(n) - k\alpha \left(1 - \frac{r}{l}\right)(1 - P_i(n)) \left\{\frac{H}{1-H}\right\} + \alpha \left(\frac{r}{l}\right)(1 - P_i(n)), \tag{34}$$

$$P_j(n+1) = P_j(n) + k\alpha \left(1 - \frac{r}{l}\right) P_i(n) \left\{\frac{H}{1-H}\right\} - \alpha \left(\frac{r}{l}\right) P_j(n) \tag{35}$$

where $0 < \alpha < 1$, $H = \min[P_1(n), \ldots, P_m(n)]$ and $0 < k\alpha < 1$.

3 Nonstationary Multi-Teacher Environment

– Up to this point, only multi-teacher environment is stationary and P-model.

– In this section nonstationary multi-teacher environment from which stochastic automata receives responses having any arbitrary number between 0 and 1 (S-model).

– In the nonstationary case the outputs are function of time and state.

– Depending upon the action u_i and the n responses $y_i^1(n,q), \dots, y_i^n(n,q)$ from the multi-teacher environment, the stochastic automaton changes the probability vector $P(n)$ by the reinforcement scheme.

– The objective of the stochastic automaton is to reduce the expectation of the sum of the penalty strengths given by,

$$I = \mathbb{E}\left\{\sum_{j=1}^{n} y_i^j(n,q)\right\} \tag{36}$$

3.1 ϵ-Optimal Reinforcement Scheme Under Nonstationary Multi-teacher Environment (MGAE)

Let $u(n) = u_i$ and the responses from the l-teacher environment are $(y_i^1, \dots, y_i^l)$. Then,

$$\begin{aligned} P_i(n+1) =& P_i(n) + \left(\frac{y_i^1 + \ldots + y_i^l}{l}\right)\sum_{j\neq i}^{m} \kappa_i(P(n)) + \\ & - \left(1 - \frac{y_i^1 + \ldots + y_i^l}{l}\right)\sum_{j\neq i}^{m} \eta_i(P(n)), \end{aligned} \tag{37}$$

$$\begin{aligned} P_j(n+1) =& P_i(n) - \left(\frac{y_i^1 + \ldots + y_i^l}{l}\right)\kappa_j(P(n)) + \\ & + \left(1 - \frac{y_i^1 + \ldots + y_i^l}{l}\right)\eta_j(P(n)), \end{aligned} \tag{38}$$

where

$$\frac{\kappa_1(P(n))}{P_1(n)} = \ldots = \frac{\kappa_m(P(n))}{P_m(n)} = \phi(P(n)) \tag{39}$$

$$\frac{\eta_1(P(n))}{P_1(n)} = \ldots = \frac{\eta_m(P(n))}{P_m(n)} = \psi(P(n)), \tag{40}$$

$$P_j(n) + \eta(P_j(n)) > 0, \tag{41}$$

$$P_i(n) + \sum_{j \neq i}^{m} \kappa(P(n)) > 0, \tag{42}$$

and

$$P_j(n) - \eta(P_j(n)) < 1, \tag{43}$$

for $j = 1, \ldots, m$ and $i = 1, \ldots, m$. The MGAE scheme is a generalized form of the GAE scheme given in the previous section where r is replaced by $(l - [y_i^1 + \ldots + y_i^l])$.

Homework Problem:
Given two performance functions given below:

$$J_1(x) = -(x-3)^2 + 10,$$

and

$$J_2(x) = -2x + 12.$$

Assume that only a noise corrupted observations are available as follows:

$$g_j(x, \omega) = J_j(x) + \omega_j, \qquad j = 1, 2$$

where ω_i is an additive white gaussian zero mean noise with variance 0.1.

1. Plot the two objective functions for $x = 1, \ldots, 5$.
2. Design a stochastic learning automaton having five actions, i.e., $u_i \in \{1, \ldots, 5\}$, to optimize (maximize) both of the optimization functions.
3. Plot the probabilities of all of the actions.
4. Compare the performance of GL_{R-I} and GNA schemes with different values of α.

Hint:
Let $\nu^j(n)$ be a measurement of $g_j(x, \omega_j)$, and

$$\overline{\nu}^j(n) = \frac{1}{n+1}(n\overline{\nu}^j(n-1) + \nu^j(n))$$

If $\nu^j(n) > \overline{\nu}^j(n-1)$, then $y_i^j = 0$, on the other hand if $\nu^j(n) < \overline{\nu}^j(n-1)$, then $y_i^j = 1$