View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Massey Research Online

Res. Lett. Inf. Math. Sci., 2005, Vol. **7**, pp 43-88
Available online at http://iims.massey.ac.nz/research/letters/

Analysis of reinforcement learning strategies for predation in a mimic-model prey environment

A.TSOULARIS[1,] & J.WALLACE[2]

[1]*I.I.M.S., Massey University, Albany Campus, Auckland, New Zealand*
A.D.Tsoularis@massey.ac.nz
[2]*School of Management, University of Bradford, Bradford, United Kingdom*
J.Wallace1@bradford.ac.uk

In this paper we propose a mathematical learning model for a stochastic automaton simulating the behaviour of a predator operating in a random environment occupied by two types of prey: palatable mimics and unpalatable models. Specifically, a well known linear reinforcement learning algorithm is used to update the probabilities of the two actions, eat prey or ignore prey, at every random encounter. Each action elicits a probabilistic response from the environment that can be either favorable or unfavourable. We analyse both fixed and varying stochastic responses for the system. The basic approach of mimicry is defined and a short review of relevant previous approaches in the literature is given. Finally, the conditions for continuous predator performance improvement are explicitly formulated and precise definitions of predatory efficiency and mimicry efficiency are also provided.

1. Introduction

A Batesian mimic is a palatable prey who gains protection from a predator through resemblance to an unpalatable species, the model. It is believed that Batesian mimicry generally harms the model by degrading its natural defensive advantage conferred by the aposematic signal. If both mimics and models are unpalatable then they both should gain protection from the predator. This type of mimicry is known as Muellerian.

A simple mathematical model of the model-mimic situation was introduced by Huheey (1964). Huheey made the fundamental assumption that the predator will reject all prey for a number of encounters following an unfavourable consumption of a model. A more sophisticated mathematical model involving a mimic, a model, and a predator was later put forward by Estabrook and Jespersen (1974). They adopted Huheey's idea of a waiting period between encounters and constructed a Markov chain to model the probabilistic encounters between the predator and the models and mimics. Bobisud and Potratz (1976) suggested that a more plausible strategy for the predator would be to remember the number of successive mimics that have been consumed after a model is consumed, and to modify its reaction on the basis of this information. Arnold (1978) investigated the avoidance behaviour of a predator in relation to the spatial prey distribution and degree of noxiousness of the models. Luedeman et al. (1981) proposed a Markov chain model that included alternative prey in addition to models and

mimics. Kannan (1983) presented an extensive analysis of three types of predator strategies: single-trial in the spirit of Estabrook and Jespersen, multi-trial in the spirit of Bobisud and Potratz, and a consume-everything strategy. Owen and Owen (1984) proposed a strategy for the predator based on recurrent sampling. Huheey (1988) presented a partial review of the quantitative methods applied to the modelling of mimicry for the previous 15 years. Speed (1993) simulated predatory behaviour based on simple learning and forgetting rules proposed earlier by Turner *et al*. (1984), and by further assuming that learning in prey is inherently Pavlovian in nature. In a series of papers, Turner and Speed (1996, 1999) proposed a generalised model of learning behaviour, based on an algorithm introduced originally by Bush and Mosteller (1955), which encompassed aspects of most of the major models as special cases.

2. The learning automaton approach to predator-model-mimic interactions

The predatory behaviour simulated by Speed and Turner (1999) consists of two main aspects: learning and forgetting. The probability of attack on the prey changes according to the rule

$$p_2 = p_1 + \alpha(\lambda - p_1),$$

where p_1 and p_2 are the attack probabilities prior and after an encounter respectively ($p_1 = p_2$ if there is no attack), α is the learning parameter ($0 \le \alpha \le 1$, with $\alpha = 0$ indicating no learning and $\alpha = 1$ indicating learning in a single trial) and λ is the asymptotic attack probability, $0 \le \lambda \le 1$. The predator starts attacking prey with a naïve probability, $p_0 = 0.5$, which is also the probability of attack after the predator has forgotten what was learnt on previous samplings. The forgetting algorithm is thus

$$p_3 = p_2 + \phi(p_0 - p_2),$$

where p_3 is the attack probability after forgetting has occurred, and ϕ is the associated forgetting parameter. So after forgetting, the frequency of attack undergoes a change

$$\Delta p = p_3 - p_2 = \phi(0.5 - p_2)$$

The further p_2 is from the naïve value, 0.5, the larger the absolute value of Δp is, which implies a larger forgetting rate, yet according to MacDougall and Dawkins (1998) empirical evidence suggests otherwise. Also the assumption made by Speed and Turner (1996, 1999), that a prey is deemed unpalatable if it is attacked asymptotically with probability less than the naïve probability, has

been questioned by Joron and Mallet (1998), who suggested that a more sensible view of palatability would be one that reduces predation upon experience, regardless of the naïve attack rate.

In this paper we model a predator as a learning automaton operating in an environment that manifests itself in the form of either palatable or unpalatable prey (Batesian mimicry). Upon each encounter the predator may choose, with a certain probability, either to consume the prey or to simply ignore it. Consumption of unpalatable models induces the predator to reduce its probability of consumption on the next encounter with a prey, otherwise the attack probability will be reinforced. Each instant the predator chooses to ignore prey it runs the risk of a missed food opportunity, should the ignored prey be palatable. The predator can discriminate prey with a certain probability which is also adjusted accordingly.

Our approach, although broadly in line with that of Turner and Speed (1996, 1999), differs in at least three respects: following the observations of Joron and Mallet, we do not use a fixed naïve attack probability. Moreover, we do not introduce a forgetting rate, rather, we allow the predator to reach an asymptotic consumption probability and an asymptotic probability of prey discrimination. This is achieved through a continuous update of action probabilities on the basis of the response the predator elicits from the environment. The need for the inclusion of an environmental response in learning can be seen thus; if the rewards to the predator for a given action and the proportion of palatable prey were known at each stage, then a payoff matrix could be set up as follows:

$$R = \begin{matrix} \\ eat \\ ignore \end{matrix} \overbrace{\begin{matrix} palatable & unpalatable \end{matrix}} \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$$

If γ is the proportion of palatable prey and p is the probability of eating a prey, the expected payoff to the predator from an encounter with a prey is:

$$\begin{aligned} E[R] = & \; p\left(\gamma r_{11} + (1-\gamma) r_{12}\right) + (1-p)\left(\gamma r_{21} + (1-\gamma) r_{22}\right) = \\ & p\left(\gamma (r_{11} - r_{21}) + (1-\gamma)(r_{12} - r_{22})\right) + \gamma r_{21} + (1-\gamma) r_{22} \end{aligned}$$

From this, the strategies for maximizing the expected payoff can be seen to be:

$$p = 1 \quad \text{if} \quad \gamma > \frac{r_{22} - r_{12}}{r_{11} - r_{12} - r_{21} + r_{22}},$$

$$p = 0 \quad \text{if} \quad \gamma < \frac{r_{22} - r_{12}}{r_{11} - r_{12} - r_{21} + r_{22}},$$

$$p \text{ any value in } [0,1] \quad \text{if} \quad \gamma = \frac{r_{22} - r_{12}}{r_{11} - r_{12} - r_{21} + r_{22}}.$$

Formulating an action plan therefore requires knowledge of the rewards and the proportion of the palatable population at each stage. Since both items are expected to vary at each stage due to the continuous interaction between the predator and prey, it is unrealistic to assume that such knowledge is readily available. Instead, by probing the environment regularly the predator is expected to learn how often it will choose the right course of action. This provides the motivation for the present study.

Finally, we emphasis that the effect that the learning automaton mode of operation of the predator has on the mimicry system is not directly appraised. Furthermore, we do not consider competition between predators.

Our work is organized as follows: in the next section a brief outline of the learning automaton concept is given. Section 4 considers the predator as a simple stochastic automaton with no learning capacity. Sections 5 and 6 are devoted to learning in fixed and varying environments respectively.

3. The concept of the learning automaton and its mathematical description

Having established the need for environmental feedback, we now present a learning automaton which uses this approach. A **learning automaton** is a deterministic or stochastic algorithm used in discrete-time systems to improve their performance in random environments (Narendra and Thathachar, 1989). A finite number of decisions (actions) are available to the system to which the environment responds either favourably or unfavourably. The purpose of the learning automaton is to increase the probability of selecting an action that is likely to elicit a favourable response based on past actions and responses. The automaton and the environment in which it operates are connected in a feedback manner as illustrated in figure 1 below:

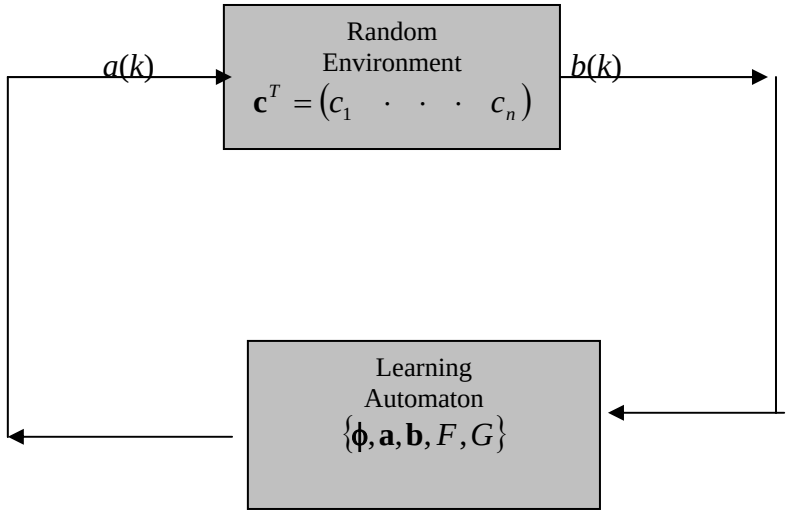


Fig.1. Feedback configuration of automaton and environment.

The system operating in a random environment can choose to perform an action from the finite set $\mathbf{a} = \{a_1, \cdots, a_n\}$. The chosen action constitutes the input to the environment which responds with an output from the set $\mathbf{b} = \{b_1, \cdots, b_m\}$. The environment is categorized as a *P*-model if the output set is binary (m = 2), a *Q*-model if m is finite and greater than 2, and as an *S*-model if **b** is a continuous random variable such that b ε (b_1, b_2), where, $b_1 \leq b_2$. When the response of the environment is unambiguously favourable or unfavourable, then a *P*-model is sufficient to describe it. When the number of possible responses is greater than 2, then the outputs are neither totally favourable nor totally unfavourable, and a *Q*- or *S*-model is needed to describe the environment. The behaviour of the environment can therefore be captured by the set, **c**, of penalty probabilities, associated with each of the potential actions available to the automaton:

$$c_i(k) = \textit{probability}(b(k) = \textit{penalty} \mid a(k) = a_i), \quad i = 1, \cdots, n$$

where $c_i(k)$ denotes the probability that the environment will respond unfavourably to the action a_i at the k^{th} encounter. If the penalty probabilities, c_i, are independent of k, the environment is stationary. The penalty probabilities are assumed unknown however, for knowledge of them would render the problem of the system (automaton) operating in a random environment, relatively trivial.

The internal structure of a stochastic automaton is characterized by the state set, $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_r\}$. The r x r stochastic state transition matrix, $F^{(\ell)} = \left(f_{ij}^{(\ell)}\right)$, determines the state at the $(k + 1)$ encounter in terms of the state and a random input, b_ℓ, at k:

$$f_{ij}^{(\ell)} = probability\left(\phi(k+1) = \phi_j \mid \phi(k) = \phi_i, b(k) = b_\ell\right) \quad i, j = 1, \cdots, r;\ \ell = 1, \cdots, m$$

The transition between states is thus an ergodic Markov chain, with the final state probabilities, $p(\phi_i) = \pi_i$, $i = 1, \cdots, r$ given by

$$\boldsymbol{\pi} = F^T \boldsymbol{\pi}$$

The r x n stochastic output matrix, $G = \left(g_{ij}\right)$, determines the action of the automaton at any encounter k in terms of the state at that encounter:

$$g_{ij} = probability\left(a(k) = a_j \mid \phi(k) = \phi_i\right), \quad i = 1, \cdots, r \quad j = 1, \cdots, n$$

The final action probabilities, $p(a_i) = p_i^*$, $i = 1, \cdots, n$ are given by

$$\mathbf{p}^* = G^T \boldsymbol{\pi}$$

For a deterministic automaton the entries of the matrices F and G are either 0 or 1. It is convenient in many cases to identify each state with a distinct action so that $r = n$ and G, after suitable reordering, is the identity matrix. As a consequence, here, $\mathbf{p}^* = \boldsymbol{\pi}$. A **fixed-structure stochastic automaton** is characterized by matrices $F^{(l)}$ and G independent of k. When the transition probabilities, $f_{ij}^{(\ell)}$ or g_{ij}, are updated at each k on the basis of the response, $b(k) = \ell$, of the environment, the automaton is called a **variable-structure stochastic automaton**. The basic idea behind the update is to increase the action probability that produces a favourable response and decrease all others; for an unfavourable response, the respective action probability is decreased and all others are increased. The algorithm for update is called a **reinforcement scheme**.

The average penalty incurred by the automaton, $M(k)$, conditioned on the state corresponding to the action probability vector, $p(k)$, at encounter k is given (Narendra and Thathachar, (1989)) by the expectation:

$$M(k) = E[b(k)\ \ is\ \ unfavourable \mid \mathbf{p}(k)] = \sum_{i=1}^{n} c_i\, p_i(k)$$

If all actions are equally likely, such an automaton is called a pure-chance automaton and suffers the expected penalty:

$$M_0 = \frac{1}{n}\sum_{i=1}^{n} c_i$$

Obviously, for an automaton to perform better than a pure-chance automaton, its expected penalty must be less than M_0. Since $M(k)$ is a random variable, we

need to examine its long term average input, $E[M(k)]$ as $k \to \infty$. A learning automaton is said to be **expedient** if:

$$\lim_{k\to\infty} E[M(k)] < M_0$$

and **optimal** if

$$\lim_{k\to\infty} E[M(k)] = \min_i c_i, \quad i = 1, \cdots, n$$

Optimality is meant to imply that the action associated with the minimum penalty probability is always chosen in the long term. In practice optimality may be unattainable and a suboptimal performance measure like ***ε*-optimality** may be more appropriate:

$$\min_i c_i < \lim_{k\to\infty} E[M(k)] < \varepsilon + \min_i c_i, \quad i = 1, \cdots, n$$

for some arbitrary $\varepsilon > 0$. Finally, a learning automaton is **absolutely expedient** if its average input monotonically decreases with time, that is if $M(k)$ is a supermartingale:

$$E[M(k+1)] < E[M(k)], \quad \text{for all } k$$

4. Predator as a fixed-structure stochastic automaton operating in stationary random model-mimic environments

A predator operating in an environment filled with unpalatable models, *M*, and palatable mimics, *X*, can be modelled as a stochastic automaton with two actions: a_1 for ignoring any prey and a_2 for consuming a prey. If the prey ignored is a mimic then the predator suffers a penalty due to loss of opportunity with probability c_1 given by:

$$c_1 = p(IX \mid a_1)$$

Similarly, a penalty with probability c_2 is incurred when the predator consumes an unpalatable model:

$$c_2 = p(EM \mid a_2),$$

where the symbols I $(\equiv a_1)$ and E $(\equiv a_2)$ stand for the actions to ignore and eat respectively.

The favourable responses by the environment occur when the predator ignores models and consumes mimics and have respective probabilities d_1 and d_2 given by:

$$d_1 = p(IM \mid a_1) = 1 - c_1$$
$$d_2 = p(EX \mid a_2) = 1 - c_2$$

Let f_{ij} and $\tilde{f}_{ij}$ be the transition probability from state (action) i = 1,2 to state (action) j = 1,2, following a favourable and unfavourable response respectively, that is,

$$f_{ij} = p(a_j \mid a_i = IM \ \ or \ a_i = EX)$$
$$\tilde{f}_{ij} = p(a_j \mid a_i = EM \ or \ a_i = IX)$$

The probability, $\bar{f}_{ij}$, of transition from state i to state j is then given by

$$\bar{f}_{ij} = d_i f_{ij} + c_i \tilde{f}_{ij}$$

Let F and $\tilde{F}$ be the state transition matrices following a favourable and unfavourable response respectively:

$$F = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{matrix} I & E \end{matrix}}^{} \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \qquad \tilde{F} = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{matrix} I & E \end{matrix}}^{} \begin{pmatrix} 1-\tilde{p} & \tilde{p} \\ \tilde{q} & 1-\tilde{q} \end{pmatrix}$$

Then the state transition matrix is as follows:

$$\bar{F} = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{matrix} I & E \end{matrix}}^{} \begin{pmatrix} 1-d_1 p - c_1 \tilde{p} & d_1 p + c_1 \tilde{p} \\ d_2 q + c_2 \tilde{q} & 1-d_2 q - c_2 \tilde{q} \end{pmatrix}$$

The long-term action probabilities, π_1 and π_2, are found from solving $\boldsymbol{\pi} = \bar{F}^T \boldsymbol{\pi}$:

$$\pi_1 = \frac{q + c_2(\tilde{q} - q)}{p + q + c_1(\tilde{p} - p) + c_2(\tilde{q} - q)} \quad , \quad \pi_2 = \frac{p + c_1(\tilde{p} - p)}{p + q + c_1(\tilde{p} - p) + c_2(\tilde{q} - q)}$$

The long term average penalty given, π_1 and π_2, is:

$$\lim_{k \to \infty} M(k) = \frac{c_1 c_2(\tilde{p} + \tilde{q} - p - q) + c_1 q + c_2 p}{p + q + c_1(\tilde{p} - p) + c_2(\tilde{q} - q)}$$

If the predator always ignores mimics and models then $c_1 \to 1$ and $c_2 \to 0$, and the long term average penalty is simply

$$\lim_{k\to\infty} M(k) = \frac{q}{\tilde{p} + q}$$

If the predator always consumes mimics and models then $c_1 \to 0$ and $c_2 \to 1$, and the average penalty is simply

$$\lim_{k\to\infty} M(k) = \frac{p}{p + \tilde{q}}$$

This "consume-all-prey" policy can be compared to the ignore-all-prey policy on the basis of the magnitude of the respective asymptotic average penalties:

$p\tilde{p} - q\tilde{q} > 0$ consume-all-prey policy is superior,
$p\tilde{p} - q\tilde{q} < 0$ ignore-all-prey policy is superior,
$p\tilde{p} - q\tilde{q} = 0$ neutral approach, both policies are equivalent.

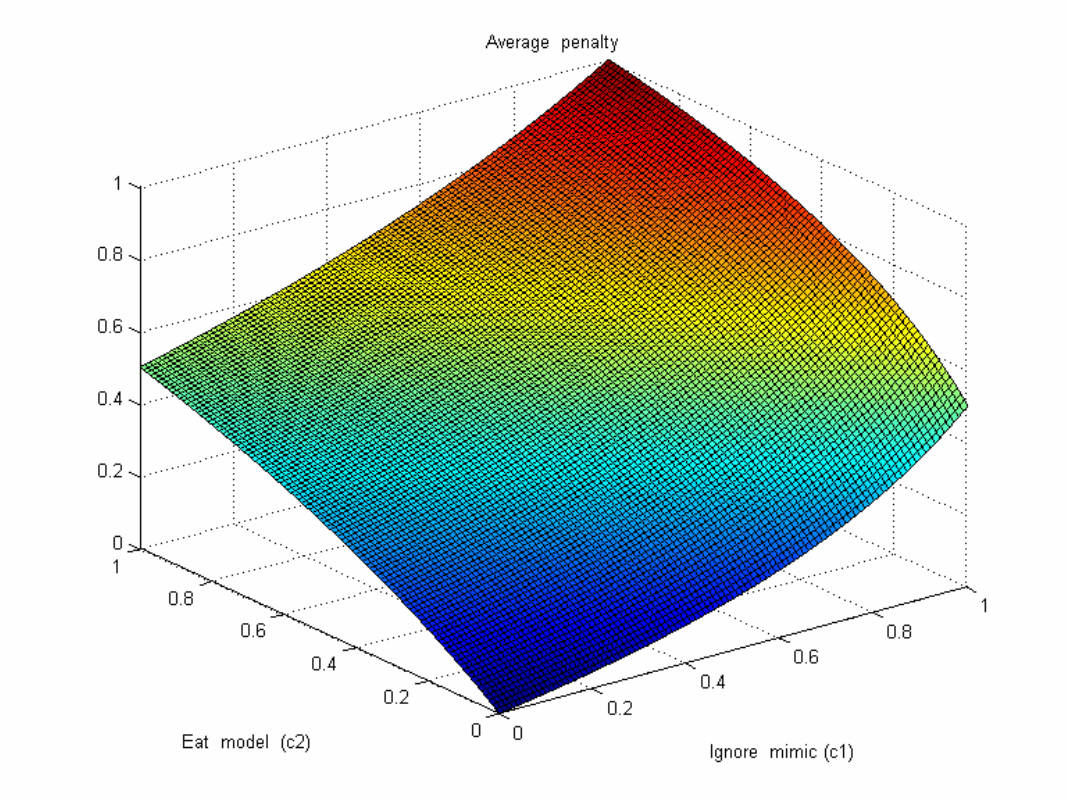


Fig.2. Average penalty surface plot for $p = 0.8, q = 0.2, \tilde{p} = 0.2, \tilde{q} = 0.8$. The predator is, in the long term, indifferent to either ignore all prey or consume all prey policies because $p\tilde{p} = q\tilde{q} = 0.16$. Either policy incurs an average penalty of 0.5.

A measure of the variation rate of the average penalty is afforded by the second derivative, $\frac{\partial^2 M}{\partial c_i^2}$, $i = 1,2$. For $c_2 = 0$,

$$\frac{\partial^2 M}{\partial c_1^2} = -\frac{2q(p+q)(\tilde{p}-p)}{[(p+q)+c_1(\tilde{p}-p)]^3}$$

and for $c_1 = 0$,

$$\frac{\partial^2 M}{\partial c_2^2} = -\frac{2p(p+q)(\tilde{q}-q)}{[(p+q)+c_2(\tilde{q}-q)]^3}$$

When transition between different states is more frequent after a favourable response, that is $p > \tilde{p}$ and $q > \tilde{q}$, the rate of the average penalty increase grows steadily for increasing c_1 and c_2. When the opposite scenario occurs, that is $p < \tilde{p}$ and $q < \tilde{q}$, the average penalty declines steadily for increasing c_1 and c_2. When either $p > \tilde{p}$ and $q < \tilde{q}$ or $p < \tilde{p}$ and $q > \tilde{q}$, the rate of the long

term penalty increase accelerates in one direction and decreases in the other, as displayed in fig. 2 above, where $p = 0.8 > \tilde{p} = 0.2,\ q = 0.2 < \tilde{q} = 0.8$.

For $p = q = \tilde{p} = \tilde{q}$ we get $\pi_1 = \pi_2 = \frac{1}{2}$, and the predator behaves essentially as a pure-chance automaton. Otherwise, expediency is guaranteed, provided

$$c_1 < c_2 \quad \text{and} \quad d_2 q + c_2 \tilde{q} > d_1 p + c_1 \tilde{p}$$

or

$$c_1 > c_2 \quad \text{and} \quad d_2 q + c_2 \tilde{q} < d_1 p + c_1 \tilde{p},$$

but not optimality, since $\lim_{k\to\infty} M(k) > min\{c_1, c_2\}$ always.

Finally, we need an average measure of the predator's efficiency in distinguishing palatable prey (mimics) from a pool of mixed prey (models and mimics). We define the **asymptotic predatory efficiency index** as the ratio of the probability of consumption of palatable prey to the total probability of encountering palatable prey:

$$e^* = \frac{\pi_2(1-c_2)}{\pi_2(1-c_2)+\pi_1 c_1} = \frac{(1-c_1)(1-c_2)p + c_1(1-c_2)\tilde{p}}{(1-c_1)(1-c_2)p + c_1(1-c_2)(\tilde{p}+q) + c_1 c_2 \tilde{q}}$$

In a complementary manner, we may define the **asymptotic mimicry efficiency index** as the ratio of the probability of ignored palatable prey to the total probability of encountering palatable prey:

$$m^* = \frac{\pi_1 c_1}{\pi_2(1-c_2)+\pi_1 c_1} = 1 - e^*$$

If the transition between states is independent of the environmental response the matrices F and $\tilde{F}$ are identical, and consequently $p = \tilde{p},\ q = \tilde{q}$. The action probabilities and the asymptotic penalty are then

$$\pi_1 = \frac{q}{p+q}\ ,\ \pi_2 = \frac{p}{p+q}\ ,\ \lim_{k\to\infty} M(k) = \frac{c_1 q + c_2 p}{p+q}$$

Expediency in this case simply equates to the action that incurs the least penalty being chosen more often, that is, $\pi_1 > \pi_2$ if $c_1 < c_2$ and $\pi_1 < \pi_2$ if $c_1 > c_2$.

If $p = q$ and $\tilde{p} = \tilde{q}$ (both matrices F and $\tilde{F}$ are doubly stochastic but not necessarily identical), then the automaton is expedient if $p < \tilde{p}$. The predator exhibits expedient behaviour when the frequency of swapping actions following immediately from an unfavourable response is higher than the frequency of swapping actions immediately following a favourable response.

A trivial strategy for the predator would be to switch action with certainty $(\tilde{p} = \tilde{q} = 1)$ whenever an unfavourable response is recorded and continue with the same action $(p = q = 0)$ whenever the response is favourable, in which case

$$F = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}} \qquad \tilde{F} = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}$$

and consequently,

$$\bar{F} = \begin{matrix} \\ I \\ E \end{matrix} \overbrace{\begin{pmatrix} d_1 & c_1 \\ c_2 & d_2 \end{pmatrix}}$$

The two final action probabilities, π_1 and π_2, in this case can also be calculated from the equation, $\boldsymbol{\pi} = \bar{F}^T \boldsymbol{\pi}$:

$$\pi_1 = \frac{c_2}{c_1 + c_2} \quad , \quad \pi_2 = \frac{c_1}{c_1 + c_2}$$

The automaton (predator) is expedient if $c_1 \neq c_2$, since

$$\lim_{k \to \infty} M(k) = c_1 \pi_1 + c_2 \pi_2 = \frac{2 c_1 c_2}{c_1 + c_2} < \frac{c_1 + c_2}{2} = M_0$$

The predatory efficiency is $e^* = 1 - c_2$, and the mimicry efficiency is $m^* = c_2$. Mimicry efficiency represents the rate that the predator consumes the wrong prey. The long term proportion of ignored palatable prey equals the long-term proportion of consumed palatable prey when $c_2 = \frac{1}{2}$, and the long-term proportion of consumed palatable prey equals the long-term proportion of consumed unpalatable prey when $c_1 = \frac{1}{2}$.

Such a fixed-structure, two-state, two-action automaton was first suggested by Tsetlin (Tsetlin 1973) and is known as an $L_{2,2}$ automaton. Despite its obvious simplicity it has found applications in many learning models (Selfridge 1978).

5. Predator as a variable-structure stochastic automaton operating in stationary random model-mimic environments

5.1. Mathematical description of reinforcement schemes

Greater flexibility can be built into modelling the predatory behaviour by considering the predator as a stochastic automaton with state transitions or action probabilities being updated at every stage using a reinforcement scheme.
In general terms, a reinforcement scheme can be represented by either updating the action probability at stage (k+1) on the basis of its previous value, the action $a(k)$ and input $b(k)$:

$$p_i(k+1) = h\left[p_i(k), a(k), b(k)\right], \quad i = 1,2,\cdots,n$$

or by updating the state transition probabilities, $f_{ij}(k+1)$, on the basis of the states, $\phi(k)$ and $\phi(k+1)$, input $b(k)$, and previous state transition probabilities $f_{ij}(k)$:

$$f_{ij}(k+1) = \varphi\left[f_{ij}(k), \phi(k), \phi(k+1), b(k)\right], \quad i, j = 1,2,\cdots,r$$

If either h or φ is linear, the reinforcement scheme is said to be linear; otherwise it is called nonlinear.

The idea of a reinforcement scheme is to increase the action probability, $p_i(k)$, and decrease all $p_j(k)$, $j \neq i$, if the action $a(k) = a_i$ results in a favourable response. For an unfavourable input, $p_i(k)$ is decreased and all the other components are increased. The same idea applies to state transition probabilities; $f_{ij}(k)$ is increased when $\phi(k) = \phi_i$, $\phi(k+1) = \phi_j$ and the input is favourable, and decreased otherwise. To maintain F as a stochastic matrix, the remaining elements of the ith row must be adjusted updated to sum to unity. Thus the state transition matrices F, $\tilde{F}$ and consequently $\bar{F}$ of the last section become state dependent. Next we adapt a well known linear reinforcement algorithm to the predator-model-mimic system.

5.2. The asymmetric $(\alpha \neq \beta)$ Linear Reward-Penalty (L_{R-P}) scheme

Linear reinforcement algorithms are based on the simple premise of increasing the probability of that action that elicits a favourable response by an amount proportional to the total value of all other action probabilities. Otherwise, it is decreased by an amount proportional to its current value. The probability updating algorithm for the two predator actions, a_1 (ignore) and a_2 (eat), with respective

to penalty probabilities c_1 and c_2, is a Markov chain and has the following form:

$$\left.\begin{aligned} p_1(k+1) &= p_1(k) + \alpha[1 - p_1(k)] \\ p_2(k+1) &= (1-\alpha)\, p_2(k) \end{aligned}\right\} \quad a(k) = a_1, \text{ response is favourable}, \ 0 < \alpha < 1$$

$$\left.\begin{aligned} p_1(k+1) &= (1-\beta)\, p_1(k) \\ p_2(k+1) &= p_2(k) + \beta[1 - p_2(k)] \end{aligned}\right\} \quad a(k) = a_1, \text{ response is unfavourable}, \ 0 < \beta < 1, \ (\beta \neq \alpha)$$

From these equations it follows that if action a_i is chosen at stage k, the probability $p_j(k)$ ($j \neq i$) is decreased, at stage k+1, by an amount proportional to its value at stage k for a favourable response, and increased by an amount proportional to $[1 - p_j(k)]$ for an unfavourable response as this is consistent with action j being more favourable. The parameters α and β are the reward and penalty parameters respectively.

In order to assess the asymptotic behaviour of the action probabilities we consider the conditional expectation of $p_1(k+1)$ given $p_1(k)$:

$$\bar{p}_1(k+1) = E[\, p_1(k+1) \mid p_1(k)] = p_1^2(k)[(\alpha - \beta)(c_1 - c_2)] + p_1(k)[1 + \alpha(c_2 - c_1) - 2\beta c_2] + \beta c_2$$

Then for the Markov chain to be ergodic, we that $\lim_{k \to \infty} E[\, p_1(k+1) \mid p_1(k)] = \bar{p}_1^*(k)$. This probability, $\bar{p}_1^*$, is found as the fixed point of the first-order nonlinear autonomous difference equation (with $c_1 \neq c_2$):

$$p_1 = p_1^2[(\alpha - \beta)(c_1 - c_2)] + p_1[1 + \alpha(c_2 - c_1) - 2\beta c_2] + \beta c_2,$$

The above quadratic equation admits a unique feasible solution for $0 < \bar{p}_1^* < 1.$:

$$\bar{p}_1^* = \frac{2\beta c_2 + \alpha(c_1 - c_2) - \sqrt{\alpha^2(c_1^2 + c_2^2) + 2c_1c_2(2\beta^2 - \alpha^2)}}{2(\alpha - \beta)(c_1 - c_2)}$$

Similarly for action a_2,

$$\bar{p}_2(k) = E[\, p_2(k+1) \mid p_2(k)] = p_2^2(k)[(\alpha - \beta)(c_2 - c_1)] + p_2(k)[1 + \alpha(c_1 - c_2) - 2\beta c_1] + \beta c_1$$

,

the unique fixed point is

$$\bar{p}_2^* = \frac{2\beta c_1 + \alpha(c_2 - c_1) - \sqrt{\alpha^2(c_1^2 + c_2^2) + 2c_1c_2(2\beta^2 - \alpha^2)}}{2(\alpha - \beta)(c_2 - c_1)}$$

The equilibrium probabilities, $\bar{p}_1^*$ and $\bar{p}_2^*$, are asymptotically stable if

$$\left|1 - \sqrt{\alpha^2(c_1^2 + c_2^2) + 2c_1c_2(2\beta^2 - \alpha^2)}\right| < 1$$

Since the expression, $\alpha^2(c_1^2 + c_2^2) + 2c_1c_2(2\beta^2 - \alpha^2)$, is a positive-semi definite quadratic form with an upper bound of 4, asymptotic stability is ensured.

Figures 3 and 4 display a typical evolution pattern of $\bar{p}_1(k)$ and $\bar{p}_2(k)$ towards $\bar{p}_1^*(k)$ and $\bar{p}_2^*(k)$, their asymptotic values respectively.

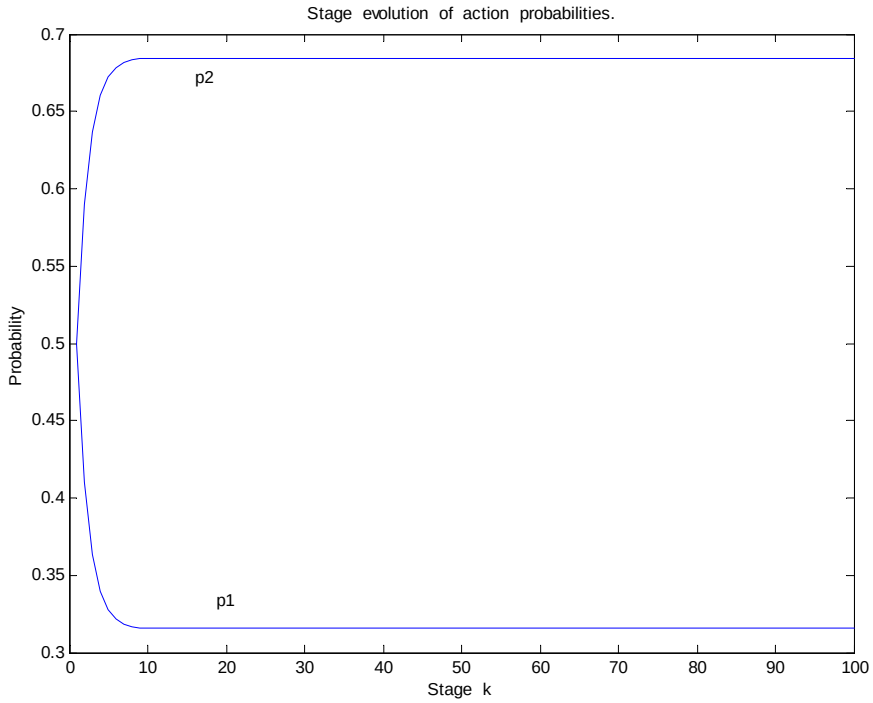


Figure 3. Plot of mean action probabilities versus stage for $\alpha = 0.5,\ \beta = 0.4,\ c_1 = 0.8,\ c_2 = 0.4$. The asymptotic mean probabilities are $\bar{p}_1^* = 0.315,\ \bar{p}_2^* = 0.685$.

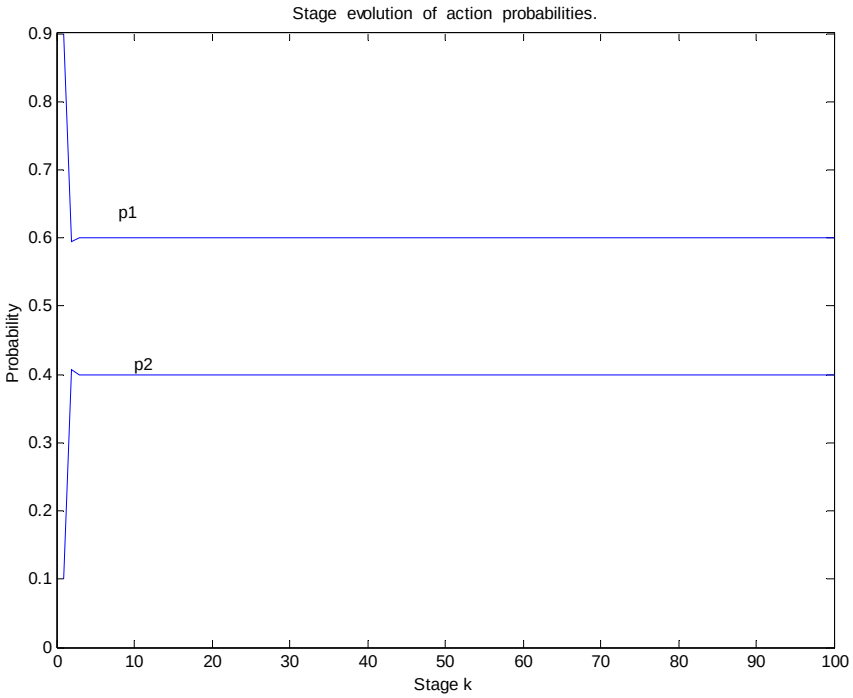


Figure 4. Plot of mean action probabilities versus stage for $\alpha = 0.3,\ \beta = 0.8,\ c_1 = 0.5,\ c_2 = 0.9$. The asymptotic mean probabilities are $\bar{p}_1^* = 0.6,\ \bar{p}_2^* = 0.4$.

The variance of the random variable $p_1(k+1)$, conditioned on $p_1(k)$ is given by

$$S_1(p_1(k)) = E\left[\left(p_1(k+1) - \bar{p}_1(k+1)\right)^2 \mid p_1(k)\right] = E\left[p_1^2(k+1) \mid p_1(k)\right] - \bar{p}_1^2(k) =$$

$$\left[p_1(k) + \alpha(1-p_1(k))\right]^2 p_1(k)(1-c_1) + (1-\beta)^2 p_1^3(k)c_1 + (1-\alpha)^2 p_1^2(k)\left[1-p_1(k)\right](1-c_2) +$$
$$\left[p_1(k) + \beta(1-p_1(k))\right]^2\left[1-p_1(k)\right]c_2 -$$

$$\left[p_1^2(k)\left[(\alpha-\beta)(c_1-c_2)\right] + p_1(k)\left[1+\alpha(c_2-c_1) - 2\beta c_2\right] + \beta c_2\right]^2 =$$

$$\alpha^2 p_1(k)\left[1-p_1(k)\right]\left[1-c_1(1-p_1(k)) - c_2 p_1(k)\right] + \beta^2\left[p_1^3(k)c_1 + (1-p_1(k))^3 c_2\right] -$$
$$\left[p_1^2(k)\left[(\alpha-\beta)(c_1-c_2)\right] + p_1(k)\left[\alpha(c_2-c_1) - 2\beta c_2\right] + \beta c_2\right]^2$$

The asymptotic variance is given by:

$$S_1(\bar{p}_1^*) = \alpha^2 \bar{p}_1^*(1-\bar{p}_1^*)\left[1-c_1(1-\bar{p}_1^*)-c_2\bar{p}_1^*\right]+\beta^2\left[(\bar{p}_1^*)^3 c_1+(1-\bar{p}_1^*)^3 c_2\right]$$

The random variables, $p_1(k)$ and $p_2(k)$, therefore converge in distribution to two random variables with means, $\bar{p}_1^*$ and $\bar{p}_2^*$, and variances, $S_1^* = S_1(\bar{p}_1^*)$ and $S_2^* = S_2(\bar{p}_2^*)$ respectively.

The average penalty at stage k+1 conditioned on the probabilities at stage k is given by

$$\overline{M}(k+1) = \overline{M}(k) - \alpha(c_1-c_2)^2\,\bar{p}_1(k)\bar{p}_2(k) - \beta(c_1-c_2)(c_1\bar{p}_1^2(k) - c_2\bar{p}_2^2(k))$$

To determine whether $\overline{M}(k)$ is monotonically increasing or decreasing we need to examine the sign of $\overline{M}(k+1)-\overline{M}(k)$ for all k. It is easier, however, to assess monotonicity from the sign of the sign of $\frac{d\overline{M}}{dt}$, provided the function $\overline{M}(t)$ can be obtained. An analytic solution to the difference equation in the mean probability is very tedious, but the continuous time solution to the associated differential equation,

$$\frac{d\bar{p}_1}{dt} = \bar{p}_1^2(t)\left[(\alpha-\beta)(c_1-c_2)\right]+\bar{p}_1(t)\left[\alpha(c_2-c_1)-2\beta c_2\right]+\beta c_2,$$

however, is straightforward and given by

$$\bar{p}_1(t) = \frac{p_1^+ e^{-Ct} - D\bar{p}_1^*}{e^{-Ct}-D},$$

where

$$p_1^+ = \frac{2\beta c_2 + \alpha(c_1-c_2) + \sqrt{\alpha^2(c_1^2+c_2^2)+2c_1c_2(2\beta^2-\alpha^2)}}{2(\alpha-\beta)(c_1-c_2)},$$

$$C = \sqrt{\alpha^2(c_1^2+c_2^2)+2c_1c_2(2\beta^2-\alpha^2)},$$

and

$$D = \frac{p_1(0)-p_1^+}{p_1(0)-\bar{p}_1^*}.$$

Similarly,

$$\bar{p}_2(t) = \frac{p_2^+ e^{-Ct} - D\bar{p}_2^*}{e^{-Ct}-D}$$

where

$$p_2^+ = \frac{2\beta c_1 + \alpha(c_2 - c_1) + \sqrt{\alpha^2(c_1^2 + c_2^2) + 2c_1c_2(2\beta^2 - \alpha^2)}}{2(\alpha - \beta)(c_2 - c_1)}$$

It can be seen clearly that, $\lim_{t\to\infty} \bar{p}_1(t) = \bar{p}_1^*$, $\lim_{t\to\infty} \bar{p}_2(t) = \bar{p}_2^*$, as expected.

Since $\overline{M}(t) = c_1\bar{p}_1(t) + c_2\bar{p}_2(t)$, the functional expression for the penalty is

$$\overline{M}(t) = \frac{M^+ e^{-Ct} - D\overline{M}^*}{e^{-Ct} - D}$$

where

$$M^+ = c_1 p_1^+ + c_2 p_2^+ = \frac{\alpha(c_1 + c_2) + C}{2(\alpha - \beta)}$$

and

$$\overline{M}^* = \lim_{k\to\infty} E[\overline{M}(k)] = c_1\bar{p}_1^* + c_2\bar{p}_2^* = \frac{\alpha(c_1 + c_2) - C}{2(\alpha - \beta)}$$

Then

$$\frac{d\overline{M}}{dt} = \frac{DC^2 e^{-Ct}}{(e^{-Ct} - D)^2(\alpha - \beta)}$$

So

$\overline{M}(k)$ is monotonically increasing ($\overline{M}(k)$ is a submartingale) if $\frac{D}{\alpha - \beta} > 0$

$\overline{M}(k)$ is monotonically decreasing ($\overline{M}(k)$ is a supermartingale) if $\frac{D}{\alpha - \beta} < 0$

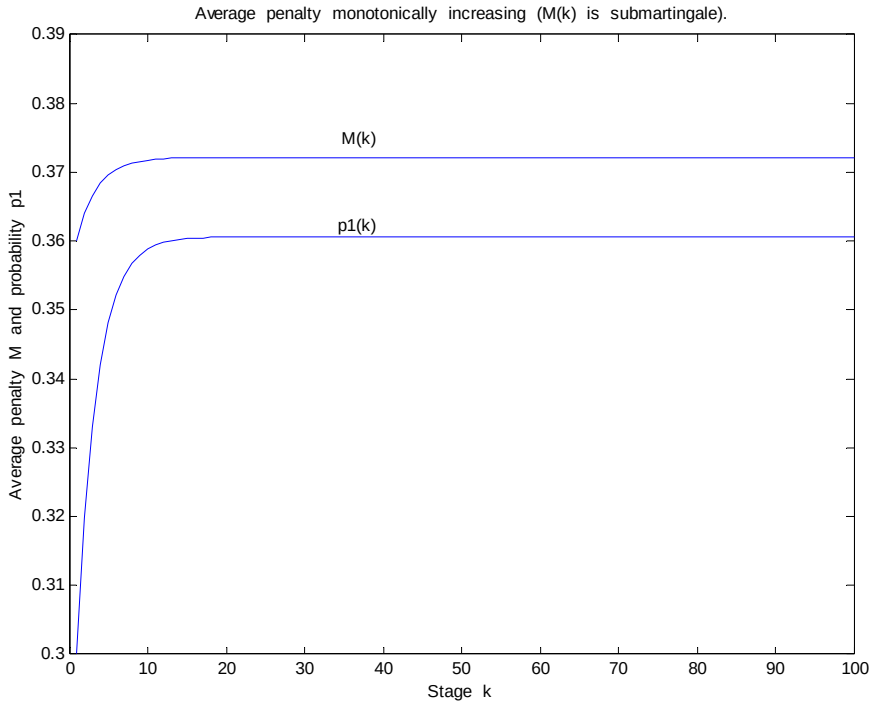


Figure 5. Plot of penalty and action probability evolution for an automaton with $\alpha = 0.5,\ \beta = 0.4,\ c_1 = 0.5,\ c_2 = 0.3,\ p_1(0) = 0.3$. Here we have the asymptotic values $\overline{p}_1^* = 0.36$, $\overline{M}^* = 0.372$, $\dfrac{D}{\alpha - \beta} = 2723.235 > 0$, so $\overline{\boldsymbol{M}}(k)$ is a submartingale.

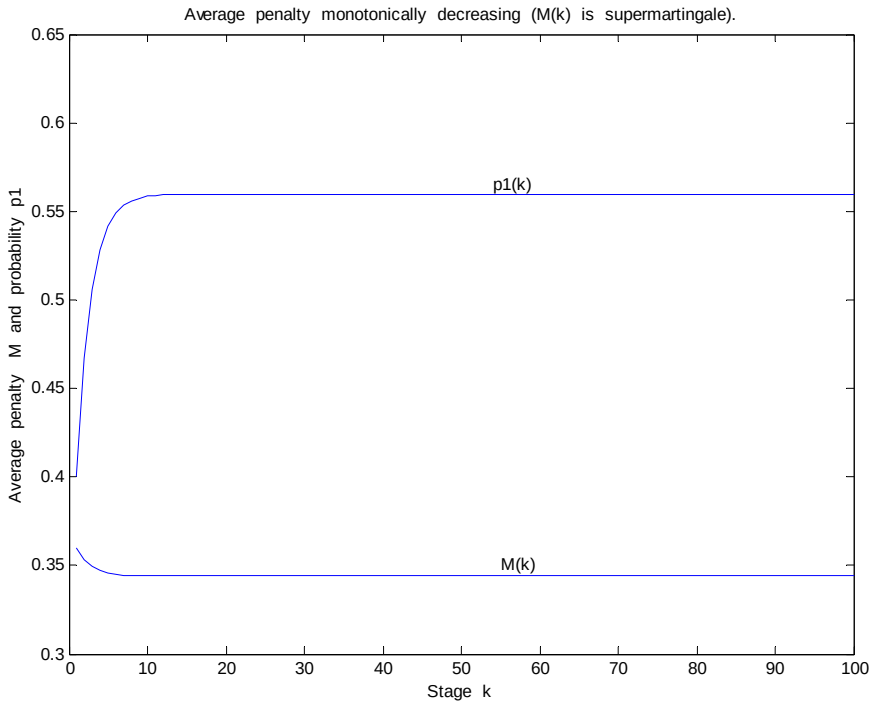


Figure 6. Plot of penalty and action probability evolution for an automaton with $\alpha = 0.4,\ \beta = 0.6,\ c_1 = 0.3,\ c_2 = 0.4,\ p_1(0) = 0.4$. Here we have the asymptotic values $\overline{p}_1^* = 0.56$, $\overline{M}^* = 0.344$, $\dfrac{D}{\alpha - \beta} = -107.20 < 0$, so $\overline{\boldsymbol{M}}\,(k)$ is a supermartingale and the automaton is absolutely expedient.

The performance of the automaton in the long term is better than that of a pure-chance automaton by virtue of the inequality, $\overline{M}^* = \dfrac{\alpha(c_1 + c_2) - C}{2(\alpha - \beta)} < M_0 = \dfrac{c_1 + c_2}{2}$. This is valid for any values of the parameters, α, β, c_1, c_2, hence the automaton is always expedient. It is absolutely expedient when $\dfrac{D}{\alpha - \beta} < 0$, and also ε-optimal because for any arbitrary $\varepsilon > 0$, proper parameter values for α and β can be chosen such that $\overline{M}^* - min\{c_1, c_2\} < \varepsilon$ holds. For instance, for $c_1 = 0.4,\ c_2 = 0.6,\ \varepsilon = 0.01$, α and β must be chosen so that the inequality $\dfrac{0.2\alpha + 0.8\beta - \sqrt{0.04\alpha^2 + 0.96\beta^2}}{2(\alpha - \beta)} < 0.01$ always holds. When $\alpha >> \beta$, the automaton exhibits nearly optimal behaviour, as $\overline{M}^* \approx min\{c_1, c_2\}$ and either $\overline{p}_1^* \approx 1$ if $c_1 < c_2$, or $\overline{p}_1^* \approx 0$ if $c_1 > c_2$. Figure 7 below displays the average penalty curve, with $c_1 = 0.4$ and $c_2 = 0.9$, plotted against all possible α and for $\beta = 0.01$. For $\alpha = 0.9$, $\overline{M}^* = 0.405 \approx c_1$.

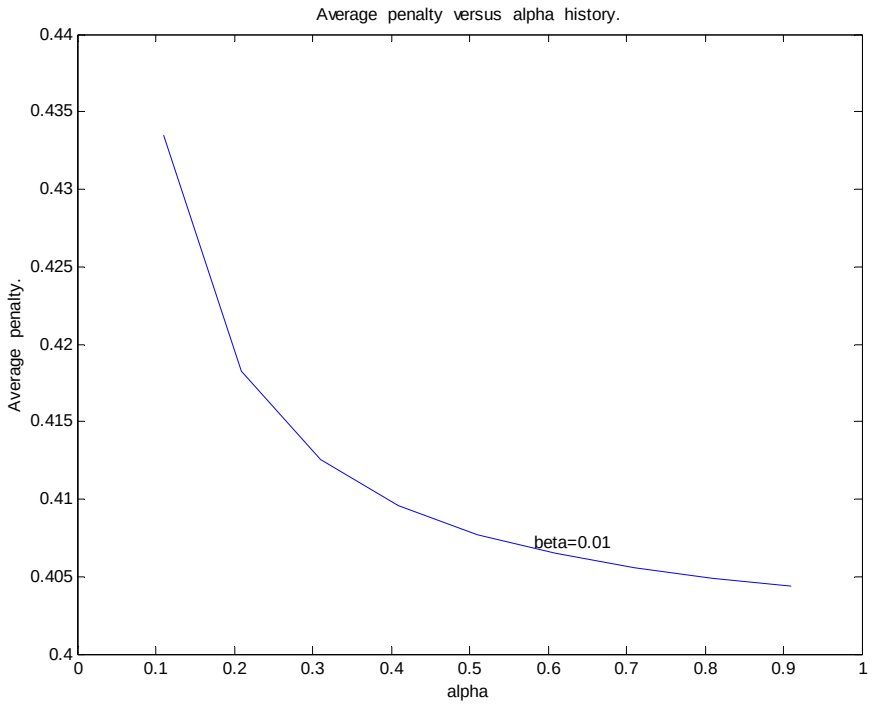


Figure 7. Penalty variation with α with $c_1 = 0.4$, $c_2 = 0.9$ and $\beta = 0.01$.

When $\beta >> \alpha$, $\overline{M}^* \approx \sqrt{c_1 c_2} > min\{c_1, c_2\}$, and hence is not optimal. Figure 8 below displays the average penalty curve, with $c_1 = 0.8$ and $c_2 = 0.2$, plotted against all possible β and for $\alpha = 0.01$. For $\beta = 0.9$, $\overline{M}^* = 0.399 \approx 0.4 = \sqrt{0.16}$.

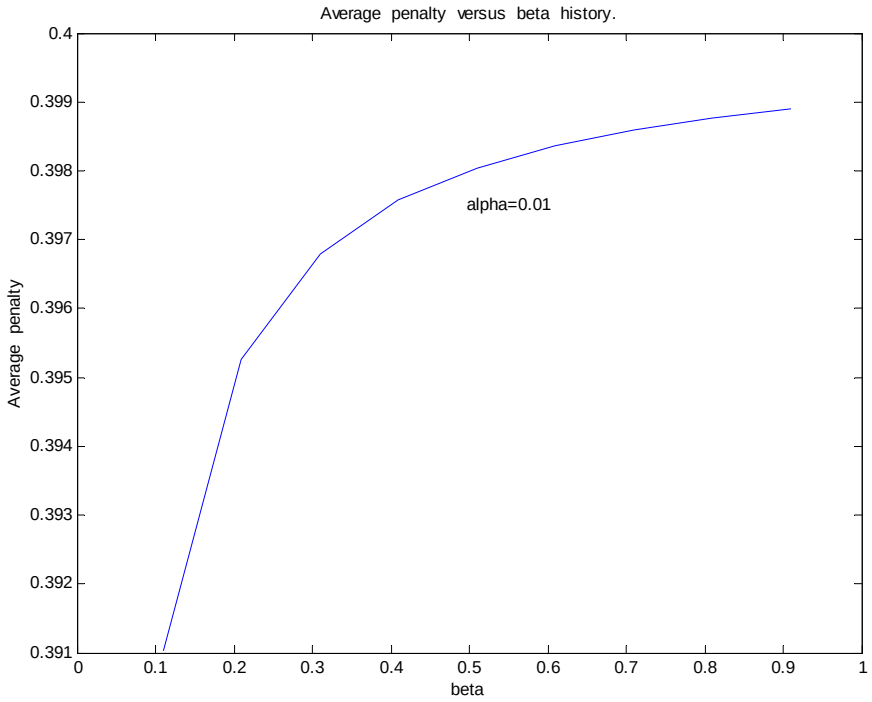


Figure 8. Penalty variation with β with $c_1 = 0.8$, $c_2 = 0.2$ and $\alpha = 0.01$.

In deriving the asymptotic properties of the automaton we have allowed the number of stages to approach infinity. In practice, however, the number of steps needed to converge to the desired action is finite. Let $\overline{M}(\overline{T}_\delta) - \overline{M}^*$ be the difference between the average penalty at time $\overline{T}_\delta$ and its asymptotic value. A natural index of the rate of convergence would be an estimate of the time needed by this difference to become equal to an arbitrary proportion, δ ($0<\delta<1$), of the difference between the final and initial penalties. We have therefore:

$$\delta = \frac{\overline{M}(\overline{T}_\delta) - \overline{M}^*}{M(0) - \overline{M}^*},$$

whence,

$$\overline{T}_\delta = \frac{1}{C} \ln\left[\frac{1}{D}\left(1 - \frac{C}{\delta(\alpha - \beta)(M(0) - \overline{M}^*)}\right)\right]$$

The time, $\overline{T}_\delta$, taken by the automaton to reach $100\delta\%$ of the difference, $M(0) - \overline{M}^*$, is longer when $\alpha < \beta$ and is dependent on the absolute difference, $|c_1 - c_2|$, rather than on the individual penalty probabilities, c_1 and c_2.

Figure 9 below illustrates the faster convergence to $\overline{M}^*$ when $\alpha < \beta$ and $\delta = [0.01\ ,\ 0.1]$.

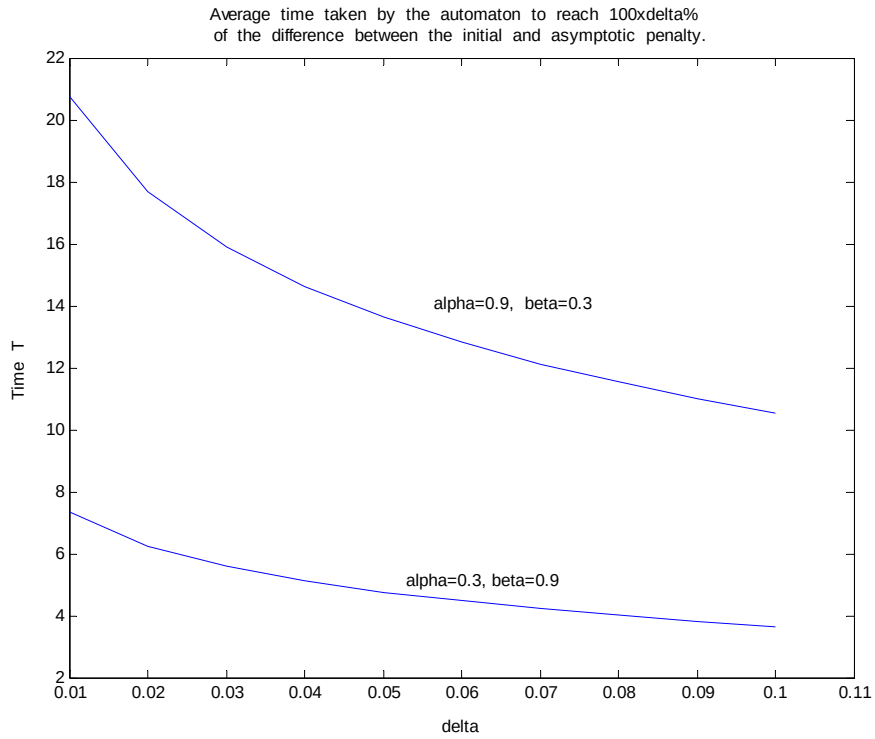


Figure 9. Two cases: (i) $\alpha = 0.9$, $\beta = 0.3$, and (ii) $\alpha = 0.3$, $\beta = 0.9$. Automaton converges to the asymptotic penalty approximately three time faster in case (ii).

In the optimal scenario, $\alpha >> \beta$, the time, $\overline{T}_\delta$, depends on the difference, $|c_1 - c_2|$, the initial probabilities, $p_1(0)$ and $p_2(0)$, α and δ, thus::

$$\overline{T}_\delta \approx \frac{1}{\alpha(c_1 - c_2)} \ln\left(\frac{1-\delta\ p_1(0)}{\delta\ p_2(0)}\right) \quad \text{if} \quad c_1 > c_2$$

$$\overline{T}_\delta \approx \frac{1}{\alpha(c_2 - c_1)} \ln\left(\frac{1-\delta\ p_2(0)}{\delta\ p_1(0)}\right) \quad \text{if} \quad c_1 < c_2$$

$\overline{T}_\delta$ increases as $p_1(0)$ increases when $c_1 > c_2$, and increases as $p_2(0)$ increases when $c_1 < c_2$ given the positive restricted ranges for α and δ.

The **variable predatory efficiency index**, $\overline{e}(k)$, is again defined as the ratio of the probability of consumption of palatable prey to the total probability of encountering palatable prey at each stage *k*:

$$\bar{e}(k) = \frac{\bar{p}_2(k)(1-c_2)}{\bar{p}_1(k)c_1 + \bar{p}_2(k)(1-c_2)}$$

This index is a monotonically increasing function of $\bar{p}_2(k)$ with an asymptotic value, $\bar{e}^*$, obtained by introducing the asymptotic probabilities in the above formula.

5.3. The symmetric ($\alpha = \beta$) Linear Reward–Penalty (L_{R-P}) scheme

The special case when $\alpha = \beta$ is called the symmetric Linear Reward-Penalty scheme (L_{R-P}) and has the following conditional expectation:

$$E[p_1(k+1) \mid p_1(k)] = [1-\alpha(c_1+c_2)]p_1(k) + \alpha c_2$$

which is a linear difference equation in $p_1(k)$, with solution

$$E[p_1(k)] = [1-\alpha(c_1+c_2)]^k p_1(0) + \frac{\left[1-(1-\alpha(c_1+c_2))^k\right]}{\alpha(c_1+c_2)}\alpha c_2$$

whence,

$$\lim_{k\to\infty} E[p_1(k)] = \bar{p}_1^* = \frac{c_2}{c_1+c_2},$$

since $|1-\alpha(c_1+c_2)| < 1$.

If $c_1 > c_2$ then $\bar{p}_1^* < \bar{p}_2^* = \frac{c_1}{c_1+c_2}$, and action a_1 is chosen asymptotically with lower probability, and vice versa. The average penalty evolves according to the difference equation

$$\bar{M}(k+1) = [1-\alpha(c_1+c_2)]\bar{M}(k) + 2\alpha c_1 c_2$$

and is monotonically decreasing (supermartingale) if

$$M(0) = c_1 p_1(0) + c_2 p_2(0) > \bar{M}^* = \frac{2c_1c_2}{c_1+c_2},$$

where $\bar{M}^*$ is the corresponding asymptotic value for $\lim_{k\to\infty} \bar{M}(k)$ otherwise it is a submartingale.

The automaton is always expedient since

$$\overline{M}^{*} = \frac{2c_1c_2}{c_1 + c_2} < \frac{c_1 + c_2}{2} = M_0 \quad (c_1 \neq c_2)$$

and has an asymptotic efficiency index given by

$$\overline{e}^{*} = 1 - c_2$$

Mimicry efficiency is thus tied to the rate the predator consumes the wrong prey. Furthermore, the symmetric $L_{R\text{-}P}$ scheme exhibits long-term properties identical to those of a purely stochastic automaton, switching actions with probability 1 whenever an unfavourable response is recorded (section 4).

5.4. The Linear Reward–Inaction ($L_{R\text{-}I}$) ($\beta = 0$) scheme

The basic idea of the linear reward-inaction scheme ($L_{R\text{-}I}$) is to increase the probability of an action if it was the last action and resulted in a favourable response ($\alpha \neq 0$) or leave the probability unchanged if unfavaourable ($\beta = 0$). The conditional expectation is described by the following nonlinear difference equation:

$$E[p_1(k+1) \mid p_1(k)] = \alpha(c_1 - c_2)p_1^2(k) + [1 + \alpha(c_2 - c_1)]p_1(k)$$

The associated asymptotic probability, $\overline{p}_1^{*}$, and efficiency index, $\overline{e}^{*}$, are found from the corresponding formulae for the asymmetric $L_{R\text{-}P}$ case with $\beta = 0$:

$$\overline{p}_1^{*} = \begin{cases} \dfrac{\alpha(c_1 - c_2) - \alpha(c_1 - c_2)}{2\alpha(c_1 - c_2)} = 0 \quad \text{if} \quad c_1 > c_2, \quad \overline{e}^{*} = 1, \quad m^{*} = 0 \\ \dfrac{\alpha(c_1 - c_2) - \alpha(c_2 - c_1)}{2\alpha(c_1 - c_2)} = 1 \quad \text{if} \quad c_1 < c_2, \quad \overline{e}^{*} = 0, \quad m^{*} = 1 \end{cases}$$

The Markov chain modelling the $L_{R\text{-}I}$ automaton has two absorbing states, action a_1 (ignore prey all the time) if $c_1 < c_2$, and action a_2 (consume every prey encountered) if $c_1 > c_2$. In the former case the predator is completely inefficient, whereas in the latter case it is 100% efficient. The average penalty, $\overline{M}(k)$, is a supermartingale because of the condition $\dfrac{D}{\alpha - \beta} = \dfrac{D}{\alpha} < 0$, with the automaton exhibiting optimum behaviour, since $\overline{M}^{*} = min\{c_1, c_2\}$.

6. Predator as a variable-structure stochastic automaton operating in nonstationary random model-mimic environments

6.1. Introduction

In the last section we analysed, in detail, a linear reinforcement learning algorithm designed to allow a predator (the automaton) to operate efficiently in an environment occupied by palatable and unpalatable prey and characterized by a constant penalty probability for each predator action. Although the assumption of a stationary environment may provide a good approximation when undergoing slow change, the concept of learning is associated with the ability to adapt in a varying environment.

In this section we analyse the performance of the learning algorithm of the last section when each penalty probability, c_i, i = 1,2, is a monotonically non-decreasing function of the respective action probability, a_i, i =1,2. We base our decision on the reasonable assumption that if the predator is ignoring all prey with a certain frequency, palatable prey amongst them are essentially ignored at a lesser rate, and consumed at a relatively greater rate. Thus,

$$c_1(p_1) = r_1 p_1, \quad 0 < r_1 < 1$$
$$c_2(p_2) = r_2 p_2, \; 0 < r_2 < 1$$

The average penalty at stage k is given by $\overline{M}(k) = r_1 p_1^2(k) + r_2 p_2^2(k)$. The pure-chance automaton has an average penalty, $M_0 = \frac{r_1 + r_2}{4}$. Due to the variation in c_1 and c_2, absolute expediency may not always be feasible in the strict sense of $E[M(k+1)] < E[M(k)]$ for all k, but may hold for some $k>k_0$.

6.2. The asymmetric ($\alpha \neq \beta$) Linear Reward-Penalty (L_{R-P}) scheme

The expectation of the action probability, $p_1(k+1)$, conditioned on $p_1(k)$, is a third-order polynomial in $p_1(k)$:

$$E[p_1(k+1) \mid p_1(k)] = (r_1 + r_2)(\alpha - \beta) p_1^3(k) + (3\beta r_2 - \alpha r_2 - r_1 - r_2) p_1^2(k) + (1 + \alpha r_2 - 3\beta r_2) p_1(k) + \beta r_2$$

Then

$$E[p_1(k+1) \mid p_1(k)] - p_1(k) =$$
$$(r_1 + r_2)(\alpha - \beta) p_1^3(k) + (3\beta r_2 - \alpha r_2 - r_1 - r_2) p_1^2(k) + (\alpha r_2 - 3\beta r_2) p_1(k) + \beta r_2$$

The asymptotic behaviour of the automaton will be determined by the nature of the roots of the cubic polynomial. Let

$$f(p_1) = (r_1 + r_2)(\alpha - \beta) p_1^3(k) + (3\beta r_2 - \alpha r_2 - r_1 - r_2) p_1^2(k) + (\alpha r_2 - 3\beta r_2) p_1(k) + \beta r_2$$

There are analytic expressions for the roots of the cubic equation, $f(p_1) = 0$, based on a method attributed to Cardan. A brief outline of the procedure is expounded in the Appendix.

As an example, let $\alpha = 0.7,\ \beta = 0.2,\ r_1 = r_2 = 0.5$. Then $H \approx -0.1142,\ G \approx -0.052,\ E \approx -3.25 \cdot 10^{-3} < 0$. From case (iii) in the Appendix, we obtain the following three real roots:

$$x^* \approx 0.654,\ -0.49,\ -0.166$$

The corresponding roots of the polynomial $f(p_1)$ are given respectively by,

$$p_1^* \approx 2,\ -0.28,\ 0.37$$

The asymptotically stable probability value in this case is $\bar{p}_1^* = 0.37$ as it satisfies the condition for asymptotic stability, namely, $f'(\bar{p}_1^*) \approx -0.52 < 0$.

Figure 10 below is a graph of the polynomial $f(p_1)$ plotted against p_1 for the above parameters. The three roots can be readily identified here. Figure 11 demonstrates the evolution of the conditional action probability $E[p_1(k+1) \mid p_1(k)]$ towards the asymptotic value of 0.37.

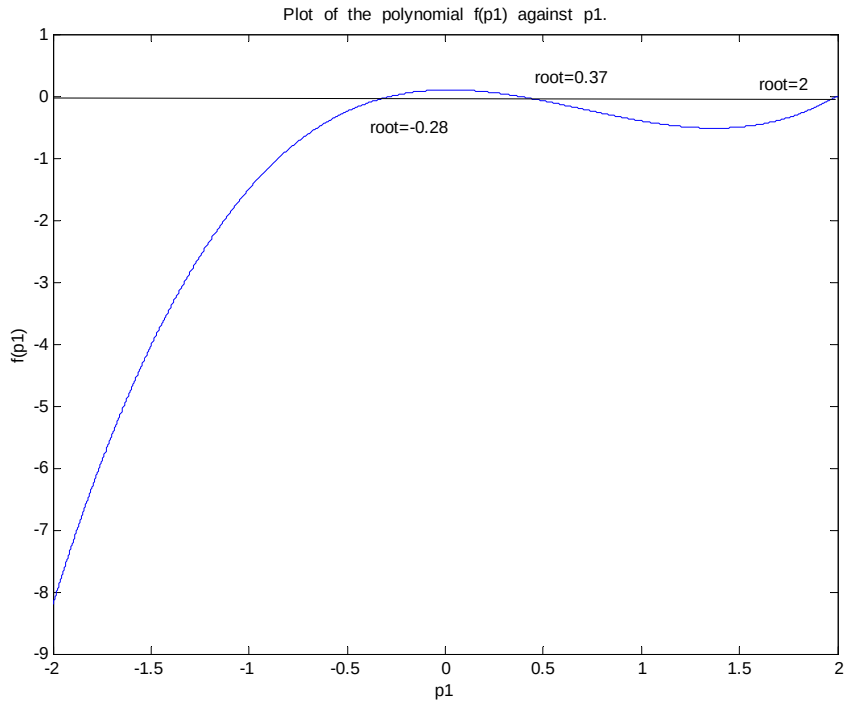


Figure 10. Plot of $f(p_1) = 0.5\,\boldsymbol{p_1^3} - 1.05\,\boldsymbol{p_1^2} + 0.05\,p_1 + 0.1$ versus p_1.

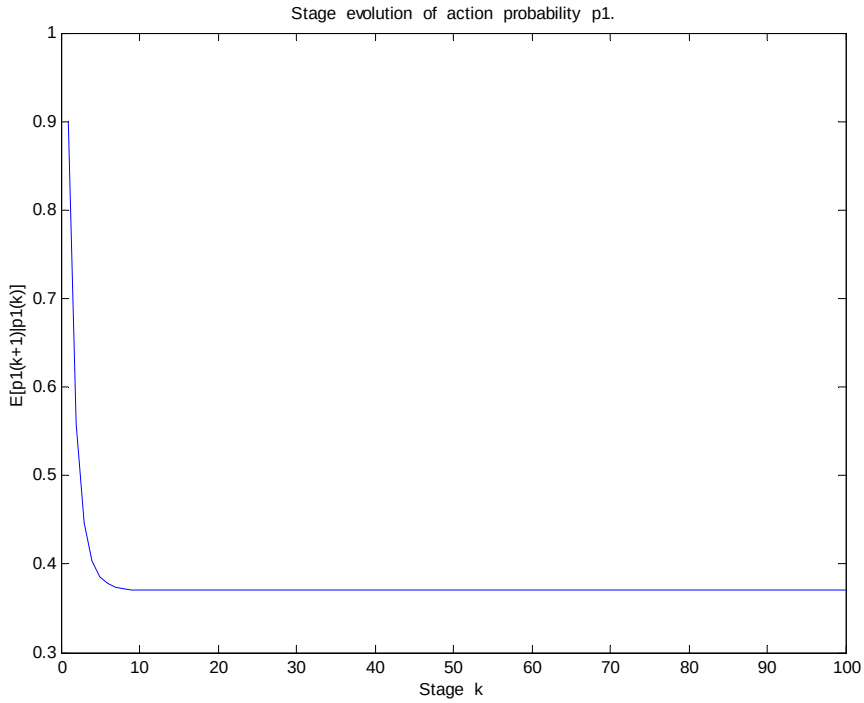


Figure 11. Evolution of $E[p_1(k+1)|p_1(k)] = 0.5\,\boldsymbol{p_1^3}(k) - 1.05\,\boldsymbol{p_1^2}(k) + 1.05\,p_1(k) + 0.1$ towards $\overline{\boldsymbol{p}}_1^* = \mathbf{0.37}$.

Figure 12 displays the plot of $f(p_1)$ when there is one real root and two complex roots. In this case, $\alpha = 0.1,\ \beta = 0.9,\ r_1 = 0.1,\ r_2 = 0.9$, and $H = 0.4245,\ G = -0.1395,\ E = 0.3254 > 0$. The unique real root of the polynomial is $\bar{p}_1^* = 0.4225$. Figure 13 displays the stage history of $E[p_1(k+1)|p_1(k)]$ towards $\bar{p}_1^* = 0.4225$. Note the oscillatory movement of $E[p_1(k+1)|p_1(k)]$ due to the complex roots of $f(p_1)$. The automaton is asymptotically stable as $f'(\bar{p}_1^*) \approx -1.637 < 0$.

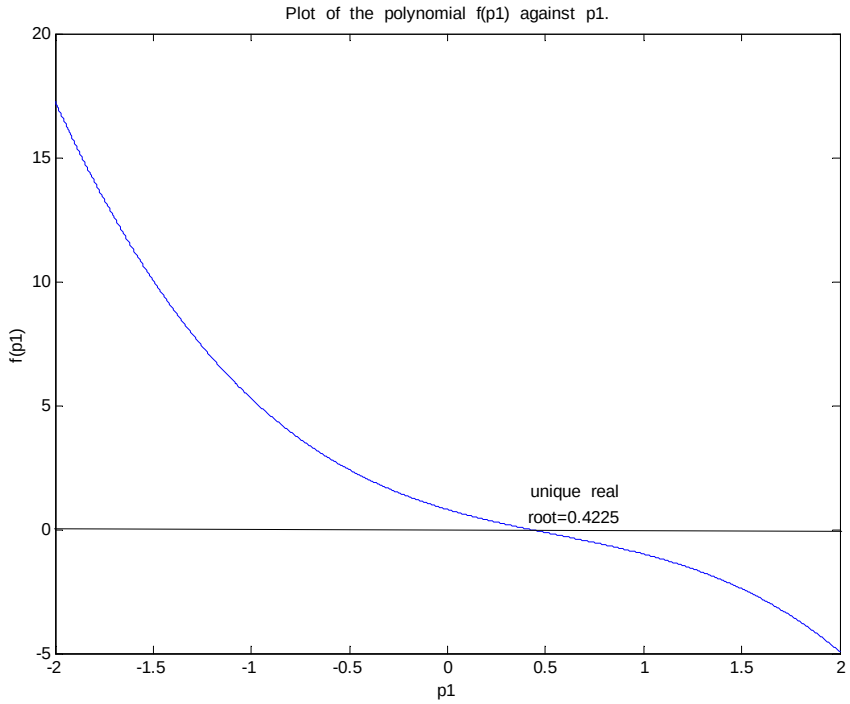


Figure 12. Plot of $f(p_1) = -0.8\,\boldsymbol{p}_1^3 + 1.34\,\boldsymbol{p}_1^2 - 2.34\,p_1 + 0.81$ versus p_1.

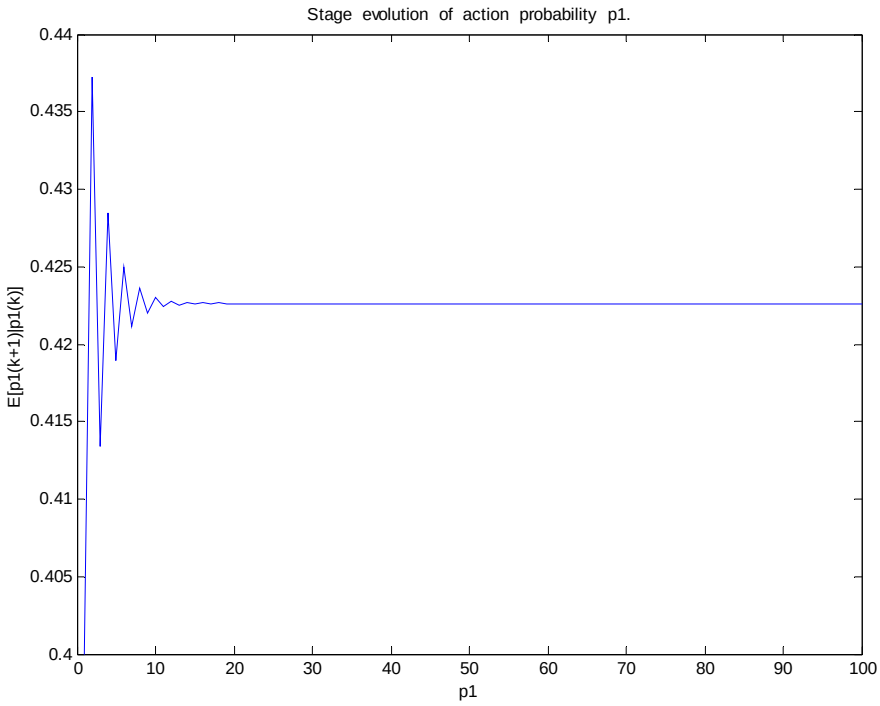


Figure 13. Evolution of $E[p_1(k+1)|p_1(k)] = -0.8\,\boldsymbol{p}_1^3(k) + 1.34\,\boldsymbol{p}_1^2(k) - 1.34\,p_1(k) + 0.81$ towards $\bar{\boldsymbol{p}}_1^* = \mathbf{0.4225}$.

The average penalty function, $\overline{M}(k)$, is either a supermartingale or submartingale except when $f(p_1)$ has complex roots. To demonstrate this, when $\alpha = 0.9,\ \beta = 0.2,\ r_1 = 0.1,\ r_2 = 0.9$, the penalty is a supermartingale and the automaton is expedient $\left(\overline{M}^* \approx 0.142 < \frac{r_1 + r_2}{4} = 0.25\right)$, as well as absolutely expedient (figure 14) whereas, for example, when $\alpha = 0.3,\ \beta = 0.9,\ r_1 = r_2 = 0.5$, the average penalty oscillates before it settles to the stable value, $\overline{M}^* \approx 0.267$ (figure 15) and is not even expedient $(0.267 > 0.25)$.

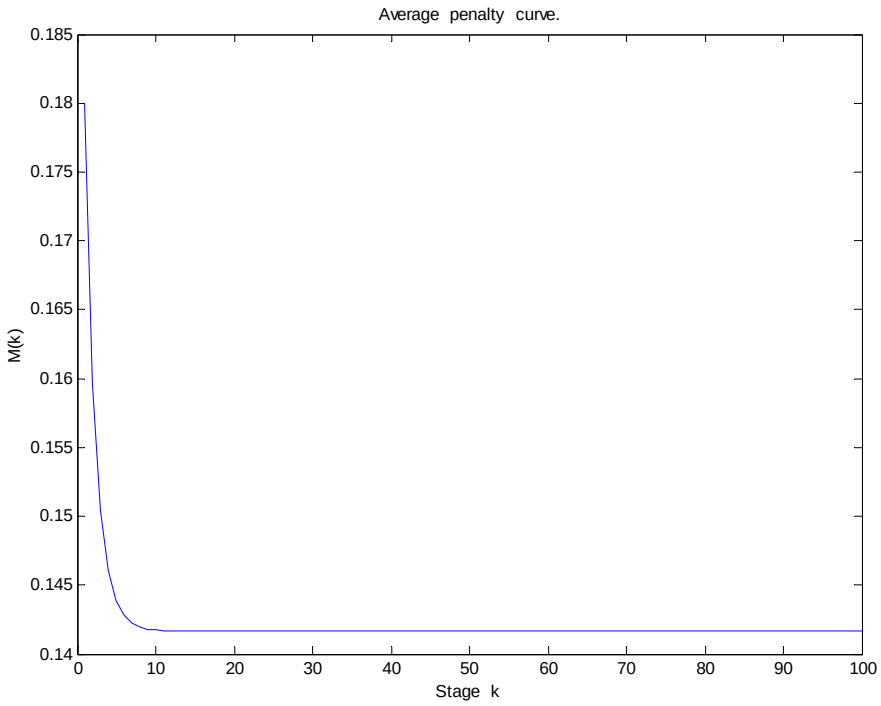


Figure 14. Average penalty is a supermartingale.

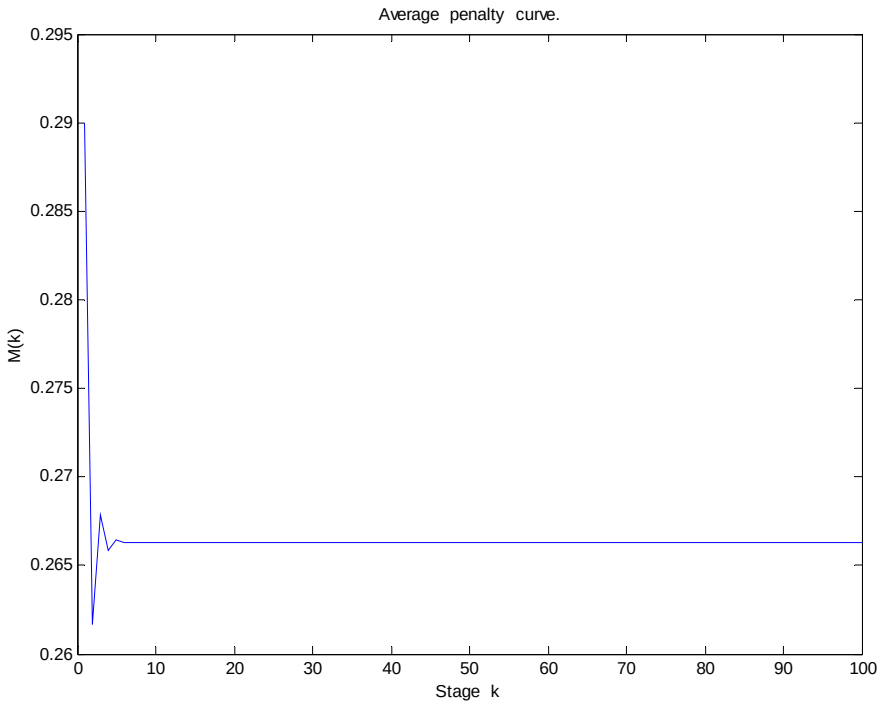


Figure 15. Average penalty oscillates before it settles to its stable value.

The penalty probabilities tend to equalize at equilibrium, that is $\bar{p}_1^* \approx \frac{r_2}{r_1 + r_2}$, $\bar{p}_2^* \approx \frac{r_1}{r_1 + r_2}$, if the following condition connecting all parameters is held:

$$\frac{1-\alpha}{\beta} = \frac{r_2(r_1 - r_2)}{r_1(r_1 + r_2)}$$

In this case the asymptotic average penalty, $\overline{M}^* = \frac{r_1 r_2}{r_1 + r_2} < M_0$ (expedient). For large α, large β and $r_1 \approx r_2$, and the automaton consequently behaves like a pure-chance automaton in the long term. Moreover, the automaton is never optimal ($\bar{p}_1^* = 0$ or 1) since always $f(0) \neq 0$ and $f(1) \neq 0$.

6.3. The symmetric ($\alpha = \beta$) Linear Reward–Penalty (L_{R-P}) scheme

When $\alpha = \beta$ and $\frac{r_1}{r_2} \neq 2\alpha - 1$, $f(p_1)$ is now a quadratic polynomial, with a unique root in the range [0,1] given by

$$\bar{p}_1^* = \frac{\alpha r_2 - \sqrt{\alpha r_2^2 (1-\alpha) + \alpha r_1 r_2}}{2\alpha r_2 - r_1 - r_2}$$

Since $f'(\bar{p}_1^*) = -\sqrt{\alpha r_2^2 (1-\alpha) + \alpha r_1 r_2} < 0$, the root is asymptotically stable. The continuous time function for $\bar{p}_1$ is,

$$\bar{p}_1(t) = \frac{p_1^+ e^{-Ct} - D\bar{p}_1^*}{e^{-Ct} - D},$$

where

$$p_1^+ = \frac{\alpha r_2 + \sqrt{\alpha r_2^2 (1-\alpha) + \alpha r_1 r_2}}{2\alpha r_2 - r_1 - r_2},$$

$$C = 2\sqrt{\alpha r_2^2 (1-\alpha) + \alpha r_1 r_2},$$

and

$$D = \frac{p_1(0) - p_1^+}{p_1(0) - \bar{p}_1^*}.$$

The continuous time average penalty is given by

$$\overline{M}(t) = r_1 \overline{p}_1^2(t) + r_2 \overline{p}_2^2(t) = (r_1 + r_2)\overline{p}_1^2(t) - 2r_2 \overline{p}_1(t) + r_2$$

and its time derivative by

$$\frac{d\overline{M}}{dt} = \frac{2C^2 e^{-Ct}}{(e^{-Ct} - D)^2} \left(\frac{D}{2\alpha r_2 - r_1 - r_2} \right) \left(\overline{p}_1(t)(r_1 + r_2) - r_2 \right)$$

The above derivative vanishes (i) asymptotically as $t \to \infty$, and (ii) when the two penalty probabilities become equal at time

$$t = \frac{1}{C} \ln \left[\frac{\left(\frac{r_2}{r_1 + r_2} - p_1^+ \right)}{D\left(\frac{r_2}{r_1 + r_2} - \overline{p}_1^* \right)} \right],$$

if such time exists.

For example, for $\alpha = 0.5,\ r_1 = 0.1,\ r_2 = 0.3,\ p_1(0) = 0.9$, the penalty curve possesses a turning point, $\overline{M}(t) = \frac{0.03}{0.4} = 0.075$ when $\overline{p}_1(t) = 0.75$ (figure 16).

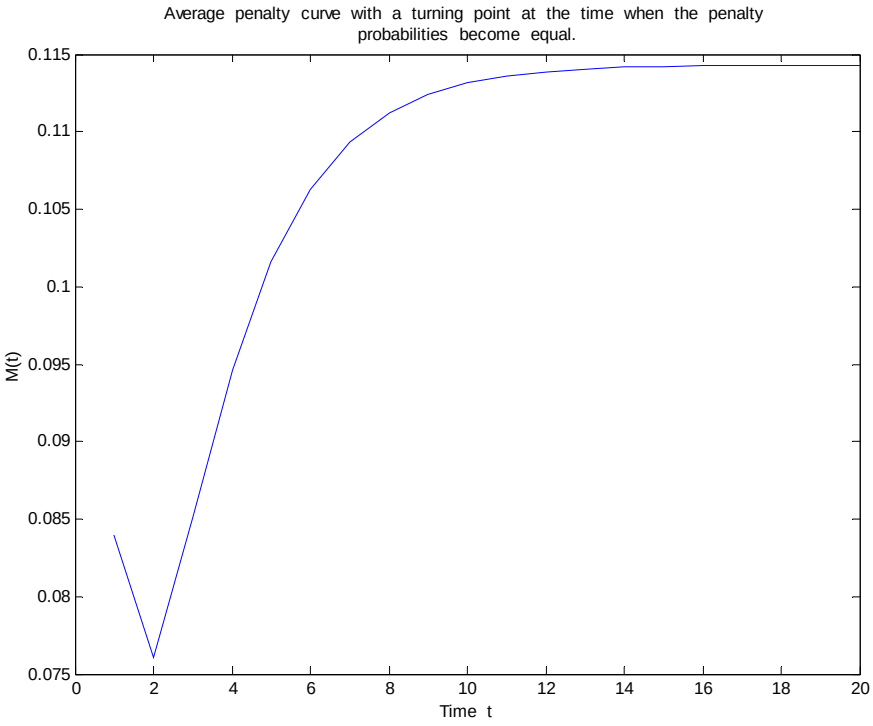


Figure 16. Average penalty curve has a turning point.

If the value $\frac{r_2}{r_1+r_2}$ is never attained by the first action probability prior to its assuming a stable value, then

$\overline{M}(t)$ is a submartingale if either $\frac{D}{2\alpha r_2 - r_1 - r_2} > 0$ ($\overline{p}_1$ increasing) and $p_1(0) > \frac{r_2}{r_1+r_2}$, or $\frac{D}{2\alpha r_2 - r_1 - r_2} < 0$ ($\overline{p}_1$ decreasing) and $p_1(0) < \frac{r_2}{r_1+r_2}$.

$\overline{M}(t)$ is a supermartingale if either $\frac{D}{2\alpha r_2 - r_1 - r_2} > 0$ ($\overline{p}_1$ increasing) and $\overline{p}_1^* < \frac{r_2}{r_1+r_2}$, or $\frac{D}{2\alpha r_2 - r_1 - r_2} < 0$ ($\overline{p}_1$ decreasing) and $\overline{p}_1^* > \frac{r_2}{r_1+r_2}$.

Next we investigate the conditions under which the penalty probabilities become equal asymptotically. Setting $p_1 = \frac{r_2}{r_1+r_2}$ in the polynomial $f(p_1)$ we obtain a cubic polynomial in r_1 and r_2:

$$f(r_1, r_2) = 2r_2^3(\alpha - 0.5)(\alpha - 1) + r_2^2 r_1(2 - 3\alpha) + r_2 r_1^2(2\alpha^2 - \alpha + 1) - \alpha r_1^3$$

(i) If $\alpha \to 1$ and $r_1 = r_2$, then $\overline{p}_1^* = \dfrac{r_2}{r_1 + r_2}$.

(ii) If $\alpha \to 0$, then $\overline{p}_1^*$ never attains the value $\dfrac{r_2}{r_1 + r_2}$.

(iii) If $\alpha = 0.5$ and $r_1 = (1+\sqrt{2})r_2$ then $\overline{p}_1^* = \dfrac{r_2}{r_1 + r_2}$.

(iv) If $\alpha > 0.5$, the polynomial $f(r_1,r_2)$ has either 0 or 2 positive roots, r_2, that can be determined according to the Cardan method outlined earlier.

(v) If $\alpha < 0.5$, the polynomial $f(r_1,r_2)$ has only 1 positive root, r_2, that can also be found easily.

Finally, when $\dfrac{r_1}{r_2} = 2\alpha - 1$ and $\alpha > \dfrac{1}{2}$, $f(p_1)$ is a linear equation with root, $\overline{p}_1^* = 0.5$, and the automaton is ultimately a pure-chance automaton.

6.4. The Linear Reward–Inaction (L_{R-I}) ($\beta = 0$) scheme

When $\beta = 0$, $f(p_1)$ is again a quadratic polynomial, with a unique root in the range [0,1] given by:

$$\overline{p}_1^* = \frac{\alpha r_2 + r_1 + r_2 - \sqrt{(\alpha r_2 + r_1 + r_2)^2 - 4\alpha^2 r_2 (r_1 + r_2)}}{2\alpha (r_1 + r_2)}$$

Since $f'(\overline{p}_1^*) = -\sqrt{(\alpha r_2 + r_1 + r_2)^2 - 4\alpha^2 r_2 (r_1 + r_2)} < 0$, the root is asymptotically stable. The continuous time function for $\overline{p}_1$ is,

$$\overline{p}_1(t) = \frac{p_1^+ e^{-Ct} - D\overline{p}_1^*}{e^{-Ct} - D},$$

where

$$p_1^+ = \frac{\alpha r_2 + r_1 + r_2 + \sqrt{(\alpha r_2 + r_1 + r_2)^2 - 4\alpha^2 r_2 (r_1 + r_2)}}{2\alpha (r_1 + r_2)},$$

$$C = \sqrt{(\alpha r_2 + r_1 + r_2)^2 - 4\alpha^2 r_2 (r_1 + r_2)}\,,$$

and

$$D = \frac{p_1(0) - p_1^+}{p_1(0) - \overline{p}_1^*}\,.$$

The time derivative of the penalty is:

$$\frac{d\overline{M}}{dt} = \frac{2C^2 e^{-Ct}}{(e^{-Ct} - D)^2 \alpha (r_1 + r_2)} D\left(\overline{p}_1(t)(r_1 + r_2) - r_2\right)$$

The analysis of the behaviour of $\overline{M}(t)$ is analogous to that for the $L_{R\text{-}P}$ scheme. As an example, let $\alpha = 0.8,\ r_1 = 0.3,\ r_2 = 0.4,\ p_1(0) = 0.9$. The penalty curve possesses a turning point, $\overline{M}(t) = \frac{0.12}{0.7} \approx 0.17$ when $\overline{p}_1(t) \approx 0.57$ (figure 17).

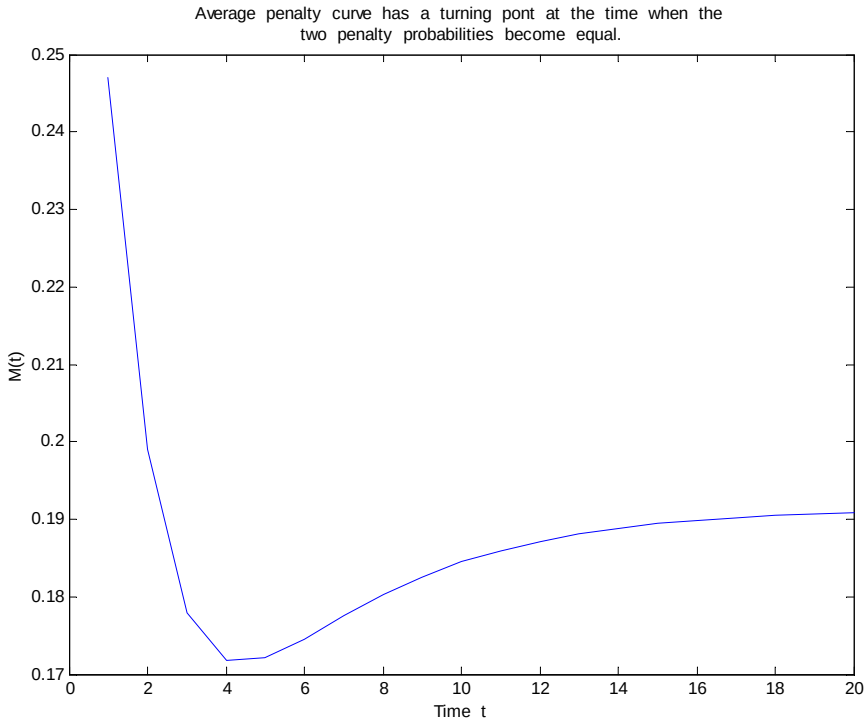


Figure 17. Average penalty curve has a turning point.

If the value $\frac{r_2}{r_1 + r_2}$ is never attained by the first action probability prior to its assuming a stable value, then

$\overline{M}(t)$ is a submartingale if either $D > 0$ ($\overline{p}_1$ increasing) and $p_1(0) > \frac{r_2}{r_1 + r_2}$, or

$$D < 0 \ (\overline{p}_1 \text{ decreasing}) \text{ and } p_1(0) < \frac{r_2}{r_1 + r_2}.$$

$\overline{M}(t)$ is a supermartingale if either $D > 0$ ($\overline{p}_1$ increasing) and $\overline{p}_1^* < \frac{r_2}{r_1 + r_2}$, or

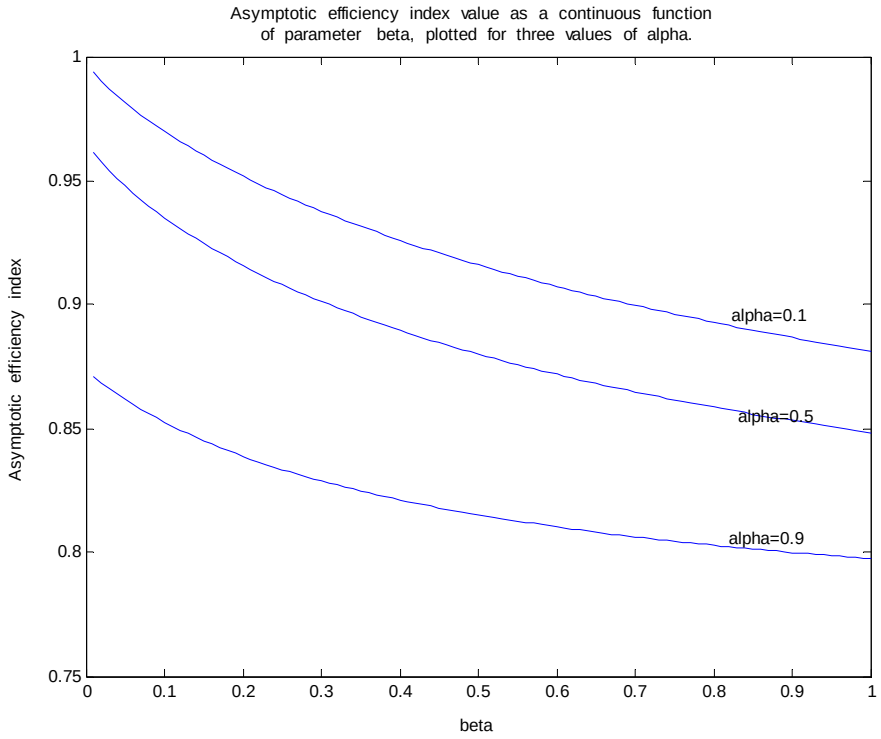
$$D < 0 \ (\overline{p}_1 \text{ decreasing}) \text{ and } \overline{p}_1^* > \frac{r_2}{r_1 + r_2}.$$

Next we investigate the conditions under which the penalty probabilities become equal asymptotically. Setting $p_1 = \frac{r_2}{r_1 + r_2}$ in the polynomial $f(p_1)$, we obtain a linear expression in r_1 and r_2:

$$f(r_1, r_2) = (\alpha - 1)(r_1 + r_2) \neq 0$$

So under the $L_{R\text{-}I}$ scheme, the penalty probabilities never become equal.

The predatory efficiency index, $\overline{e}(k)$, for the asymmetric $L_{R\text{-}P}$ with variable penalty structure scheme assumes a more complicated form than that of the asymmetric $L_{R\text{-}P}$ with fixed penalties and is not given explicitly here. Instead, we provide graphical output of simulated variation of the asymptotic index value, $\overline{e}^*$, with the learning parameters α and β in the following two figures. Figure 18 depicts $\overline{e}^*$ as a continuous convex function of β for three values of α, $\alpha = 0.9,\ 0.5,\ 0.1$. Figure 19 depicts $\overline{e}^*$ as a continuous concave function of α for three values of β, $\beta = 0.9,\ 0.5,\ 0.1$. Figure 20 displays the linearity of the index function for the symmetric version of the $L_{R\text{-}P}$ $(\alpha = \beta)$. In all cases, $r_1 = 0.6,\ r_2 = 0.4,\ p_1(0) = 0.5$. Evidently, the efficiency of predation is at its highest (and mimicry efficiency at its lowest) when both α and β assume low values simultaneously. Using an analogy from utility theory, and considering the predatory efficiency index to be the utility function, we could infer, for the given values of r_1 and r_2, that fixing parameter α and adjusting parameter β is equivalent to risk seeking predatory behaviour, fixing parameter β and adjusting parameter α amounts to risk averse behaviour, whereas fixing both parameters is comparable to risk neutrality.



α.

Figure 18. Asymptotic efficiency index as a convex function of β for fixed

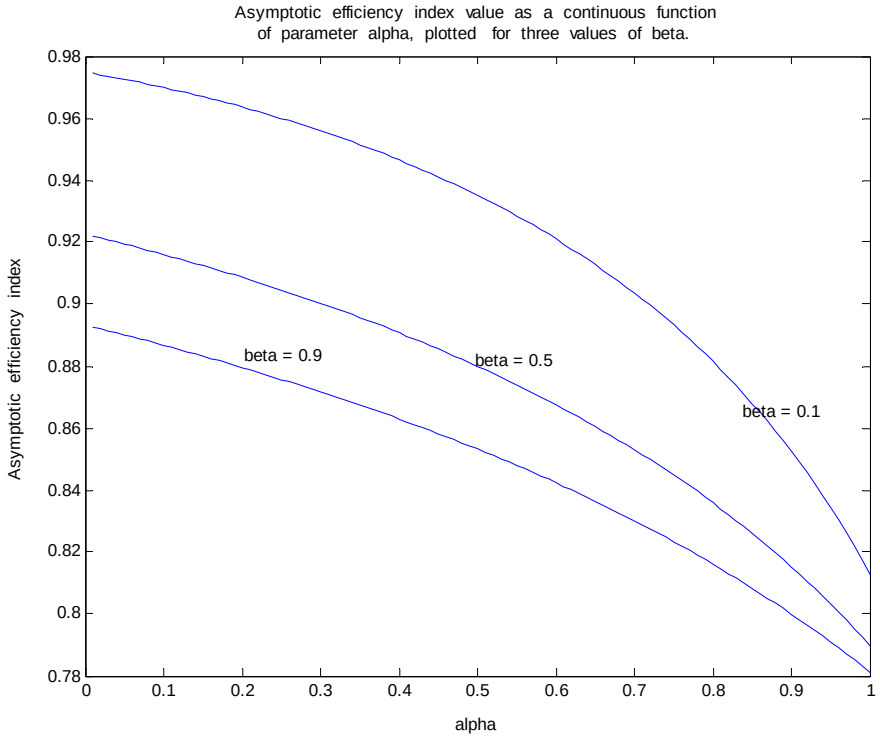


Figure 19. Asymptotic efficiency index as a concave function of α for fixed β.

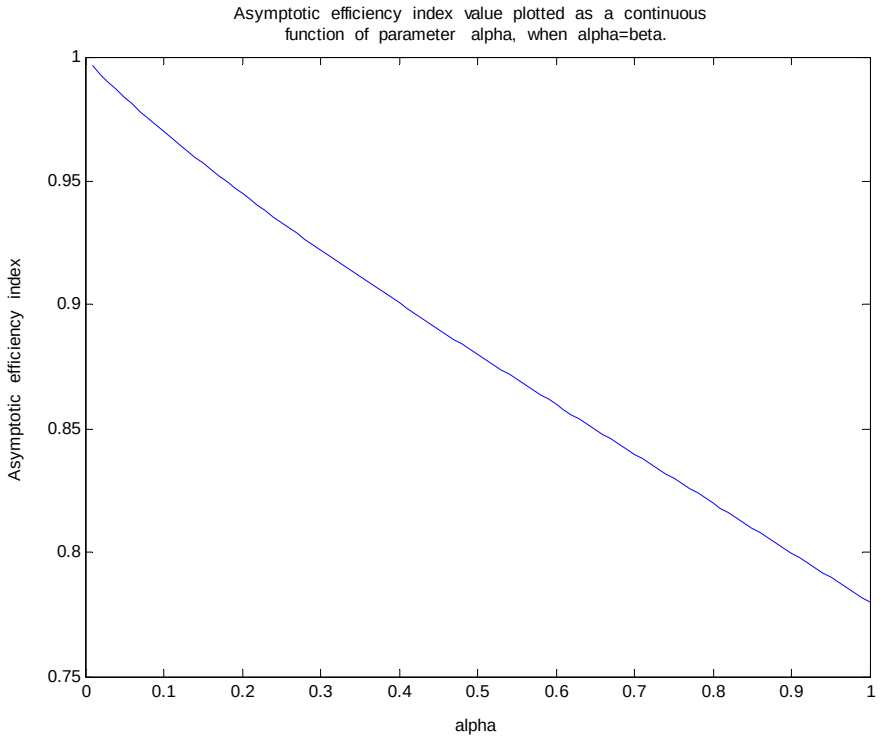


Figure 20. Asymptotic efficiency index as an approximately linear function of α when $\alpha = \beta$.

7. Discussion

In this paper we have explored the concept of a predator as a learning automaton feeding on prey that can be broadly categorized as either palatable (the mimics) or unpalatable (the models). The predator's actions is to either attack the prey or simply ignore it. Each action elicits a probabilistic response from the environment that is classified as favourable or unfavourable. A response is deemed favourable if either the prey consumed is of the palatable type or if the prey ignored is unpalatable and deemed unfavourable if either the prey ignored is palatable or the prey consumed is unpalatable. This distinction made when ignoring prey is related to the predator's ability to discriminate effectively against models. If the predator senses that the prey ignored is of palatable nature it will decrease the frequency of avoidance and vice versa. Furthermore, the distinction aids in quantifying the expected frequency of missed palatable food as part of the average overall penalty.

In section 4 we analysed various predator strategies when no learning takes place and derived a condition that dictates the best (that is, minimum average penalty) policy. In sections 5 and 6 we identified the predator with a learning automaton capable of learning by means of a reinforcement algorithm, the Linear Reward-Penalty scheme, $L_{R\text{-}P}$. In section 5 the environmental penalty probabilities were held constant, whereas in section 6 they were allowed to vary in proportion to the respective action probabilities.

As a measure of predator performance we have consistently used the stage dependent average linear penalty,

$$\overline{M}(k) = \bar{p}_1(k)c_1(k) + \bar{p}_2(k)c_2(k)$$

where $\bar{p}_1$ and $\bar{p}_2$ are the average probabilities of avoiding and consuming respectively at the k^{th} encounter, and c_1, c_2 are the respective penalties incurred at the encounter. A monotonically decreasing penalty indicates a steady improvement in the predator's overall performance. We have obtained explicit parameter conditions that will render $\overline{M}(k)$ a supermartingale and thus ensure such performance. In contrast Estabrook and Jespersen, (1974) and subsequent researchers who improved on their work, used the expected long-term benefit per encounter to the predator as a performance index. This however pertained solely to consumption. Under our approach such benefit is given by

$$S^* = \bar{p}_2^* d_2 - b\bar{p}_2^* c_2 = \bar{p}_2^*(1 - c_2 - bc_2)$$

where b is a parameter quantifying the unpalatability of the model, and $\bar{p}_2^*$ is the long-term mean probability of prey consumption. Some benefit will accrue if

$b < \frac{1-c_2}{c_2}$. The maximum long-term benefit can be found by introducing the expression for $\bar{p}_2^*$ in S^*. An expression containing the two learning parameters, α and β, is then obtained and can be differentiated with respect to these to yield the stationary values α^* and β^*. Perhaps a more realistic approach than maximizing the long-term benefit would be the maximization of the benefit function at each stage k by the predator. The problem then becomes a typical multi-stage decision problem with two decision variables $\alpha(k)$ and $\beta(k)$, and can be formulated as an Optimal Control programme:

$$\max_{\alpha(k),\beta(k)} \sum_{k=1}^{N} S(k)$$

subject to the constraints

$$\mathbf{p}(k+1) = \mathbf{f}(\mathbf{p}(k), \alpha(k), \beta(k)), \quad \text{learning algorithm}$$

$$\left.\begin{aligned} 0 \le \alpha(k) \le 1 \\ 0 \le \beta(k) \le 1 \end{aligned}\right\} \text{control (learning parameter) constraints}$$

Solution of the Optimal Control problem will produce the desired optimal sequence, $\alpha^*(k)$, $\beta^*(k)$, $k = 1, \cdots, N$.

Alternatively, if the benefit to the predator is to include ignored palatable prey, the performance criterion may be chosen as the minimization of the cumulative average penalty, $\sum_{k=1}^{N} \overline{M}(k)$, or the direct minimization the long-term penalty, $\overline{M}^*$.

The long-term efficiency of the predator, $\bar{e}^*$, is the fraction of encounters with palatable prey that are actually consumed. The percentage the predator falls short of being 100% efficient constitutes the efficiency of the mimics, $\overline{m}^*$, which is an important factor in their effort to survive. The change in $\bar{e}^*$ depends on the magnitude of the learning parameters α and β, with the environmental parameters, c_1 and c_2, affecting the rate of change.

In this work we have assumed, for the sake of model simplicity, that the environmental penalty probabilities are either constant or proportional to the respective action probabilities, and have confined our attention to two types of prey only. We have also chosen a linear learning algorithm to model the predator's behaviour. Despite the simplicity of the model some rich behaviour is seen to evolve. In general, the environmental penalties are likely to depend on prey density, spatial and temporal prey distribution, varying degrees of prey unpalatability and appearance, and seasonal variations in the predator's behavioural patterns. For example, to reflect seasonal characteristics in any model

an interesting and valid choice of penalty functions might be $c_i(t) = C_i \sin(\varpi_i t + \theta_i),\ i = 1,2$.

We have endeavoured here to construct a simple theoretical framework for predator learning from which more comprehensive models can originate in the future. Good data about predator psychology is difficult to obtain and consequently how the predators actually learn is still unclear (Speed 1999). The mathematical modelling of the predator-model-mimic complex is still at a very speculative stage. We feel that the learning automaton methodology can be a useful tool in making theoretical predictions that can be tested when comprehensive data become available.

APPENDIX

The value of the polynomial,

$$f(p_1) = (r_1 + r_2)(\alpha - \beta)p_1^3(k) + (3\beta r_2 - \alpha r_2 - r_1 - r_2)p_1^2(k) + (\alpha r_2 - 3\beta r_2)p_1(k) + \beta r_2$$

at each endpoint of the range of interest, [0,1], is

$$f(0) = \beta r_2 > 0$$
$$f(1) = -(r_1 + r_2)(1 - \alpha) - \beta r_1 < 0$$

From the Intermediate Value Theorem there must be at least one root, $\bar{p}_1^*$, of the polynomial in the range [0,1]. We are going to prove that $\bar{p}_1^*$ is unique. From the Descartes' rule of signs we know that the number, n_p, of positive roots of $f(p_1)$ is at most equal to the number of variations, s_p, in sign of the coefficients of $f(p_1)$. Moreover, the difference, $s_p - n_p$, is a nonnegative even integer. There are three cases to consider:

(i) If $\alpha \geq 3\beta$ then $s_p = 2$. Since $s_p - n_p = 2 - n_p$ must be a nonnegative even integer, then either $n_p = 0$ or $n_p = 2$. As we have already established that there is at least one root in the range [0,1], $n_p \neq 0$. If $n_p = 2$ then the graph of $f(p_1)$ must cross the axis of p_1 twice in the range [0,1] and consequently $f(0)$ and $f(1)$ will have the same sign. Since they are opposite in sign we conclude

that there are at most two positive roots but only one, namely $\bar{p}_1^*$, in the range [0,1].

(ii) If $\beta < \alpha < 3\beta$ then again $s_p = 2$, and the reasoning of (i) applies here too.

(iii) If $\alpha < \beta$ then either $s_p = 1$ or $s_p = 3$. Since $s_p - n_p$ must be a nonnegative even integer, then either $n_p = 1$ or $n_p = 3$. If $n_p = 1$ then the only positive root is $\bar{p}_1^*$ in the range [0,1]. If $n_p = 3$ then the derivative, $f'(p_1)$, of $f(p_1)$ with respect to p_1 must vanish at least twice in [0,1]. This implies that $f'(p_1)$ must have more than one turning point in [0,1]. The second derivative, $f''(p_1)$, however, is a linear function of p_1 and can only vanish at most one point in [0,1]. So when $\alpha < \beta$, $f'(p_1)$ does not have any positive roots in [0,1] and consequently $n_p \neq 3$, i.e., $n_p = 1$. We conclude again that $f(p_1)$ vanishes only at one point, $\bar{p}_1^*$, in the interval [0,1].

Let

$$z_0 = \beta r_2$$
$$z_1 = \frac{\alpha r_2 - 3\beta r_2}{3}$$
$$z_2 = \frac{3\beta r_2 - \alpha r_2 - r_1 - r_2}{3}$$
$$z_3 = (r_1 + r_2)(\alpha - \beta)$$

$$H = z_1 z_3 - z_2^2$$
$$G = z_0 z_3^2 - 3 z_1 z_2 z_3 + 2 z_2^3$$
$$E = G^2 + 4H^3$$

The roots, p_1^*, of the cubic polynomial $f(p_1)$ are found from the roots, x^*, of the cubic polynomial $g(x) = x^3 + 3Hx + G$ via the affine transformation

$$p_1^* = \frac{x^* - z_2}{z_3}$$

There are three cases to consider:

(i) If $E > 0$ then $g(x)$ has two complex roots and one real positive root given by

$$x^* = \left(\frac{-G+\sqrt{E}}{2}\right)^{\frac{1}{3}} - H\left(\frac{-G+\sqrt{E}}{2}\right)^{-\frac{1}{3}}$$

(ii) If $E = 0$ then $g(x)$ has three real roots, one repeated. They are given by either

$$-2\sqrt{-H}\ ,\ \sqrt{-H}\ ,\ \sqrt{-H}$$

or

$$2\sqrt{-H}\ ,\ -\sqrt{-H}\ ,\ -\sqrt{-H}$$

(iii) If $E < 0$ then $g(x)$ has three real distinct roots, given by

$$2\sqrt{-H}\cos\frac{\theta}{3},\ \ 2\sqrt{-H}\cos\frac{\theta+2\pi}{3},\ \ 2\sqrt{-H}\cos\frac{\theta+4\pi}{3}$$

$$\text{where}\ \ \cos\theta = -\frac{G}{2\sqrt{-H^3}}\,.$$

References

J.E.Huheey, Studies of warning coloration and mimicry. IV. A mathematical model of model-mimic frequencies, Ecology, 45(1) Winter 1964 185-188.

G.F.Estabrook and D.C.Jespersen, Strategy for a predator encountering a model-mimic system, The American Naturalist, 108(962) July-August 1974 443-457.

L.E.Bobisud and C.J.Potratz, One-trial versus multi-trial learning for a predator encountering a model-mimic system, The American Naturalist, 110(971) January-February 1976 121-128.

S.J.Arnold, The evolution of a special class of modifiable behaviors in relation to environmental pattern, The American Naturalist, 112(984) March-April 1978 415-427.

J.K.Luedeman, F.R.McMorris, and D.D.Warner, Predator encountering a model-mimic system with alternative prey, The American Naturalist, 117 1981 1041-1048.

D.Kannan, A Markov chain analysis of predator strategy in a model-mimic system, Bulletin of Mathematical Biology, 45(3) 1983 347-400.

R.E.Owen and A.R.G.Owen, Mathematical paradigms for mimicry: recurrent sampling, Journal of Theoretical Biology, 109 1984 217-247.

J.E.Huheey, Mathematical models of mimicry, The American Naturalist, 131 (Supplement) June 1988 S22-S41.

M.S.Speed, Muellerian mimicry and the psychology of predation, Animal Behaviour, 45 1993 571-580.

J.R.G.Turner, E.P.Kearney, and L.S.Exton, Mimicry and the Monte-Carlo predator: the palatability spectrum and the origins of mimicry, Biological Journal of the Linnean Society 23 1984 247-268.

J.R.G.Turner and M.P.Speed, Learning and memory in mimicry. I. Simulations of laboratory experiments, Phil. Trans. R.Soc. Lond. B 351 1996 1157-1170.

M.P.Speed and J.R.G.Turner, Learning and memory in mimicry: II. Do we understand the mimicry spectrum? Biological Journal of the Linnean Society 67 1999 281-312.

R.R.Bush and F.Mosteller, Stochastic models for learning, J.Wiley & Sons, New York, 1955.

M.Joron and J.L.B.Mallet, Diversity in mimicry: paradox or paradigm?, TREE 13(11) November 1998 461-466.

K.Narendra and M.A.L.Thathachar, Learning Automata: An Introduction, Prentice Hall, Englewood Cliffs NJ, 1989.

M.L.Tsetlin, Automaton Theory and Modeling of Biological Systems, Academic Press, New York, 1973.

W.L.Ferrar, Higher Algebra, Oxford University Press, Oxford, 1962.

M.P.Speed, Robot predators in virtual ecologies: the importance of memory in mimicry studies, Animal Behaviour, 57 1999 203-213.