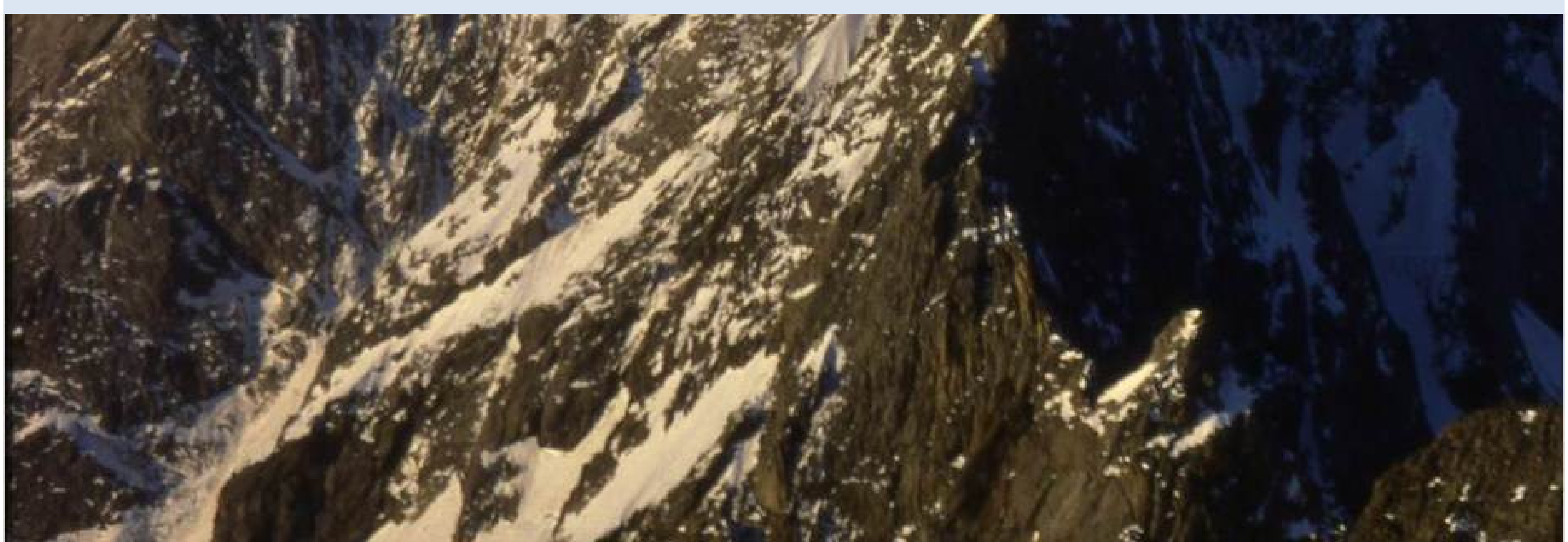
The Evolution of Stochastic Automata

By

David Eric Probert
Churchill College, Cambridge

June 1976



THE EVOLUTION OF STOCHASTIC AUTOMATA

BY

DAVID ERIC PROBERT

OF

CHURCHILL COLLEGE

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy, at the University of Cambridge.

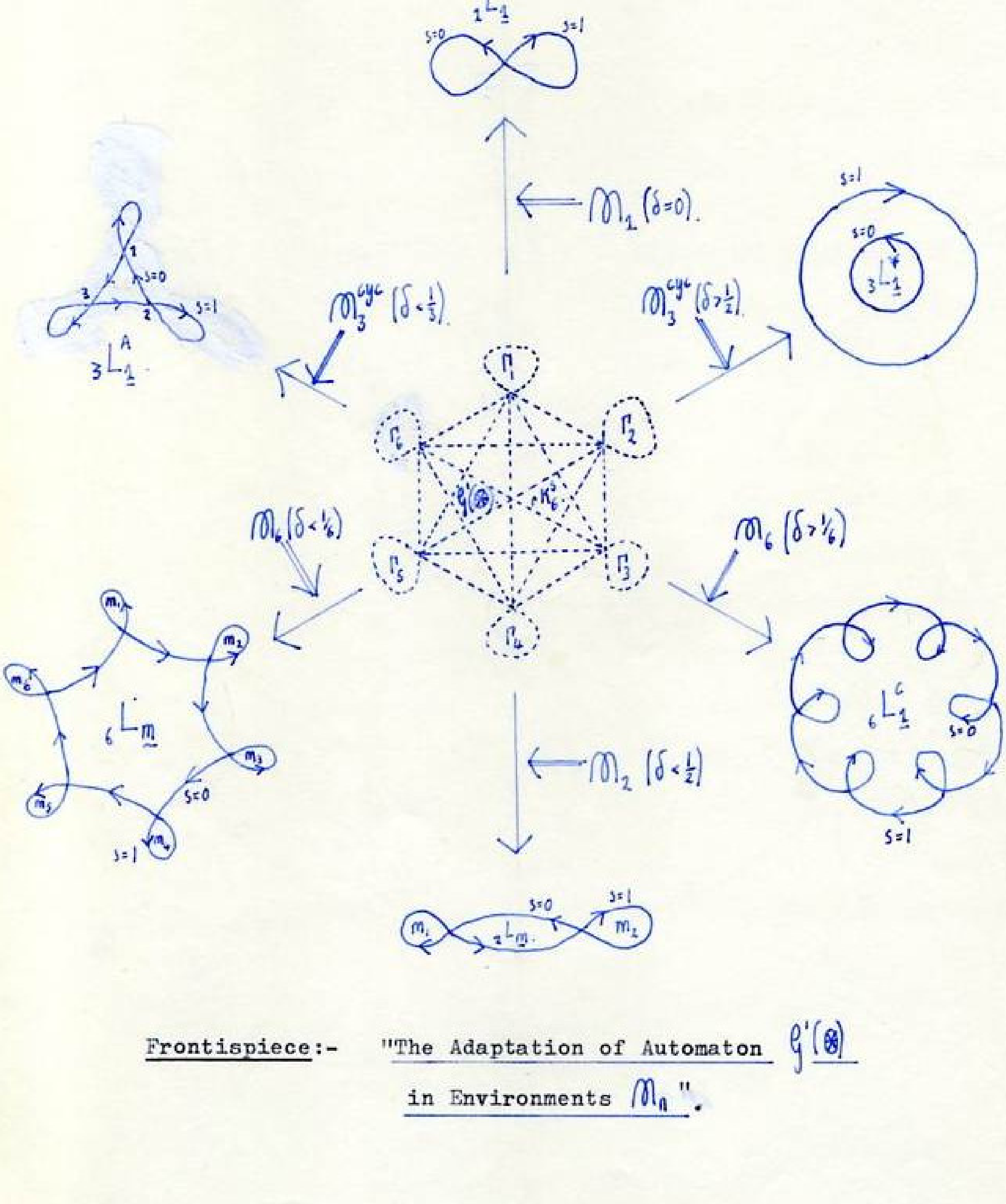
This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration.

I hereby declare that this dissertation entitled,

"The Evolution of Stochastic Automata",

has not been submitted for a degree or diploma at any other university.

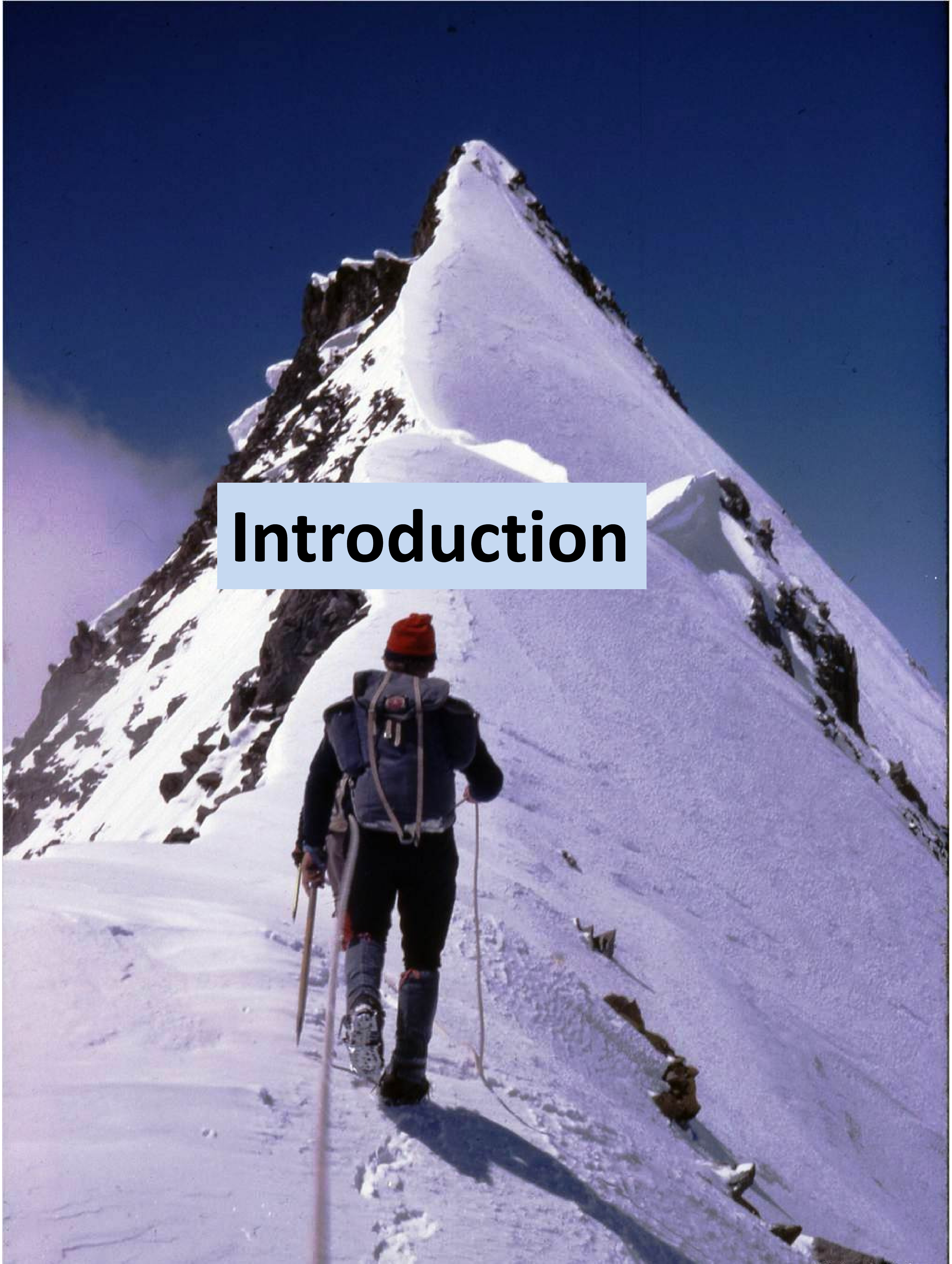
Signed:- D.E. Probert   Date:- 16th June 1976.



Frontispiece:- "The Adaptation of Automaton $\mathcal{G}'(\circledast)$ in Environments $\mathfrak{M}_n$".

Contents.

Introduction



0. Introduction.

Is there knowledge? it will vanish away; for our knowledge and our prophecy alike are partial and the partial vanishes when wholeness comes.

1 Corinthians 13 v 9.

0. Introduction.

0.1 Background, Motivation and Philosophy.

0.1.1 During the past 25 years a large number of papers have appeared on structural adaptation through reinforcement.

$$\text{chaos} \xrightarrow{\overset{\mathcal{M}}{\Downarrow}} \text{structure adapted to environment } \mathcal{M}$$

Much of this work can ultimately be traced back to the ideas of Skinner (operant conditioning) in the 1930's, and Wiener (feedback and cybernetics) and Von Neumann (automata) in the 1940's, which have since developed in many disciplinary niches.

0.1.2 The approach adopted in this thesis attempts to unify the conceptual basis in a basic evolving stochastic automaton named the π-cell, defined in 1.1. This element is essentially provided by the work of Bush and Mosteller (1955), but the framework of mathematical psychology has changed little since then, and has culminated in the profound mathematical treatment of Norman (1972). Other significant contributions have been provided by Lamperti and Suppes (1960) based on the work of Luce (1959), on the β-rule. This has since been placed on a more general basis by Kanal (1962) and Marley (1967), and much of learning theory was incorporated into random systems with complete connections by Iosifescu and Theodorescu (1969).

0.1.3 Norman (1975) actually relates the Hardy-Weinberg equations of mathematical genetics to a stochastic reinforcement mechanism similar to those of psychological learning theory. Indeed, the laws of mathematical genetics of Fisher (1930) are specific reinforcement mechanisms arising from survival or extinction of genotypes, with the selective viabilities acting as the environmental stimuli.

This has a natural comparison with Skinner's (1938) operant conditioning, where reinforcement arises from the reward or penalty received on executing an action , depending stochastically on its expediency within the environment.

Both genetics and learning theory can be represented by unstructured automata since in both fields we consider evolving distributions over genotypes or choices rather than allowing the automata to have an underlying network of transitions. In chapters 1 and 2 we consider such unstructured automata, whilst in chapter 3 we extend the theory to structured automata.

0.1.4. In 1961, the Russian mathematician Tsetlin published a pioneering paper on fixed structured automata, which opened up a new field of research linked with economic behaviour. Then in 1963 Vorontsova and Varshavskii considered the possibility of starting from an arbitrary structure, and defining reinforcement rules which give expedient adaptation to an environment. These structures evolving under reinforcement are essentially networks of π-cells with underlying digraphs (reward-penalty). It was then only possible to use computer simulations rather than theoretical techniques, since evolution by reinforcement raises many technical difficulties, which were treated by Norman (1968). The deterministic automata developed in Russia have been considered as models for biological systems, Tsetlin (1974), queuing systems and synchronisation, Varshavskii, Meleshina and Tsetlin (1965,1968) and power regulation, Stefanyuk and Tsetlin (1967).

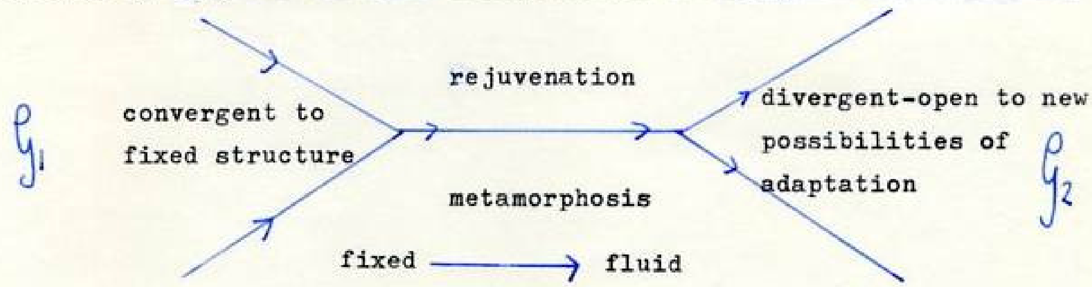
0.1.5. Thus, structural adaptation need not be solely biological, for we could consider the evolution of urban systems and cities, Bacon (1974), the differentiation of rôles within society, Whittle (1971), or the development of knowledge itself through research.

So we can consider our scientific knowledge as an evolving entity in which we abstract laws and build models from information we receive from the environment, and test these through further actions (experiments). A simple example of this is the development of the uniform polyhedra. The ancient Greeks knew of the Platonic solids, yet Kepler (1619) and subsequently Poinsot (1809) considered certain stellations and showed them to be uniform. Finally Coxeter et Al (1954) enumerated 75 such uniform polyhedra with increased complexity of kernal structure and stellation depth, and Skilling (1975) showed by direct computer search that the list was indeed complete. I have taken this example because the stellation process is similar to the process of increasing "memory depth" in structured automata, whilst preserving the "SOSA" property (chapter 3), and the kernal corresponds to an action-switch (3.4.5.).

0.1.6. But all structural adaptation appears to stem ultimately from "life". The reward-penalty stimuli used in our π-cell models are themselves derived concepts and more fundamentally we could view the actions executed by a structure,and the consequent structural reinforcement,as relating directly to its own survival. Pollution, in its broadest sense, is then that which tends to inhibit the reinforcement mechanism within the environment. Bernal (1967), Calvin (1969) and Cairns-Smith (1971) all consider how "life" could spontaneously arise on primeval earth.

Structures "feed" on -ve entropy, Schrödinger (1944), giving rise to hierarchical stability through "food" chains and ecological communities within the environment. In the absence of useful energy there is erosion and decay through the action of the 2nd law of thermodynamics. Glansdorff and Prigogine (1971) consider such an approach.

Structures may ultimately "stagnate" and become fixed, yet they may achieve higher forms of adaptation through a metamorphosis achieved by rejuvenation. (fluidisation of form)



$\mathcal{G}_i$ represent structures.

In chapter 3, we consider a stochastic automaton as an evolving entity with an initial random structure. Yet, through reinforcement, this probabilistic "fluid" form asymtotically becomes a deterministic "fixed" structure, acting expediently in its markovian environment. The limiting structure is not necessarily unique, so that it may be possible for the automaton to increase its payoff (adaptation) through a metamorphosis as outlined above, where fluidisation is translated mathematically as allowing state transitions to be probabilistic rather than deterministic.

0.1.7. In 1965, American control engineers Fu and Mc Murty became interested in the possibility of using learning automata as an alternative to the standard hill-climbing techniques, since virtually no prior information is required on environmental parameters and the rules themselves are simple. It is also quite easy to set several automata loose on a multimodal function in attempting to find a global maxima, as in Jarvis (1975). The paper of Fu and Mc Murty, based on the work of Vorontsova and Varshavskii (1963), initiated a new field which remained isolated from mathematical psychology until 1971. During the period 1966-71, Chandrasekaren and Shen (1967) formulated games between unstructured automata and performed computer simulations, later extended by Viswanathan and Narendra (1974). They demonstrated oscillatory behaviour which is treated analytically in this thesis. These unstructured games are also related to the structured games of Tsetlin (1963) and the π-cell games of chapter 2 seem a natural gaming approach.

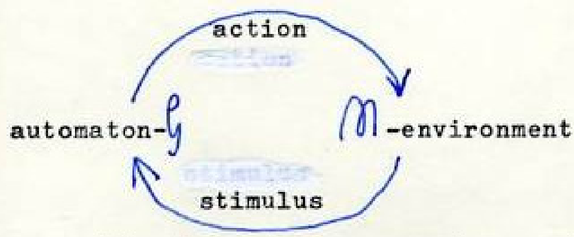
When the engineering literature was merged with the work of Norman, several fallacious stability arguments were revealed by a counter-example of Kushner, published by Viswanathan and Narendra (1971), and are considered further by Narendra and Thathachar (1974). The stability criteria used by the engineers were only valid in a deterministic rather than a probabilistic process, when flow against a probabilistic drift is possible for <u>all</u> time. The optimal reinforcement rules developed in this thesis actually do only require a deterministic stability criteria which is the technique of boundary learning (1.7). So the conditions are given under which the ideas of Chandrasekaren do carry over from a probabilistic to a deterministic control theoretic framework.

0.1.8. A series of papers by Cover and Hellman (1970) on hypothesis testing by finite memory, used static automata resembling those of Tsetlin. The finite memory constraint was contested by Chandrasekaren (1970) and at present no truly satisfactory definition is to be found, apart from in the information theory of Shannon (1949).

The π-cells themselves have ∞ memory in that a real number is held to arbitrary accuracy and similarly in the work of Cover and Hellman, a randomiser is used to generate arbitrarily small probabilities. Further, even with "∞ computing power" we obtain undecidable propositions in logic (as with Turing's automaton), which are essentially obtained through a +ve feedback of the theory on to itself,(symbolically), which gives knowledge a certain relativity.

This seems related to structural adaptation when ultimately, in setting up a mathematical framework, we may attain such intrinsic incompleteness. "Is an evolving automaton of sufficient complexity (a universal automaton) able to explain its own existence and motivate its continued evolution?" This is a deep philosophical question debated by philosophers through the ages, including Sartre and Camus in our own century.

Returning to the statistical framework, a completely independant theory of evolutionary operation was developed in America, primarily for optimising the operation of industrial chemical plants. A thorough treatment of this concept (EVOP) is covered in Box and Draper (1962 and 1969). The basic philosophy is very similar to that in this thesis in that we have operational research based on the biological theory of Darwin.



It seems possible that viewing a system, perhaps an industry, as an evolving entity in its environment, with competing systems, will prove a useful future frame of reference in operational research. This paradigm is indeed pursued in Day and Groves (1975) as a basis for future economic theory.

0.1.9. Models of brain mechanisms were initiated by Mc Culloch and Pitts in 1943, with their basic threashold neuron. This field of work has held a virtually independent existence, developing into the computer based field of pattern recognition and automatic clustering. Rosenblatt defined the α-perceptron in 1957, which was developed into the T.L.U.(threashold logic unit) used in the monograph of Nilsson (1965). These models do exhibit structural adaptation but I believe that a probabilistic setting is more natural for abstracting the notion of environmental uncertainty. However, this thesis does adopt an elemental approach based on the π-cell, instead of the neuron, as the unifying element for previous work both in psychology and control engineering.

0.1.10. Cellular automata, treated rigorously by Codd (1968), have been used by Conway as a model of "life", Gardner (1971), which is deterministic yet generates unpredictable patterns. However, in this thesis we are concerned with models that explicitly have actions executed by an automaton, with the resulting stimuli acting as the next input. In "life", we generate structural forms related to the specified rules of state transition, yet there is no structural adaptation within the environment. Such automata were formulated by Von Neumann (1948) as a model of self-reproduction. Richardson (1976) has also recently considered the self-replication of molecules, which would appear essential for structural reinforcement.

Kauffman (1969) considered cellular behaviour modelled with random genetic nets, also based on elemental "binary automata" with underlying deterministic digraph. These have been used as a practical model for learning by Aleksander (1971).

0.1.11. Recent interest in morphogenesis has been aroused by the treatise of Thom (1975), which embraces both biological and physical structural adaptation. We shall briefly consider certain similarities between boundary learning and catasrophes in 1.13, giving the theory of the evolving stochastic automaton similarities with topological morphogenesis. Indeed, in chapter 3, we shall see how a π-cell network gives a simple model for the underlying mechanism of cellular differentiation, without any form of centralised control.

0.1.12. The work of Bush and Mosteller (1955) was based on a learning π-cell, but the foundations have since become obscured by the analysis of reinforcement rules and parameter estimation. The π-cell is defined to use any uniformly learning rule (1.2 and 1.8). Having returned to the basic learning entity, the concept is extended to cover:-

a) Automata games.

b) Adaptation in dynamic environments.

c) Cellular differentiation.

d) Hierarchical adaptation.

My aim in this thesis is to emphasise the new conceptual framework rather than to dwell on the rigorous mathematical derivations, which are still incomplete, particularly for π-cell networks which require deep probabilistic ideas in their analysis.

For a deeper treatment of the background literature, I refer the reader to the excellent survey paper of Narendra and Thathachar (1974).

0.2 Summary.

The Π-cell corresponds to an evolving unstructured automaton and gives a generalisation to the work of Norman (1972). In successive chapters we shall analyse:-

1). The singleton Π-cell evolving in environment $\mathcal{M}$.

Existing reinforcement rules mathematical psychology are either conditionally optimal or ϵ-optimal. A theory of optimal reinforcement rules is developed and their properties investigated in both static and dynamic environments. The optimality is shown to depend on a uniform learning property and behaviour near the absorbing barriers. This is in contrast to conditionally optimal rules, Luce (1959), Lamperti and Suppes (1960), which are non-uniformly learning, and ϵ-optimal rules, Norman (1968), which are centrally rather than boundary learning (1.7).

2). Games between Π-cells.

The Π-cell is essentially a time dependant pie graph that "adapts" to its environment $\mathcal{M}$. Symbolically we shall use $\otimes_i$ to designate the i^{th} Π-cell. A game between Π-cells is now easily explained as the i^{th} Π-cell $\otimes_i$ acting in an environment $\mathcal{M}_i = \{\otimes_{j \neq i}\}$. Each Π-cell only knows the result of its own strategy and knows nothing of the behaviour of competing Π-cells. However, 2.5.1 gives Nash point convergence when such a point exists, whilst 2.3.3 gives optimal time-averaged payoff for zero-sum games with an equilibrium point of mixed strategies, in the deterministic approximation to the automata trajectories.

3). Networks of Π-cells.

The network consists of a set of Π-cells with an underlying probabilistic digraph $\mathcal{G}$, which is described by the two markov

transition matrices σ^s_{ij}. The superscript $s \in \{0,1\}$ represents the stimulus received, and this determines which transition matrix, σ^0_{ij} or σ^1_{ij} to use. Initially, each automaton state is associated with a specific action, but this is developed to allow a distribution over all actions in each state. The state space is partitioned into into sets of states using a particular π-cell, $\Gamma_i = \{ x_k : \otimes_i$ is used, where x_k denotes state $k \}$. The mechanism of reinforcement and mode of operation of the automata is described in detail in 3.1. Limiting structures are considered in 3.2 $\rightarrow$ 3.6.

In 3.7, the π-cell network is considered as a model for cellular differentiation. However, it has not yet been possible to give rigorous proofs owing to the complexity of the process but it is planned to carry out a program of computer simulations in the future to guide the theoretical insight.

Finally, in 3.9, we briefly consider the concept of hierarchies of π-cells, so that a π-cell network is a level-1 hierarchy $\mathcal{G}^1(\otimes)$ and a π-cell becomes a level-0 hierarchy $\mathcal{G}^0(\otimes)$. This provides a further unifying link and a basis for further research.

Another fruitful area for future research appears to be the community behaviour of automata for co-operation rather than the competition of chapter 2. This has been considered in particular by Chaikovskii (1968) and Golovchenko (1974), based on the work of Tsetlin et al (1963,1964,1965) and published in full in the collected works of Tsetlin (1974).

Community behaviour is considered in the examples of 2.5 and briefly in the "sheep effect" of 3.11.

I should like to acknowledge the help, encouragement and inspiration of my supervisor, Peter Whittle, during this period of research, which was undertaken during the tenure of an S.R.C. grant (1973-1976).

Chapter One
Unstructured Learning Automata

Chapter 1.

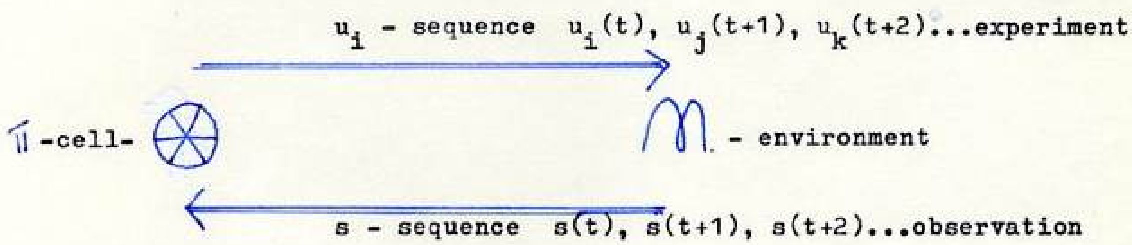
Is there a thing of which it is said," See, this is new"?

It has been already, in the ages before us.

Ecclesiastes 1 v 10.

1. Unstructured Automata.

1.1. The Singleton π-Cell.



Definition 1.1.1.

A π-cell is a specific form of evolving automaton with the following properties:-

Input stimuli $s \in \{0,1\}$ $s=0$ penalty, $s=1$ reward.

Output actions u_i $1 \le i \le n$.

State $\underset{\sim}{\pi} = (\pi_1 \ldots \pi_n)$, $\pi_i(t) = \Pr(\text{ output action } u_i \text{ at time } t \in N)$.

Transition $\left.\begin{array}{l} \pi_j(t+1) = \pi_j(t) + T_{ij} \text{ on receiving } s(t)=0 \\ \pi_j(t+1) = \pi_j(t) + S_{ij} \quad " \quad " \quad s(t)=1 \end{array}\right\}$ action i used at time t.

and normalise $\sum_i \pi_i(t+1) = 1$.

We now define the the environment $\mathfrak{M}(\Delta_{\alpha\beta}, q_{u_i}^{\alpha,s})$.

$\Delta_{\alpha\beta} = \Pr(E_\alpha(t) \to E_\beta(t+1))$. where E_α is the environment state α.

$q_{u_i}^{\alpha,s} = \Pr(\text{stimulus is } s \mid u_i(t) \text{ and } E_\alpha(t))$.

and $q_{u_i}^{\alpha,1} = q_{u_i}^{\alpha} = 1 - p_{u_i}^{\alpha}$

Until section 1.12 we shall have a static $\mathfrak{M}$:- $\Delta_{\alpha\beta} = \delta_{\alpha\beta}$ (delta function)

Thus under static $\mathfrak{M}$ the environment state remains the same for <u>all</u> time.

Now define $R(\pi(t)) = \sum_i q_i \pi_i(t)$. the expected reward. The transition rule is <u>uniformly learning</u> if $R(\pi)$ is a <u>sub-martingale</u>.

In 1.2 we find conditions under which T_{ij} and S_{ij} are U.L. (uniformly learning).

The transition rule is usually refered to as a reinforcement rule by the mathematical psychologists.

The rule is <u>optimal</u> if $\lim\limits_{t\to\infty} \pi_i(t) = 1$ only if $q_i \geq q_j$ for all j. Vorontsova (1965) considered non-linear reinforcement rules in continuous time, for 2 actions, and gave conditions for a family of rules to be optimal. However, in the survey paper of Narendra (1974), it was still unknown whether such rules gave optimal discrete time behaviour. Linear rules of the form $T_{ij} = 0$, $S_{ij} = \theta$ $S_{ii} = \theta(1-\pi_i)$ have been shown by Norman (1968) to be ε-optimal, that is:-

$$\lim_{\theta\downarrow 0} \lim_{t\to\infty} \pi_i(t) = 1 \text{ when } q_i > q_j \text{ for all } j \neq i.$$

But no discrete time rules have been proven optimal independant of $\underset{\sim}{q}$.

In this chapter, we shall prove the existence of a family of optimal discrete time rules for n-actions, and show that this optimality is only dependant on boundary behaviour. This boundary learning is essential when we consider π-cell networks and we test absorbing barriers at boundaries for probabilistic stability with respect to the optimal reinforcement rules.

1.2. <u>Uniform Learning</u>.

For any random variable X(t), we denote the expected increment by $\Delta X(t) = (X(t+1) \mid X(t) \ldots X(0)) - X(t)$.

Thus a rule is U.L. if $\Delta R(\pi(t)) \geq 0$. Rather than state the rules that we shall be using, I shall indicate why we are restricted to a certain family, through a series of lemmas.

Theorem 1.2.1. (Whittle)

The necessary and sufficient conditions on feasible S_{ij} and T_{ij} to give $\Delta R \geq 0$ are:-

a). $\sum_i \pi_i S_{ij} = \sum_i \pi_i T_{ij} = 0.$

and b). $\Lambda \geq 0$ where $\lambda_{ij} = \frac{1}{2}(\pi_i(S_{ij}-T_{ij})+\pi_j(S_{ji}-T_{ji}))$.

Proof

We have $\Delta R = \sum_i q_i \Delta\pi_i = \sum_i q_i(\pi_j q_j S_{ji} + \pi_j p_j T_{ji})$.

$= q'\Lambda q + \sum_j (\sum_i \pi_i T_{ij}) q_j$. i

$= p'\Lambda p + \sum_j (\sum_i \pi_i S_{ij}) p_j$. ii

For necessity, we just consider a small perturbation $\underset{\sim}{\epsilon}$ from $q_i = q_j \; \forall \; i,j$. Put $\underset{\sim}{q} = c\underset{\sim}{1} + \underset{\sim}{\epsilon}$, then $\Delta R = \underset{\sim}{\epsilon}'\Lambda\underset{\sim}{\epsilon} + \sum_j (\sum_i \pi_i T_{ij})\epsilon_j$ as $\underset{\sim}{1}'\Lambda\underset{\sim}{1} = 0$ and $\sum_j S_{ij} = 0$ by normalisation. Now to prove $\Delta R \geq 0$, we just require $\underset{\sim}{\epsilon}$ sufficiently small and then we choose $\underset{\sim}{\epsilon}$ to give $\sum_j (\sum_i \pi_i T_{ij})\epsilon_j < 0$ which is always possible if $\sum_i \pi_i T_{ij} \neq 0$. (Note that $\underset{\sim}{\epsilon}'\Lambda\underset{\sim}{\epsilon} \sim O(|\underset{\sim}{\epsilon}|^2)$ and can be neglected w.r.t. $\sum_{ij} \pi_i T_{ij}\epsilon_j$.) Hence $\sum_i \pi_i T_{ij} = 0$ and $\Lambda \geq 0$ are necessary. Similarly $\sum_i \pi_i S_{ij} = 0$ from ii above. The conditions above clearly sufficient.

Lemma 1.2.2.

U.L. $\Rightarrow$ ($\Delta\pi_k \equiv 0$ for all k if $q_i = q_j \; \forall$ i,j).

Proof. We let $q_i = q_k$ for all i and evaluate $\Delta\pi_k$.

$\Delta\pi_k = q_k \sum_i \pi_i S_{ik} + p_k \sum_i \pi_i T_{ik} = 0$ by condition a) above, and we actually also have:- (If $\underset{\sim}{q} = c\underset{\sim}{1}$ for any c $\Rightarrow \Delta\pi_k \equiv 0 \; \forall$ k) $\Leftrightarrow$ a), in 1.2.1.

Lemma 1.2.3.

The linear rule $T_{ij} = 0$, $S_{ij} = -\theta_{ij}\pi_j$, $S_{ii} = \theta_{ii}(1-\pi_i)$ is U.L. iff $\theta_{ij} = \theta \; \forall$ i,j.

Proof.

$\sum \pi_i = 1 \Rightarrow \sum_j (\theta_{ii} - \theta_{ij})\pi_j = 0$ and hence $\theta_{ii} = \theta_{ij} \; \forall$ i,j and lemma 1.2.2 gives $\sum_i (\theta_{jj} - \theta_{ij})\pi_i = 0$ and so we have $\theta_{jj} = \theta_{ij} \; \forall$ i,j.

Combining the above, $\theta_{ij} = \theta$ = constant.

And $\Delta R = \sum_{ij} \frac{1}{2}\theta\pi_i(q_i-q_j)^2\pi_j \geq 0$ gives sufficiency.

We cannot extend 1.2.3 to give U.L. $\Rightarrow \theta_{ij}(\underset{\sim}{\pi}) = \theta(\underset{\sim}{\pi})\ \forall i,j$, for non-linear rules but the final three lemmas in this section go as far as we can.

Lemma 1.2.4.

The non-linear rule $T_{ij} \equiv 0$, $S_{ij} = -\theta_i(\underset{\sim}{\pi})\pi_j$ and $S_{ii} = \theta_i(\underset{\sim}{\pi})(1-\pi_i)$ is U.L. iff $\theta_i(\underset{\sim}{\pi}) = \theta(\underset{\sim}{\pi})\ \forall i$.

Proof.

U.L. $\Rightarrow \sum_i \pi_i S_{ij} = 0$ and hence $\sum_j \pi_j (\theta_i(\underset{\sim}{\pi}) - \theta_j(\underset{\sim}{\pi})) = 0 \quad \forall i$.

Hence $\theta_i(\underset{\sim}{\pi}) = \theta_j(\underset{\sim}{\pi})\ \forall$ i.j and $\Delta R = \sum \frac{1}{2}\theta(\underset{\sim}{\pi})\pi_i(q_i - q_j)^2 \pi_j$ gives sufficiency.

Lemma 1.2.5.

If we restrict the no of actions to n = 2, then the non-linear rule $T_{ij} = 0$, $S_{ij} = -\theta_{ij}(\underset{\sim}{\pi})\pi_j$ $i \neq j$ and $S_{ii} = \theta_{ii}(\underset{\sim}{\pi})(1-\pi_i)$ is U.L. iff $\theta_{ij}(\underset{\sim}{\pi}) = \theta(\underset{\sim}{\pi})$ i,j.

Proof.

We have sufficiency as in 1.2.4 and for the necessity.

$\sum \pi_i = 1$ gives $\theta_{11} = \theta_{12}$ and $\theta_{22} = \theta_{21}$

whilst $\sum_i \pi_i S_{ij} = 0$ gives $\theta_{11} = \theta_{21}$ and $\theta_{22} = \theta_{12}$

Hence we have $\theta_{ij}(\underset{\sim}{\pi}) = \theta(\underset{\sim}{\pi})\ \forall i,j$.

Lemma 1.2.6.

The non-linear rule $T_{ij} \equiv 0$, $S_{ij} = -\theta_{ij}(\underset{\sim}{\pi})\pi_j$ $i \neq j$ and $S_{ii} = \theta_{ii}(\underset{\sim}{\pi})(1-\pi_i)$ $n > 2$ is U.L. if $\theta_{ij}(\underset{\sim}{\pi}) = \theta_{ji}(\underset{\sim}{\pi})$ and $\theta_{ii}(\underset{\sim}{\pi}) = \sum_j \theta_{ij}(\underset{\sim}{\pi})\pi_j$ (normalisation).

Proof.

$$\Delta R = \sum_i q_i \Delta\pi_i(t) = \sum_i q_i \pi_i \sum_j \pi_j (\theta_{ii} q_i - \theta_{ij} q_j)$$

$$= \sum_i \theta_{ii} q_i^2 \pi_i - \sum_{i,j} \theta_{ij} q_i q_j \pi_i \pi_j = \sum_{i,j} \theta_{ij} q_i \pi_i \pi_j (q_i - q_j)$$

$$= \frac{1}{2} \sum_{i,j} \theta_{ij}(\underset{\sim}{\pi}) \pi_i \pi_j (q_i - q_j)^2 \geq 0 \qquad 0 < \theta_{ij}(\underset{\sim}{\pi}) < 1.$$

and hence the rule is U.L. and normalisation gives $\theta_{ii}(\underset{\sim}{\pi}) = \sum_j \theta_{ij}(\underset{\sim}{\pi})\pi_j$

Actually it is necessary to have $\sum_i \pi_i \theta_{ij} = \sum_i \theta_{ji}\pi_i$ using $\sum_i \pi_i S_{ij} = 0$. Thus asymtotically as $\pi_i \uparrow 1$ we must have $\forall j$, $\theta_{ij}(\underset{\sim}{\pi}) \to \theta_{ji}(\underset{\sim}{\pi})$. So 1.2.6. is really the best practical rule we can formulate.

U.L. is sufficient for later theorems but not necessary throughout $\pi_i \in I = [0,1]$ and optimality of the $\theta_{ij}(\underset{\sim}{\pi})$ rules will be characterised by certain boundary properties. In 1.6.7. we briefly discuss the possible use of T_{ij} in addition to S_{ij}. First we consider the convergence of U.L. rules, which follows from semi-martingale theorems.

1.3. <u>Convergence</u>.

To ensure that $\exists$ i s.t. $\lim_{t\to\infty} \pi_i(t) = 1$, we shall use the symmetric $\theta_{ij}(\underset{\sim}{\pi})$ rules which are the natural extension of the 2-action $\theta_{ij}(\underset{\sim}{\pi}) = \theta(\underset{\sim}{\pi})$ rules. Under the same condition, $\theta(\underset{\sim}{\pi}) = 0$ if and only if some $\pi_i = 1$, we obtain boundary convergence for $\theta(\underset{\sim}{\pi})$ family, but apart from 2-actions, conditions for optimality are still unknown. e.g. $\theta(\underset{\sim}{\pi}) = \prod_{i=1}^{n}(1 - \pi_i)$. All U.L. rules converge by s/mg theorems, but the difficulty lies in obtaining a.s. absorption in boundary, $\pi_i \to 1$ for some i.

On taking action i at time t, $u_i(t)$.

$$\pi_j(t+1) = \pi_j(t)(1 - \theta_{ij}(\underset{\sim}{\pi})) + \delta_{ij}\theta_{ij}(\underset{\sim}{\pi}) \qquad s(t)=1$$

$$\pi_j(t+1) = \pi_j(t). \qquad s(t)=0$$

We denote this family of rules by $\mathcal{R}$, where $\theta_{ij}(\underset{\sim}{\pi})$ iff $\pi_i = 0$ or $\pi_j = 0$.

<u>Theorem</u> 1.3.1.

Under $\mathcal{R}$, $\lim_{t\to\infty} \pi_i(t) \in \{0,1\}\ \forall i.$

<u>Proof</u>

First we must prove that the limit exists.

$\Delta R \geq 0$ and hence $\lim_{t\to\infty} R(\pi(t)) \to R^*$ by s/martingale theorem.

And $\Delta\pi_k = \pi_k \sum_j \theta_{kj}(\underset{\sim}{\pi})\pi_j(q_k - q_j)$ for k s.t. $q_k \geq q_j\ \forall j$.

Since π_k is bounded $\pi_k \xrightarrow{a.s.} V_k \in [0,1]$ by s/mg theorem.

a). First suppose $\Delta\pi_k > 0$ and $V_k \in (0,1)$

Now by Doob (1953), as $|\pi_k(t)|$ is uniformly integrable, we have

$\lim_{t\to\infty} E|\bar{\pi}_k(\infty) - \bar{\pi}_k(t)| = 0.$ i

But $E(\bar{\pi}_k(t+1)|\bar{\pi}_k(t)) - \bar{\pi}_k(t) = \Delta\bar{\pi}_k > 0$ for $\bar{\pi}_k \in (0,1)$

and so $E(\bar{\pi}_k(t+1)|\bar{\pi}_k(0)) - E(\bar{\pi}_k(t)|\bar{\pi}_k(0)) = \int_\Omega \Delta\bar{\pi}_k(t)\, d\mu > 0$

and $\lim_{t\to\infty} \Delta\bar{\pi}_k(t) = \lim_{t\to\infty} \int_\Omega \Delta\bar{\pi}_k(t)\, d\mu > 0$ ii

(see Sawaragi and Baba (1974) for (Ω, F_∞, μ)).

where the limit exists as $\Delta\bar{\pi}_k$ converges a.s. and is bounded absolutely

But $\lim_{t\to\infty} E(\bar{\pi}_k(t)) = E(\bar{\pi}_k(\infty)) = \nu_k$ by i and so $\lim_{t\to\infty} \Delta\bar{\pi}_k(t) = 0$

We have a contradiction unless $\nu_k \in \{0,1\}$.

b) Now let $\Delta\bar{\pi}_k \equiv 0$, which for U.L. rules occurs iff $q_i = q_j \;\forall j$.

Again we use Doob, since $|\bar{\pi}_k(t)|^2$ is U.I. $\lim_{t\to\infty} |\bar{\pi}_k(\infty) - \bar{\pi}_k(t)|^2 = 0$

But the variance of increment $= E(\bar{\pi}_k(t+1) - \bar{\pi}_k(t) - \Delta\bar{\pi}_k(t))^2$.

$= E(\bar{\pi}_k(t+1) - \bar{\pi}_k(t))^2 > 0$ for $\bar{\pi}_k \in (0,1)$.

and $E((\bar{\pi}_k(t+1) - \bar{\pi}_k(t))^2 | \underset{\sim}{\bar{\pi}}(t)) = q_k\bar{\pi}_k \left[(\theta_{kk}(1-\bar{\pi}_k))^2 + (\sum_{i\neq k} \theta_{ik}^2 \bar{\pi}_i\bar{\pi}_k) \right]$

So $\lim_{t\to\infty} E((\bar{\pi}_k(t+1) - \bar{\pi}_k(t))^2 | \underset{\sim}{\bar{\pi}}(0)) = \lim_{t\to\infty} \int_\Omega q_k\bar{\pi}_k \left((\theta_{kk}(1-\bar{\pi}_k))^2 + \sum_{i\neq k} \theta_{ik}^2 \bar{\pi}_i\bar{\pi}_k \right) d\mu$

> 0 as $\nu_k \in (0,1)$.

Yet $\lim_{t\to\infty} E(\bar{\pi}_k^2(t+1)) = E(\bar{\pi}_k^2(\infty))$ and we have a contradiction.

Hence both conditional expectation and variance vanishing give us boundary convergence of $\bar{\pi}_k$.

c) If the rule is optimal, we have $\bar{\pi}_i \to 1$ only if $q_i \geq q_j \;\forall j$ by defn.

d) Now re-order suffices to give $q_1 \geq q_2 \geq \dots \geq q_n$, to give $\bar{\pi}_i \to \nu_i \in \{0,1\} \;\forall i$. Note that $q_n \leq q_j$ j gives $\bar{\pi}_n \to \nu_n \in \{0,1\}$ by the same reasoning as a).

Suppose $\lim_{t\to\infty} \bar{\pi}_1(t) = 1$ then result is immediate

Thus let $\bar{\pi}_1(t) \to 0$ and consider sign($\Delta\bar{\pi}_2$)

We have $\Delta\bar{\pi}_2 = \sum_j \bar{\pi}_2\theta_{2j}(q_2 - q_j)\bar{\pi}_j$ and sgn($\Delta\bar{\pi}_2$) can only alternate infinitely often if $\lim_{t\to\infty} \bar{\pi}_2(t) \in \{0,1\}$.

Suppose sgn($\Delta\bar{\pi}_2$) alternates finitely often, then for $t > t^*$ say, we have $\pi_2(t)$ is a s/martingale and $\pi_2 \to \nu_2 \in [0,1]$ and by a) and b) $\nu_2 \in \{0,1\}$.

Now by induction, if $\lim_{t\to\infty} \pi_i(t) = 0$ for $0 < i \leq r$ we require $\lim_{t\to\infty} \pi_{r+1}(t) \in \{0,1\}$, which is easily proved by adapting above. If $\lim_{t\to\infty} \pi_{r+1}(t) = 1$ we are done, else continue until all suffixes are exhausted, noting if $\lim_{t\to\infty} \pi_i(t) = 0$, $0 < i \leq n-1$, then $\pi_n(t) \to 1$ by normalisation. //

An alternative proof of a) and b) could be based on # up-crossings over any rational interval is $\overset{a.s}{<} \infty$, Breiman (1968), and hence to prove $\nu_i \in \{0,1\}$, we need only note conditional variance > 0 at $\nu_i \in (0,1)$. and so $\pi_i \to \nu_i \in (0,1) \Rightarrow$ # up crossings $\overset{a.s}{=} \infty$ across any sufficiently short interval $[\nu_i - \epsilon, \nu_i + \epsilon]$ giving contradiction.

Since optimality will be shown to arise from $\theta_{ij}(\underset{\sim}{\pi}) \downarrow 0$ as $\pi_i \downarrow 0$ or $\pi_j \downarrow 0$, sufficiently fast, it may be thought that $\theta(t)$ would also give optimal rules. However, even though $\pi_i \to \nu_i \in [0,1]$ there is now no reason to restrict $\nu_i \in \{0,1\}$; we may only have time dependence occurring through $\underset{\sim}{\pi}(t)$ to give our $\theta_{ij}(\underset{\sim}{\pi}(t))$ rules

Similarly for $\theta(\underset{\sim}{\pi}(t))$ rules over > 2 actions; if $\theta(\underset{\sim}{\pi}(t)) \downarrow 0$ as $\pi_i \downarrow 0$, then we need not have boundary absorption.

<u>Corollory</u> 1.3.2.

Under U.L. $\theta(\underset{\sim}{\pi})$ rules with $\theta(\underset{\sim}{\pi}) \downarrow 0$ iff $\pi_i \uparrow 1$ for some i, then $\lim_{t\to\infty} \pi_k(t) \in \{0,1\}$ $\forall$ k.

<u>Proof</u>

Use a) and b) of 1.3.1, which hold since conditional variance only vanishes at boundary. Indeed b) is sufficient but it is a) which brings in the concept of probabilistic drift, which is the key to discussions of optimality. //

In the next sections, we shall initially just consider 2-actions which are extended to $n > 2$ in 1.6. Laksmirvarahan and Thathachar (1973) considered uniformly learning $\theta(\underset{\sim}{\pi})$ rules over n-actions, yet seemed unaware that they are not necessarily absorbed in $\pi_i = 1$.

boundaries. We use 2-action $\theta(\underset{\sim}{\pi})$ rules which are not necessarily symmetric in π_1 and π_2 . Previous authors have assumed $\theta(\pi_1) = \theta(\pi_2) = K(\pi_1\pi_2)^\alpha$ say, when it is not at all necessary. In the general n-action $\theta_{ij}(\underset{\sim}{\pi})$ rules, we need only have the condition $\theta_{ij}(\underset{\sim}{\pi}) \downarrow 0$ iff $\pi_i \downarrow 0$ or $\pi_j \downarrow 0$.

Now we have our π-cell $\underset{\sim}{\Delta}$

$\circledast \sim$

Output	$u_i(t)$
Input	$s(t) \in \{0,1\}$ with resp prob $\{p_{u_i}, q_{u_i}\}$
State	$\pi_i(t) = \Pr(\ u_i(t),\ t \in N\ )$
Transition	$\pi_j(t+1) = \pi_j(t)(1-\theta_{ij}(\underset{\sim}{\pi})) + \delta_{ij}\theta_{ij}(t)$ when $s(t)=1$ and $u_i(t)$.
	$\pi_j(t+1) = \pi_j(t)$ when $s(t)=0$.
	and $\theta_{ij}(\underset{\sim}{\pi}) = \theta_{ji}(\underset{\sim}{\pi})$, $\sum_i \pi_i(t)=1$.
	$\theta_{ij}(\underset{\sim}{\pi}) \downarrow 0$ iff $\pi_i \downarrow 0$ or $\pi_j \downarrow 0$.
	e.g $\theta_{ij}(\underset{\sim}{\pi}) = K(\pi_i\pi_j)^\alpha$
U.L.	$\Delta R \geq 0$ with equality iff $\underset{\sim}{q} = c\underset{\sim}{1}$.

We now determine conditions under which a π-cell asymtotically a.s. picks the optimal action. Then we shall have the complete properties of an isolated learning $\circledast$, with its set of reinforcement rules.

1.4 Comparison

We shall now prove that all rules in our non-linear family are at least ϵ-optimal. In the literature, such as the paper of Sawaragi (1974), each rule has been treated separately, whilst the ϵ-optimality arise fundamentally from the comparison theorem 1.4.5, proved through a series of lemmas. This is extended to n-actions in 1.6.6.

We define the operator $U_{\theta(\underset{\sim}{\pi})}\psi(\underset{\sim}{\pi}(t)) = E(\psi(\underset{\sim}{\pi}(t+1)) \mid \underset{\sim}{\pi}(t))$, first introduced by Norman (1968) for the linear rules $\theta(\underset{\sim}{\pi}) = \theta$ = constant.

Theorem 1.4.1

If $\gamma_{\theta(\pi)}(\pi_\cdot) = Pr(\lim_{t\to\infty} \pi_\cdot(t) = 1 \mid \pi_\cdot(0) = \pi_\cdot)$ then $U_{\theta(\pi)}\gamma_{\theta(\pi)} = \gamma_{\theta(\pi)}$ and if $\gamma_{\theta(\pi)}$ is continuous, it is unique.

Proof.

We know by 1.3.1 that $\lim_{t\to\infty} \pi_i(t) \in \{0,1\}$, and hence iterating $U_{\theta(\pi)}$ we obtain $\lim_{n\to\infty} U^n_{\theta(\pi)}\gamma_{\theta(\pi)} = \mathcal{E}(\gamma_{\theta(\pi)}(\pi_\cdot(\infty)) \mid \pi_\cdot(0)) = 1.\gamma_{\theta(\pi)}(\pi)$ and we have the result from $\lim_{n\to\infty} U^n\gamma = U \lim_{n\to\infty} U^{n-1}\gamma = U\gamma = \gamma.$

Indeed, if $\psi(i) = i$ for boundary $i \in \{0,1\}$, then $\lim_{n\to\infty} U^n_{\theta(\pi)}\psi(\pi) = \gamma_{\theta(\pi)}(\pi).$ easily, by the same argument.

Finally, let $\gamma^*_{\theta(\pi)}$ be another solution to $U_{\theta(\pi)}\gamma^* = \gamma^*$ which is continuous. Then $\lim_{n\to\infty} U^n_{\theta(\pi)}\gamma^* = \gamma^*$, by cty, but we also have $\lim_{n\to\infty} U^n_{\theta(\pi)}\gamma^*_{\theta(\pi)} = \gamma_{\theta(\pi)}$ and this gives uniqueness. We could write this fundamental equation more concisely as $\Delta\gamma_{\theta(\pi)}(\pi_\cdot) \equiv 0.$ //

Definition. 1.4.2.

A function $\psi(\pi)$ on $[0,1]$ with $\psi(i) = i$ at $i \in \{0,1\}$ is super-regular (sub-regular) if $\psi(\pi) \geqslant (\leq)\ U\psi(\pi)$ $\forall \pi \in [0,1]$.

(i.e. $\Delta\psi(\pi) \leq (\geqslant) 0$)

We may find that $U_{\theta(\pi)}$ has no continuous solutions, in particular, if the rule is optimal, we have $\gamma_{opt}(\pi_\cdot) \equiv 1$ for $\pi \in (0,1]$, $\gamma(0) = 0$.

To prove that $U_{\theta(\pi)}$ is the operator for an optimal rule, we construct a continuous sub-regular family $\psi_\delta(\pi)$ with $\lim_{\delta\downarrow 0} \psi_\delta(\pi) = \gamma_{opt}(\pi)$ the discontinuous limit. Then since $\lim_{n\to\infty} U^n\psi_\delta = \gamma\ \ \forall\delta$, the sub-regularity $U\psi_\delta > \psi_\delta$ ensures $\gamma_{\theta(\pi)}(\pi) = \gamma_{opt}(\pi)$. The sub-regular family would give us a contradiction if we asserted any other solution, apart from γ_{opt}, as giving the absorption probabilities.

Now we prove the convexity of $\gamma_\theta(\pi)$, for the linear rule, with $q_2 \geqslant q_1$.

We have $U_\theta\phi(\pi) = (\pi_1 p_1 + \pi_2 p_2)\phi(\pi) + \pi_1 q_1 \phi(T_1\pi) + \pi_2 q_2 \phi(T_2\pi)$

where $T_2\pi = (1-\theta)$ and $T_1\pi = T_2\pi + \theta$.

or $$U_\theta\phi(\bar\pi) - \phi(\bar\pi) = \bar\pi_1 q_1(\phi(T_1\bar\pi) - \phi(\bar\pi)) + \bar\pi_2 q_2(\phi(T_2\bar\pi) - \phi(\bar\pi))$$

with $\phi(i) = i \quad i \in \{0, 1\}$.

and also define $V_{\bar\pi}$ by,

$$V_{\bar\pi^*}\phi(\bar\pi) = (\bar\pi_1^* p_1 + \bar\pi_2^* p_2)\phi(\bar\pi) + \bar\pi_1^* q_1 \phi(T_1\bar\pi) + \bar\pi_2^* q_2 \phi(T_2\bar\pi).$$

with $U\phi(\bar\pi) = V_{\bar\pi}\phi(\bar\pi)$.

<u>Lemma</u> 1.4.3.

$\gamma_\theta(\bar\pi)$ is monotone increasing on $\bar\pi \in [0,1]$

<u>Proof</u>.

Take ϕ as any monotone increasing function with $\bar\pi' \le \bar\pi'' \Rightarrow \phi(\bar\pi') \le \phi(\bar\pi'')$ then we show the same is true for $U\phi$ and hence γ_θ, as $\lim_{n\to\infty} U^n\phi = \gamma$.

Now $$U_\theta\phi(\bar\pi'') - U_\theta\phi(\bar\pi') = V_{\bar\pi''}\phi(\bar\pi'') - V_{\bar\pi'}\phi(\bar\pi')$$
$$\ge V_{\bar\pi'}\phi(\bar\pi'') - V_{\bar\pi'}\phi(\bar\pi') \ge 0 \qquad (*)$$

where * follows from monotonicity of ϕ and the following:-

$$V_{\bar\pi''}\phi(\bar\pi'') - V_{\bar\pi'}\phi(\bar\pi'') = (\bar\pi_1'' - \bar\pi_1')q_1(\phi(T_1\bar\pi'') - \phi(\bar\pi'')) + (\bar\pi_2' - \bar\pi_2'')q_2(\phi(\bar\pi'') - \phi(T_2\bar\pi'')) \ge 0 \quad //$$

<u>Lemma</u> 1.4.4.

$\gamma_\theta(\bar\pi)$ is convex on $\bar\pi \in [0,1]$

<u>Proof</u>.

Assume $q_1 < q_2$; as $\gamma_\theta(\bar\pi) = \bar\pi$ at $q_1 = q_2$.

Then we take ϕ convex and show that $U\phi$ and hence γ_θ is convex.

Take $\alpha, \beta \ge 0 \quad \alpha + \beta = 1$. Then for $\bar\pi' < \bar\pi''$.

$$\alpha V\phi(\bar\pi') + \beta V\phi(\bar\pi'') - V\phi(\alpha\bar\pi' + \beta\bar\pi'') = \alpha V_{\bar\pi'}\phi(\bar\pi') + \beta V_{\bar\pi''}\phi(\bar\pi'') - V_{\alpha\bar\pi' + \beta\bar\pi''}\phi(\alpha\bar\pi' + \beta\bar\pi'')$$

$$\ge \alpha V_{\bar\pi'}\phi(\bar\pi') + \beta V_{\bar\pi''}\phi(\bar\pi'') - V_{\alpha\bar\pi' + \beta\bar\pi''}(\alpha\phi(\bar\pi') + \beta\phi(\bar\pi''))$$ by convexity of

$$= \alpha V_{\bar\pi'}\phi(\bar\pi') + \beta V_{\bar\pi''}\phi(\bar\pi'') - (\alpha V_{\bar\pi'} + \beta V_{\bar\pi''})(\alpha\phi(\bar\pi') + \beta\phi(\bar\pi''))$$ and linearity of T_j.

$$= \alpha\beta(V_{\bar\pi''} - V_{\bar\pi'})(\phi(\bar\pi'') - \phi(\bar\pi')) = \alpha\beta\delta D$$ where $\delta = \bar\pi_1'' - \bar\pi_1' = \bar\pi_2' - \bar\pi_2'' > 0$

and $$D = q_2[\phi(T_2\bar\pi') - \phi(\bar\pi') - \phi(T_2\bar\pi'') + \phi(\bar\pi'')] - q_1(\phi(T_1\bar\pi') - \phi(\bar\pi') - \phi(T_1\bar\pi'') + \phi(\bar\pi'')).$$

$= q_2A_2 - q_1A_1$ say and $D \ge 0$ if $A_2 > A_1$ or

$$\phi(T_1\bar\pi') - \phi(T_2\bar\pi') \le \phi(T_1\bar\pi'') - \phi(T_2\bar\pi'') \qquad (**)$$

But $T_1\bar\pi - T_2\bar\pi = \theta$ = const and $\bar\pi' \le \bar\pi''$ and hence (**) follows by convexity.

//

Theorem 1.4.5.

If $\theta(\bar{u}) \le \theta$ then $U_{\theta(\bar{u})} \gamma_\theta \le \gamma_\theta$.

Proof

We use $\theta^* = \theta(\bar{u})$ and note that γ_θ convex gives:-

$[\gamma(\bar{u}+d) - \gamma(\bar{u})]/d$ increases with d and $[\gamma(\bar{u}) - \gamma(\bar{u}-d)]/d$ decreases with d.

Thus
$$\begin{aligned} U_{\theta^*}\gamma_\theta(\bar{u}) - \gamma_\theta(\bar{u}) &= \bar{u}_1 q_1 (\gamma_\theta(T_1^*\bar{u}) - \gamma_\theta(\bar{u})) - \bar{u}_2 q_2 (\gamma_\theta(\bar{u}) - \gamma_\theta(T_2^*\bar{u})) \\ &\le \theta^*\bar{u}_1\bar{u}_2 \left(q_1(\gamma_\theta(T_1^*\bar{u}) - \gamma_\theta(\bar{u}))/(T_1^*\bar{u} - \bar{u}) - q_2(\gamma_\theta(\bar{u}) - \gamma_\theta(T_2^*\bar{u}))/(\bar{u} - T_2^*\bar{u}) \right) \\ &\le \theta^*\bar{u}_1\bar{u}_2 \left(q_1(\gamma_\theta(T_1\bar{u}) - \gamma_\theta(\bar{u}))/(T_1\bar{u} - \bar{u}) - q_2(\gamma_\theta(\bar{u}) - \gamma_\theta(T_2\bar{u}))/(\bar{u} - T_2\bar{u}) \right) \\ &= \frac{\theta^*}{\theta}(U_\theta\gamma_\theta - \gamma_\theta) = 0 \end{aligned}$$
//

Corollory. 1.4.6.

The family of 2-action $\theta(\bar{u})$ U.L. rules is at least ϵ-optimal.

Proof.

We have $\lim_{\theta \downarrow 0} \gamma_\theta(\bar{u}) = 0$ and $\lim_{n \to \infty} U^n_{\theta(\bar{u})}\gamma_\theta = \gamma_{\theta(\bar{u})} \le \gamma_\theta$ and we can always find $\theta \ge \theta(\bar{u})$, by compactness and $\theta(\bar{u}) \downarrow 0$ at boundaries. Thus $\lim_{\theta \downarrow 0} \gamma_{\theta(\bar{u})} = 0$ where we suppose $\theta(\bar{u}) = \theta f(\bar{u})$ with θ the learning parameter and $f(\bar{u})$, the learning function. //

It is conjectured that if $\theta^*(\bar{u}) \le \theta^{**}(\bar{u})$ then $\gamma_{\theta^*} \le \gamma_{\theta^{**}}$, but this remains unproven. However, we do achieve monotonicity of $\gamma_{\theta(\bar{u})}$ under quite general conditions.

Lemma 1.4.7.

i) If $\frac{d}{d\bar{u}_1}(T_i\bar{u}_1) > 0$, for i = 1 and 2, then $\gamma_{\theta(\bar{u})}$ is monotone.

ii) If $\exists\ \bar{u}_1$ s.t. $\frac{d}{d\bar{u}_1}(T_i\bar{u}_1) < 0$ for i = 1 or 2, then $\exists\ \phi(\bar{u}_1)$ s.t:-

a) $\phi(\bar{u})$ monotone $\uparrow \not\Rightarrow U\phi(\bar{u})$ monotone $\uparrow$.

b) $\phi(\bar{u})$ non-monotone $\not\Rightarrow U\phi(\bar{u})$ non-monotone.

iii) If $\theta(\bar{u}_1) = \theta(\bar{u}_2)$ then if $\frac{d}{d\bar{u}_1}(T_1\bar{u}_1) > 0\ \forall \bar{u}_1$, $\gamma_{\theta(\bar{u})}$ is monotone .

Proof.

i) We assume $\phi(\bar{u})$ is monotone increasing with $\phi(i) = i$, $i = 0, 1$.

and show that this is preserved by U and hence that $\gamma_{\theta(\bar{\pi})}$ is monotone.

$$U\phi(\bar{\pi}^*) - U\phi(\bar{\pi}) = (\bar{\pi}_1^* p_1 + \bar{\pi}_2^* p_2)\,\phi(\bar{\pi}^*) - (\bar{\pi}_1 p_1 + \bar{\pi}_2 p_2)\,\phi(\bar{\pi}).$$
$$+ \bar{\pi}_1^* q_1\, \phi(T_1\bar{\pi}_1^*) - \bar{\pi}_1 q_1\, \phi(T_1\bar{\pi})$$
$$+ \bar{\pi}_2^* q_2\, \phi(T_2\bar{\pi}_1^*) - \bar{\pi}_2 q_2\, \phi(T_2\bar{\pi}) \qquad \text{where } \bar{\pi}_1^* > \bar{\pi}_1.$$

Now $\quad \frac{d}{d\bar{\pi}_1}(T_i\bar{\pi}_1) > 0 \quad i = 1,2 \implies T_1\bar{\pi}^* > T_1\bar{\pi}$
$\qquad\qquad\qquad\qquad\qquad\qquad T_2\bar{\pi}^* > T_2\bar{\pi}$

Thus $\quad U\phi(\bar{\pi}^*) - U\phi(\bar{\pi}) > (\bar{\pi}_1^* - \bar{\pi}_1)(q_2 - q_1)\,\phi(\bar{\pi}) + q_1(\bar{\pi}_1^* - \bar{\pi}_1)\,\phi(\bar{\pi}) + q_2(\bar{\pi}_2^* - \bar{\pi}_2)\,\phi(T_2\bar{\pi}).$

$$= (\bar{\pi}_1^* - \bar{\pi}_1)\,q_2\,(\phi(\bar{\pi}) - \phi(T_2\bar{\pi})) \geq 0$$ by monotonicity of ϕ .

Hence $\gamma_{\theta(\bar{\pi})}$ is monotone.

ii) I shall just prove that if $T_1\bar{\pi}_1^* < T_1\bar{\pi}_1$ with $\bar{\pi}_1^* > \bar{\pi}_1$ and ϕ is monotone, then $U\phi$ is not always monotone, where $(\bar{\pi}_1^* - \bar{\pi}_1) < \epsilon$ sufficiently small such that $T_1(\bar{\pi}_1 - \bar{\pi}_1^*) > 0$ is possible with $d/d\bar{\pi}_1\,(T_1\bar{\pi}) < 0$.

Define $0 < \phi(T_2\bar{\pi}) = \phi(T_2\bar{\pi}^*) = \phi(\bar{\pi}) = \phi(\bar{\pi}^*) = \phi(T_1\bar{\pi}^*) < \phi(T_1\bar{\pi}) = 1$ with ϕ cts.

and so $\quad U\phi(\bar{\pi}^*) - U\phi(\bar{\pi}) < q_1(\bar{\pi}_1 - \bar{\pi}_1^*)\,\phi(\bar{\pi}) - \bar{\pi}_1 q_1 + \bar{\pi}_1^* q_1\,\phi(\bar{\pi}) = \bar{\pi}_1 q_1\,(\phi(\bar{\pi}) - 1) < 0.$

and hence $U\phi$ is non-monotone , since $U\phi(i) = i,\ i = 0,1$, holds as for ϕ . The other assertion follows similarly, choosing $\phi(\bar{\pi})$ appropriately.

iii) If $d/d\bar{\pi}_1\,(T_1\bar{\pi}_1)\big|_{\bar{\pi}_1^*} > 0$ then $\frac{d}{d\bar{\pi}_1}(T_2\bar{\pi}_1)\big|_{\bar{\pi}_2^*} > 0.$

since $\quad d/d\bar{\pi}_1\,(\bar{\pi}(1-\theta(\bar{\pi})) + \theta(\bar{\pi}))\big|_{\bar{\pi}_1} = 1 - \theta(\bar{\pi}_1) + (1-\bar{\pi}_1)\,d\theta/d\bar{\pi}_1\big|_{\bar{\pi}_1} = 1 - \theta(\bar{\pi}_2) - \bar{\pi}_2\,d\theta/d\bar{\pi}_1\big|_{\bar{\pi}_2}$
$$= \frac{d}{d\bar{\pi}_1}\left(\bar{\pi}(1-\theta(\bar{\pi}))\right)\big|_{\bar{\pi}_2} = d/d\bar{\pi}_1\,(T_2\bar{\pi})\big|_{\bar{\pi}_2}.$$

Hence centrally symmetric rules give monotone $\gamma_{\theta(\bar{\pi})}$ if $d/d\bar{\pi}_1\,(T_1\bar{\pi}_1) > 0$. //

The condition $|d/d\bar{\pi}_1\,(T_1\bar{\pi}_1)| < 1$ is necessary for Norman's (1972) distance-diminishing rules, but this is only satisfied by linear rules, with $d/d\bar{\pi}\,(T_1\bar{\pi}_1) = 1 - \theta$ in our U.L. $\theta(\bar{\pi})$ family. It may be that such a Lifshitz condition is required in order for the method of proving a property P carries across from ϕ to $U\phi$ and hence to γ , to operate.

More subtle techniques may be required to prove the more general comparison theorem.

The comparison theorem 1.4.5. says that the slower we learn, the more likely we are to follow the expected drift. This is the converse situation to gambling in which the optimum stategy is to play boldly and attempt to go against the drift as in Dubins and Savage (1965). In learning we wish (drift/diffusion) = $r(\pi) \uparrow \infty$, whilst in gambling we need $r \downarrow 0$, where $\phi''(\pi) + r(\pi)\phi(\pi) = 0$ gives the diffusion approximation to the process, with $\phi(\pi)$ generating absorption probabilities.

For the optimal rules in the following sections, we attempt to learn very slowly near boundaries, so that we achieve asymtotic reflection from sub-optimal boundaries. We are then absorbed only at the optimal stable boundary, where the drift is with us.

1.5. Boundary Behaviour.

We shall now partition the non-linear U.L. rules $\mathcal{R}$ to obtain those which are actually optimal $\mathcal{R}_0$, by examining the behaviour of $\theta(\pi)$ near to boundaries $\pi_i \in \{0, 1\}$.

<u>Definition</u> 1.5.1.

A rule has boundary behaviour α at $\pi_i = 0$ if $\theta(\pi) \sim O(\pi_i^{\alpha})$ as $\pi_i \downarrow 0$.

I shall only consider rules which can be classified according to their α-dependence, thus eliminating unwanted pathological rules. The following lemmas give strong reasons to conjecture that $\alpha = 1$ is the transitional class occurring between $\alpha > 1$ optimal rules and $0 \le \alpha < 1$ ϵ -optimal rules.

<u>Lemma</u> 1.5.2.

The continuous time deterministic approximation to $\theta(\pi)$ rules gives convergence with exponential behaviour in t iff $\alpha = 0$.

<u>Proof</u>.

We have $\Delta\pi_i = \theta(\pi)\pi_i(1-\pi_i)(q_1 - q_2)$ and w.l.o.g. put $\theta(\pi) = \theta\pi_i^{\alpha}(1-\pi_i)^{\alpha}$.

$= \gamma \bar{\pi}_i^{\beta} (1-\bar{\pi}_i)^{\beta}$ with $\beta = \alpha + 1$, $\gamma = \theta(q_1 - q_2) > 0$.

Then we take deterministically $dx/dt = \gamma x^{\beta}(1-x)^{\beta}$. a)

Now when $x \uparrow 1$ $\dot{x} \sim \gamma(1-x)^{\beta}$ or $\int \frac{dx}{(1-x)^{\beta}} = \int \gamma dt = \gamma t + c$

Hence for $\beta > 1$ $\frac{1}{(1-x)^{\alpha}} = \gamma t + c$ and $x = 1 - (\gamma t + c)^{-1/\alpha}$ b)

Whilst for $\beta = 1$ $\log(1-x) = -(\gamma t + c)$ or $(1-x) = De^{-\gamma t}$, $x = 1 - De^{-\gamma t}$ c)

Now redefine time-scale:- $T = \gamma t + c$ and put $y = 1 - x$

Thus $x = 1 - 1/T^{1/\alpha}$, $\alpha > 0$ $\Big\}$ $yT^{1/\alpha} = 1$ power behaviour

$x = 1 - e^{-T}$, $\alpha = 0$ $\Big\}$ $y = e^{-T}$ exponential.

and in case $\alpha = 1$, we obtain $yT = 1$. //

The linear case is special due to θ =const being distance-diminishing, whilst for $\alpha > 0$ $\lim_{\bar{\pi} \to 0,1} \theta(\bar{\pi}) = 0$ so such a Lifshitz condition cannot be imposed.

<u>Lemma</u> 1.5.3.

The continuous time stochastic diffusion approximation is optimal iff $\alpha > 1$, and is conditionally optimal for $\alpha = 1$.

<u>Proof</u>.

We solve $\phi'' + 2a(\bar{\pi})/b(\bar{\pi})\, \phi'(\bar{\pi}) = 0$ where $a(\bar{\pi})$ = expected drift.
and $b(\bar{\pi})$ = variance.

Then $a(\bar{\pi}) = \theta(\bar{\pi})\, \bar{\pi}(1-\bar{\pi})(q_1 - q_2)$

$b(\bar{\pi}) = q_1 \theta^2(\bar{\pi})\, \bar{\pi}(1-\bar{\pi})^2 + q_2 \theta^2 \bar{\pi}^2 (1-\bar{\pi}) = (q_1(1-\bar{\pi}) + q_2 \bar{\pi})\, \theta^2(\bar{\pi})\, \bar{\pi}(1-\bar{\pi})$.

and for $|q_1 - q_2|$ small put $q_1 + \bar{\pi}_1(q_2 - q_1) = q_1 + o(1)$ and this will not affect the conclusion of optimality for $\alpha > 1$.

Thus $r(\bar{\pi}) = \frac{2a(\bar{\pi})}{b(\bar{\pi})} = \frac{\text{drift}}{\text{diffusion}} \sim 2(q_1 - q_2)/q_1\theta(\bar{\pi}) = 2(1 - q_2/q_1)/\theta(\bar{\pi})$

and for $q_1 > q_2$ (so that $\bar{\pi}_1 = 1$ is optimum) put $(1 - q_2/q_1) = \theta k$
where $\theta(\bar{\pi}) = \theta f(\bar{\pi})$ Thus $\frac{\phi''}{\phi'} = -\frac{2k}{f(\bar{\pi})}$ and for $\alpha = 1$ we put w.l.o.g. $f(\bar{\pi}) = \bar{\pi}(1-\bar{\pi})$.

and $\log \phi' = \int \frac{-2k}{\pi(1-\pi)} = -2k \log\left(\frac{\pi}{1-\pi}\right) + C = \log\left(\frac{1-\pi}{\pi}\right)^{2k} + C$

$\phi' = A\left(\frac{1-\pi}{\pi}\right)^{2k}$ and $\phi(i) = i \quad i = 0, 1.$

Then for $\alpha = 1$ at both boundaries $\phi(x) = \int_0^x \left(\frac{1-\pi}{\pi}\right)^{2k} d\pi \Big/ \int_0^1 \left(\frac{1-\pi}{\pi}\right)^{2k} d\pi$

Clearly when $2k \geq 1$, $\phi(x) \equiv 1$ for $x \neq 0$ and diffusion is optimal since $\int_0^x 1/\pi^{2k}\, d\pi$ is divergent iff $2k \geq 1$.

Whilst $2k < 1$ gives $\phi(x) < 1$ for $x < 1$ and ϕ is monotone $\uparrow$; we have ϵ-optimality.

(Note: formal bounds for $\alpha = 1$ will be proved in 1.5.17.)

Now since $\int_0^x 1/\pi^\alpha\, d\pi$ is divergent for $\alpha > 1$, we can easily verify $\phi(x) \equiv 1$ for $x \neq 0$ here also. Similarly $\int_0^x 1/\pi^\alpha\, d\pi$ is convergent for $\alpha < 1$, giving ϵ-optimality.

Where $\phi(x) = \left[\int_0^x \exp \int_{y_0}^y -2k/f(\pi)\, d\pi\, dy \Big/ \int_0^1 \exp \int_{y_0}^y -2k/f(\pi)\, d\pi\, dy\right]$ is easily found.

Finally, we find this diffusion limit for $\alpha = 0$.

$$\phi''/\phi' = -2k \qquad \log \phi' = -2kx + C \qquad \phi' = Ae^{-2kx}, \qquad \phi = Be^{-2kx} + C.$$

And boundary conditions give $\phi(x) = (1 - e^{-2kx}) / (1 - e^{-2k})$ //

Norman (1971) proves that the discrete time $\gamma_\theta(\pi)$ converges weakly to $\phi(x)$ for this linear rule. The theorem requires $r(\pi)$ bounded and hence breaks down for $\alpha > 0$, when $r(\pi) \uparrow \infty$ at boundaries. Vorontsova (1965) proved a result similar to 1.5.3. for a family of rules with penalty and reward reinforcement. However, for $\alpha < 1$, such rules cannot be normalised to give uniform learning and thus our reward-inaction scheme appears the more fundamental. I shall now state a result of Norman (1968) which gives bounds on the discrete time absorption probabilities for the linear rule.

Lemma 1.5.4.

i) Let $\phi_{z,\theta}(\bar{\pi}) = (1-e^{-z\bar{\pi}/\theta})/(1-e^{-z/\theta})$ then $\exists$ +ve y and z such that $\phi_{y,\theta} \le \gamma_\theta(\bar{\pi}) \le \phi_{z,\theta}$ $\forall \theta, \bar{\pi} \in [0,1]$

ii) $\gamma_\theta(\bar{\pi}) \geq (1-e^{-z\bar{\pi}/\theta})$ where $q_1 > q_2$.

Proof.

Norman (1968), proves $\gamma_{x,\theta}(\bar{\pi}) = e^{x\bar{\pi}/\theta}$ can be super or sub-regular depending on x, and the result i) then follows easily by noting that the class of super and sub-regular functions are closed under addition and multiplication by +ve constants.

For ii) we have $(1-e^{-z/\theta}) < 1$ and hence also $\lim_{\theta \downarrow 0} \gamma_\theta(\bar{\pi}) = 1$ giving ϵ-optimality.

//

In the literature, such as Sawaragi (1974), we find that the same method is applied to rules of classes $\alpha > 0$. However, we may observe that Norman's $\phi_{z,\theta}(\bar{\pi})$ arises precisely form the weak convergence limit of 1.5.3, which only has exponential behaviour for $\alpha = 0$, as in the deterministic solution 1.5.2. For $\alpha > 0$, we should only expect to be able to put tight bounds on $\gamma_{\theta(\bar{\pi})}(\bar{\pi})$ by using the weak convergence limit $\phi(x)$ appropriate to it. The use of the $\alpha = 0$ limit for $\alpha > 0$ rules will only give us ϵ-optimality, which has already been proved for all $\alpha \geq 0$ in theorem 1.4.5.

Lemma 1.5.5.

The $\lim_{n \to \infty}$ Pr(attain $\bar{\pi}_1 \geq \frac{1}{n}$ before absorption at $\bar{\pi}_1 = 0$, $\bar{\pi}_1(0) = \frac{1}{2n}$, using $\theta(\bar{\pi})) = 1$ when $\alpha > 1$ and $q_1 > q_2$.

Or, concisely we write, $\lim_{n \to \infty} Pr_{\theta(\bar{\pi})}(1/n\, BO \mid \tfrac{1}{2n}) = 1.$

Proof.

$$Pr_{\theta(\bar{\pi})}(1/n\, BO \mid \tfrac{1}{2n}) \geq P_{\theta'(\bar{\pi})}(1BO \mid \tfrac{1}{2n})$$ with $\theta'(\bar{\pi}) = \theta(\bar{\pi})$ on $\bar{\pi} \in [0, \frac{1}{n})$, $\theta'(\bar{\pi}) = \theta(\frac{1}{n})$ on $\bar{\pi} \in [\frac{1}{n}, 1]$.

But $\gamma_{\theta > 1/n}(\bar{\pi}) \le \gamma_{\theta'(\bar{\pi})}(\bar{\pi})$ by 1.4.5. as $\theta'(\bar{\pi}) \le \theta(\frac{1}{n})$ will hold w.l.o.g. (If $\theta(\bar{\pi})$ oscillates near $\bar{\pi} = \frac{1}{n}$ we just bound $\theta(\bar{\pi})$ suitably above by some θ =const)

Then by 1.5.4. $\gamma_{\theta(1/n)}(\bar{\pi}) \geq 1 - e^{-z\bar{\pi}/\theta(1/n)}$ for +ve z $\forall \bar{\pi}, n$.

Thus $P_{\theta(\bar{\pi})}(1/n\, BO \,|\, \tfrac{1}{2n}) \geq 1 - e^{-z\bar{\pi}/\theta(1/n)}$ now put $\bar{\pi} = \frac{1}{2n}$

and $\theta(1/n) = \theta(1/n)^{\alpha}(1 - 1/n)^{\alpha} < \theta/n^{+\alpha}$ and hence,

$$P_{\theta(\bar{\pi})}(1/n\, BO \,|\, 1/2n) \geq 1 - e^{-z/2n \cdot 1/\theta n^{-\alpha}} = 1 - \exp\left(-\frac{z}{2\theta} n^{\alpha-1}\right).$$

and $\lim_{n\uparrow\infty} \exp\left(-\frac{z}{2\theta} n^{(\alpha-1)}\right) = 0$ when $\alpha > 1$, and hence result. //

This lemma shows why we may expect asymtotic reflection from sub-optimal boundaries. For $\alpha < 1$, using the diffusion limit and weak convergence for $\alpha = 0$ bounds, we can similarly show $P_{\theta(0)}(1/n\, BO \,|\, 1/2n) = 0$ giving absorption at sub-optimal boundaries. However, for $\alpha = 1$, we can obtain no such bounds, for non-rigorously, we can put $\alpha = 1$ in $\gamma_{\theta(1/n)}(\bar{\pi})$ to obtain only $P_{\theta(\bar{\pi})}(1/n\, BO \,|\, 1/2n) \geq 1 - e^{-z/2\theta}$ = const, which demonstrates how difficult this transition case really is to understand.

Definition. 1.5.6.

If at time t we choose action i and at time t+1 we choose action $j \neq i$, then we have an alternation.

The following lemma shows how the alternation concept is closely linked to that of optimality. However, only if we exclude the transition case $\alpha = 1$, is # alternations $= \infty$ nec and suff for optimality. For the transition rules, we shall prove this to be neither nec nor suff for convergence to optimum action, since 1.5.17. will show $\alpha = 1$ to be conditionally optimal as in the diffusion.

For the next lemma we consider boundary behaviour as $\bar{\pi}_i \uparrow 1$. We let $\theta(\bar{\pi}_i) \sim O(1 - \bar{\pi}_i)^{\alpha}$ as $\bar{\pi}_i \uparrow 1$ which determines absorption behaviour whilst 1.5.1 effectively defines reflection behaviour. However, after 1.5.7. all boundaries will be assumed to have the same α-absorption and α-reflection behaviour, so the distinction is only of mathematical interest.

Lemma 1.5.7.

At boundary i, $\theta(\bar{\pi}_i) \sim O((1-\bar{\pi}_i)^{\alpha})$ as $\bar{\pi}_i \uparrow 1$.

i) If $\alpha < 1$ and $\lim_{t \to \infty} \bar{\pi}_i(t) = 1$ then $\text{Halts} \overset{a.s.}{<} \infty$ and if $\alpha < 1$ at all boundaries, this is strengthened to $\mathcal{E}\,\text{Halts} < \infty$.(And indeed all moments are finite.)

ii) If $\alpha > 1$ and $\lim_{t \to \infty} \bar{\pi}_i(t) = 1$ then $\text{Halts} \overset{a.s.}{=} \infty$.

iii) If $\alpha = 1$ and $\lim_{t \to \infty} \bar{\pi}_i(t) = 1$ then $\text{Halts} \overset{a.s.}{=} \infty$ iff $\theta q_i \leq 1$, where q_i = reward probability,as usual.

Proof.

We define $i_t(\bar{\pi}_j)$ =Prob(We always choose action j $| \ \bar{\pi}_j(t) = \bar{\pi}_j$).

Now we immediately have the basic recurrence equation,

$$i_t(\bar{\pi}_j) = \bar{\pi}_j \left(p_j i_t(\bar{\pi}_j) + q_j i_t(\bar{\pi}_j + \theta(\bar{\pi})(1-\bar{\pi}_j)) \right). \qquad 1)$$

and hence $i_t(\bar{\pi}_j) = \left(1 - (1-\bar{\pi}_j)/(1-p_j\bar{\pi}_j)\right) i_t(\bar{\pi}_j + \theta(\bar{\pi})(1-\bar{\pi}_j))$ with $p_j = 1 - q_j$

I shall now omit suffixes j and t and write $\epsilon(\bar{\pi}) = (1-\bar{\pi})/(1-p\bar{\pi})$ and $T_1\bar{\pi} = \bar{\pi} + \theta(\bar{\pi})(1-\bar{\pi})$.

Thus $i(\bar{\pi}) = (1-\epsilon(\bar{\pi}))\, i(T_1\bar{\pi}) = \prod_{n=0}^{\infty} (1-\epsilon(T^n\bar{\pi}))$.

where $\lim_{n \to \infty} i(T^n\bar{\pi}) = 1$ since $\theta(\bar{\pi})(1-\bar{\pi}) = 0$ iff $\bar{\pi} = 0, 1$.

Now $(1-\bar{\pi})/(1-p\bar{\pi}) \sim 1/(1-p) \left((1-\bar{\pi}) - p/q (1-\bar{\pi})^2 + \ldots \right)$ and hence as $\bar{\pi}_i \uparrow 1$ we can neglect $o(1-\bar{\pi})$ terms, except in the case $\alpha = 1$ when we must be much more careful.

Note that $1-\epsilon(\bar{\pi}) = q\bar{\pi}/(1-p\bar{\pi})$, which is a transformed distribution function.

Now for $\alpha \neq 1$, we have $i(\bar{\pi}) > 0$ iff $\sum_{n=0}^{\infty} (1-\bar{\pi}_n) < \infty$

We prove convergence or divergence for α by a comparison test with $\sum 1/n^{\beta}$, β chosen appropriately, and ii) follows immediately from i). We use the condition $i(\bar{\pi}) > 0$ iff $\sum_{n=0}^{\infty} \epsilon(T^n\bar{\pi}) < \infty$.

i) a) Put $\bar{\pi}_n = 1 - 1/n^{\beta}$, for some fixed $n >> 1$ and so $\theta(\bar{\pi}) \sim \theta(1-\bar{\pi})^{\alpha}$

Define $b_n = 1 - \bar{\pi}_n = 1/n^{\beta}$ and so $b_{n+1} = 1 - \bar{\pi}_{n+1} = 1/n^{\beta} - \theta/n^{(\alpha+1)\beta}$.

$$b_{n+1} = 1/n^{\beta}\left(1 - \theta/n^{\alpha\beta}\right).$$

We take the ratio $b_n/b_{n+1} \sim 1 + \theta/n^{\alpha\beta}$ whilst $a_n/a_{n+1} = \frac{1/n^\beta}{1/(n+1)^\beta} \sim 1 + \beta/n$

But $\sum 1/n^\beta < \infty$ for $\beta > 1$ and if $\alpha\beta < 1$, $b_n/b_{n+1} > a_n/a_{n+1}$ for some $n > n_0$.

Hence, since the above holds for any $n \gg 1$ and we can always find $\beta > 1$ s.t. $\alpha\beta < 1$ when $\alpha < 1$, we have(by Gauss Ratio Test,)

$\alpha < 1 \Rightarrow \sum_{n=1}^{\infty} (1 - \bar{\pi}_n) < \infty$ and hence $i(\bar{\pi}) > 0$

Now since the process is markovian and $i(\bar{\pi}) \uparrow 1$ as $\bar{\pi} \uparrow 1$, we must a.s. have t_0 s.t. for $t > t_0$, we take only action i and $\#alts \overset{a.s}{<} \infty$.

b) Now we have to prove $\mathcal{E}\#alt \overset{a.s}{<} \infty$ for $\alpha < 1$ at all boundaries, and indeed $\mathcal{E}(\#alt)^n < \infty$.

Prob(at least 1 more alternation starting from next choice) $< 1 - \epsilon$

where $\min_{\bar{\pi}} \mathcal{E}_{i}(i(\bar{\pi}_i)) > \epsilon > 0$ which exists by continuity of $i(\bar{\pi}_i)$ and $\lim_{\bar{\pi}_i \uparrow 1} i(\bar{\pi}_i) = 1$.

Now take successive alternations, starting process again at $\bar{\pi}_{n+1}$ if r^{th} alternation occurs at $\bar{\pi}_{n-1}, \bar{\pi}_n$.

Then $Pr(\#alt > 2r) < (1-\epsilon)^r/\epsilon$ (allowing possibility of alt at n, n+1 trials)

Clearly also $Pr(\#alt > 2r \text{ and } \#alt < 4r) < (1-\epsilon)^r/\epsilon$.

Then $\mathcal{E}\#alt = \sum k p_k < 3 + \sum_{s=0}^{\infty} 2^{s+2} (1-\epsilon)^{2^s}/\epsilon < \infty$ by comparison with G.P.

since $Pr(2^s < \#alt < 2^{s+1}) < (1-\epsilon)^{2^{s-1}}/\epsilon$.

Also given $\delta > 0, \exists n_0$ s.t. $Pr(\#alt > n_0) < \delta$, follows by above, and hence we prove rigorously the last statement in a).

Also $\mathcal{E}(\#alt)^n = \sum k^n p_k < C + \sum_{s=0}^{\infty} 2^{n(s+2)} (1-\epsilon)^{2^s}/\epsilon < \infty$ by comparison with G.P. where $C < 2^n + 1$.

ii) For $\alpha > 1$, we find $\beta : \alpha\beta > 1$ with $\sum 1/n^\beta = \infty$, which is always possible, and hence $\sum_{n=1}^{\infty} (1 - \bar{\pi}_n) = \infty$ and $i(\bar{\pi}) = 0 \Rightarrow \#alt \overset{a.s}{=} \infty$.

iii) We put $(1 - \bar{\pi}_t) = 1/x$ giving $(1 - \bar{\pi}_t)/(1 - p\bar{\pi}_t) = 1/(p + qx) = 1/n$.

Note that in this lemma π_t refers to action i at trial t, since we are considering boundary i throughout.

Define $a_t = (1-\pi_t)/(1-p\pi_t)$ Then $i(\pi)=0$ iff $\sum_t a_t = \infty$.

We prove divergence using the ratio test; comparison with $\sum 1/n$.

We just compare the tail of a_t with $b_n = 1/n$ and by integral or ratio test; tail above $y = 1/n \Rightarrow$ divergent

and tail strictly below $y = 1/n \Rightarrow$ convergent

Now
$$(1-\pi_{t+1}) / (1-p\pi_{t+1}) = (1-\pi_t)(1-\theta(\pi_t)) / (1-p(\pi_t + (1-\pi_t)\theta(\pi_t)))$$
$$= (1-\theta(\pi_t)) / (p + q/(1-\pi_t) - p\theta(\pi_t)).$$
$$= (1-\theta(\pi_t)) / (n - p\theta(\pi_t)).$$

Thus $a_t / a_{t+1} = (1 - 1/n\, p\theta(\pi_t)) / (1-\theta(\pi_t))$

But we have $\theta(\pi_t) = O(1-\pi_t)$ as $\pi_t \uparrow 1$ and so let $\theta(\pi_t) = \theta(1-\pi_t)$.

Then $\theta(\pi_t) = \theta q / n(1-p/n)$ and $a_t/a_{t+1} = 1 + \theta q/n + O(1/n^2)$ which gives divergence when $\theta q \leq 1$

Thus $i(\pi_i) > 0$ iff $\theta q_i > 1$ and iff $\#alt < \infty$.

//

Remarks 1.5.8

a) For the linear rule it is actually possible to express alternations in terms of $\gamma(\pi)$ as in Norman (1968). We shall give an easier proof here.

Lemma 1.5.9.

For the linear rule $\chi(\pi) = (2-\theta(q_1+q_2))(\pi - \gamma(\pi)) / \theta(q_2-q_1)$.

and $\chi(\pi) = 2(1-\theta q)\pi(1-\pi)/\theta^2 q$ when $q_1 = q_2$

$\chi(\pi) = \mathcal{E}\,\#$ alternations in response, $\pi(0) = \pi$

Proof.

$\Delta\chi(\pi) = -\pi_1\pi_2(2-\theta(q_1+q_2))$ is easily found.

and $\Delta\pi = \theta(q_1 - q_2)\pi_1\pi_2$ enables us to substitute when $q_1 \neq q_2$.

and we have $\Delta\gamma(\bar{\Pi})=0$ and $\gamma(\bar{\Pi})$ is the unique fixed function satisfying the given boundary conditions.

So $\chi(\bar{\Pi}) = \left(2-\theta(q_1+q_2)\right) / \theta(q_2-q_1) \left(\bar{\Pi} - c\gamma(\bar{\Pi})\right)$ where c = 1 from boundary values.

If $q_1 = q_2$ then we find $\Delta \bar{\Pi}_1\bar{\Pi}_2 = -\theta^2 q \bar{\Pi}_1\bar{\Pi}_2$

and $\Delta\chi = -2\bar{\Pi}_1\bar{\Pi}_2(1-\theta q)$

So again we just use uniqueness of $\gamma(\bar{\Pi})$ for result.

//

b) This method only works for θ =const, else we get non-linearities. However, if $\bar{\Pi} \sim 1/n$, $\theta(\bar{\Pi}) \sim \theta(1/n)^{\alpha}$ then $\chi \sim n^{(\alpha-1)}$ on putting $\gamma(\bar{\Pi}) \sim 0$, $q_2 > q_1$. Thus as $n \to \infty$ $\chi = \infty$, $\alpha > 1$ whilst at the transition case $\alpha = 1$,

$\chi < \infty$ $\alpha < 1$

we get $\chi \sim 2/\theta(q_2-q_1)$ which gives no information. This is a non-rigorous argument and 1.5.7. is still required to give tight conditions on $\theta(\bar{\Pi})$ in a fully rigorous manner, and to deal with the subtle case of $\alpha = 1$.

Lemma 1.5.10.

$i(\bar{\Pi}_1)\uparrow$ when $d/d\bar{\Pi}\,(T_1\bar{\Pi}) \geq 0$ $\forall \bar{\Pi} \in [0,1]$.

Proof.

We have $i(\bar{\Pi}_1) = \left(1 - \bar{\Pi}_2/(1-p\bar{\Pi}_1)\right) i(T_1\bar{\Pi})$

Thus if $\bar{\Pi}_1' > \bar{\Pi}_1$, $i(\bar{\Pi}_1') - i(\bar{\Pi}_1) \geq \left(i(T_1\bar{\Pi}_1') - i(T_1\bar{\Pi}_1)\right)\left(1 - \bar{\Pi}_2/(1-p\bar{\Pi}_1)\right)$.

and $d/d\bar{\Pi}\,(T_1\bar{\Pi}) \geq 0 \Rightarrow T_1\bar{\Pi}' > T_1\bar{\Pi}$.

Now iterate to get $i(\bar{\Pi}_1') - i(\bar{\Pi}_1) \geq \left(i(T_1^n\bar{\Pi}_1') - i(T_1^n\bar{\Pi}_1)\right) \prod_{r=0}^{n-1} \left(1 - \frac{T_1^r\bar{\Pi}_2}{(1-pT_1^r\bar{\Pi}_1)}\right) \geq 0$

Hence $i(\bar{\Pi}_1)\uparrow$

//

In future sections, we shall consider rules with the same α - behaviour at all boundaries, so that the two definitions we have used become equivalent.

For the transition case iii) 1.5.7., we see that it is possible for $\theta q_2 < 1 < \theta q_1$ and yet in 1.5.17.we shall prove that the rule may still be either optimal or ϵ-optimal. Thus we may have an ϵ-opt rule with $\#alt = \infty$ with finite probability, or an opt rule with $\#alt \overset{a.s.}{<} \infty$. If $\theta q_2 < \theta q_1 < 1$, then we shall also find ϵ-opt rules with $\#alt \overset{a.s.}{=} \infty$.

Corollory 1.5.11.

If $i(\Pi_i) > 0$ and $\exists_j$ s.t. $q_j > q_i$ then the rule is at best ϵ-optimal, and we say it is in class R_ϵ.

Proof.

By comparison theorem 1.4.5, all U.L.$\theta(\Pi)$ rules are at worst ϵ-opt, but $i(\Pi_i) > 0$ with finite probability we only take sub-optimal action i. Hence all rules with $\alpha < 1$, at any boundary, (since we do not know which is optimum) are in R_ϵ, and also rules with $\alpha = 1,\ q_i\theta > 1$ at boundary i. //

We must now consider those rules for which $i(\Pi_i) \equiv 0$ at every sub-optimal boundary; $q_{opt} > q_i \ \forall i \neq opt$. First we prove that the class of optimal rules R_1 includes those with $\alpha > 1$ at every boundary. Then for $\alpha = 1$, we delineate regions which give conditional optimality depending on θ and $|q_1 - q_2|$, and other regions which are at best, ϵ-optimal.

Although the class of rules $\alpha = 1$,with its conditional optimality (1.5.17) would not be of great practical use, its study does help us to understand the behaviour of $\alpha \neq 1$ more fully, as ϵ-opt "fades" into optimality. Both have fundamentally different workings as discussed in section 7.

Theorem 1.5.12.

i) The class of rules with $\alpha > 1$ at each boundary is optimal.

ii) This optimality is independent of the behaviour of $\theta(\Pi)$ in finitely many compact subsets of I $= [0,1]$

Proof.

We construct a sub-regular family $\rho_{k\delta}(x)$, guided in choice by 1.5.3.

Define $\rho_{k\delta}(x) = \psi_{k\delta}(x) / \psi_{k\delta}(1)$ where $\psi_{k\delta}(x) = \int_0^\delta f_k(\delta)\,d\bar{u} + \int_\delta^x f_k(\bar{u})\,d\bar{u} \qquad x \geq \delta$

$\qquad = \int_0^x f_k(\delta)\,d\bar{u} \qquad 0 \leq x < \delta.$

and where

$$f_k(\bar{u}) = \exp\left[k / [\bar{u}(1-\bar{u})]^{\alpha-1} \right] \qquad \bar{u} \in [0, \tfrac{1}{4}]$$
$$f_k(\bar{u}) = \exp\left[2k(1-2\bar{u}) / (\bar{u}(1-\bar{u}))^{\alpha-1} \right] \qquad \bar{u} \in [\tfrac{1}{4}, \tfrac{3}{4}]$$
$$f_k(\bar{u}) = \exp\left[-k / [\bar{u}(1-\bar{u})]^{\alpha-1} \right] \qquad \bar{u} \in [\tfrac{3}{4}, 1].$$

Our $f_k(\bar{u})$ is chosen, since at the boundaries $\bar{u} = 0, 1$, it approximates the integrand in $\phi(x) = \int_0^x \exp \int_{y_0}^{y} -2k\theta/_{\theta(\bar{u})}\, d\bar{u}\, dy \sim \int_0^x f_k(\bar{u})\, d\bar{u}.$

We truncate the integral at $x = \delta$, to remove the divergence, and as $\delta \downarrow 0$, we obtain $\rho_{k\delta}(x) \uparrow \gamma_{\theta(\bar{u})}(\bar{u}) \equiv 1$ on $\bar{u} \neq 0$. Our task is to prove, for $\alpha > 1$, $U_{\theta(\bar{u})}\rho_{k\delta}(\bar{u}) - \rho_{k\delta}(\bar{u}) \geq 0$, $\bar{u} \in [0,1]$, and hence $\gamma_{\theta(\bar{u})}(\bar{u}) \geq \rho_{k\delta}(\bar{u}) \quad \forall \underline{\bar{u}}, \delta.$ (for some $k > k^* > 0$ for every δ).

Lemma 1.5.13

$\psi(x) = \int_0^x d\bar{u}$ is sub-regular for all $\theta(\bar{u})$, when $q_1 \geq q_2$.

Proof.

$$U\psi - \psi = q_1\bar{u}_1 \int_{\bar{u}_1}^{T_1\bar{u}_1} d\bar{u} - q_2\bar{u}_2 \int_{T_2\bar{u}_1}^{\bar{u}_1} d\bar{u} = \theta(\bar{u})\,\bar{u}_1\bar{u}_2\,(q_1 - q_2) \geq 0.$$

//

This holds since $\gamma_{\theta(\bar{u})}(\bar{u}) = \bar{u}$ is the solution to $U_{\theta(\bar{u})}\gamma_{\theta(\bar{u})} = \gamma_{\theta(\bar{u})}$ when $q_1 = q_2$.

Lemma 1.5.14

$\exists k > 0$ s.t. $D_k(\underline{\bar{u}}) = q_1\bar{u}_1 \int_{\bar{u}_1}^{T_1\bar{u}_1} f_k(\bar{u})\,d\bar{u} - q_2\bar{u}_2 \int_{T_2\bar{u}_1}^{\bar{u}_1} f_k(\bar{u})\,d\bar{u} > 0$

Proof.

Here we formally prove $\rho_{k\delta}(x)$ is sub-regular, and most of the theorem arises from this lemma. We apply the δ-truncation later so that we may write $U_{\theta(\bar{u})}\psi_{k\delta}(\bar{u}) - \psi_{k\delta}(\bar{u}) = q_1\bar{u}_1 \int_{\bar{u}_1}^{T_1\bar{u}_1} f_k\,d\bar{u} - q_2\bar{u}_2 \int_{T_2\bar{u}_1}^{\bar{u}_1} f_k\,d\bar{u}$ when $T_2\bar{u}_1 > \delta$

We partition $I = [0,1]$ into $I_1 = [0, \epsilon)$, $I_2 = [\epsilon, 1-\epsilon^*]$ and $I_3 = (1-\epsilon^*, 1]$.

i) Take $I_2 = [\epsilon, 1-\epsilon^*]$ a compact subset of $[0,1]$.

Now in $\lim_{k\downarrow 0} f_k(\underline{\Pi}) \to 1$ and $\lim_{k\downarrow 0} q_1\bar{\Pi}_1 \int_{\bar{\Pi}_1}^{T_1\bar{\Pi}_1} f_k d\bar{\Pi}_1 - q_2\bar{\Pi}_2 \int_{T_2\bar{\Pi}_1}^{\bar{\Pi}_1} f_k d\bar{\Pi}_1 = q_1 - q_2 > 0$

By compactness, we can use a finite open covering from any ∞ open covering. So choose intervals S_i s.t. $\exists\, k_i$ and $D_{k_i}(\underline{\Pi})$ on S_i; then take a finite covering of I_2 and put $k^* = \min_i k_i > 0$ for i in our open cover.

ii) Now take $I_1 = [0,\epsilon)$ with $\epsilon < 1/4$.

$$D_k(\underline{\Pi}) > q_1\bar{\Pi}_1 (T_1\bar{\Pi}_1 - \bar{\Pi}_1) f_k(T_1\bar{\Pi}_1) - q_2\bar{\Pi}_2(\bar{\Pi}_1 - T_2\bar{\Pi}_1) f_k(T_2\bar{\Pi}_1).$$

$$= \left(q_1 \exp\left[k/(T_1\bar{\Pi}_1 . T_1\bar{\Pi}_2)^{\alpha-1}\right] - q_2 \exp\left[k/(T_2\bar{\Pi}_1 . T_2\bar{\Pi}_2)^{\alpha-1}\right]\right) \theta(\underline{\Pi})\,\bar{\Pi}_1\bar{\Pi}_2.$$

$$= \theta(\underline{\Pi})\bar{\Pi}_1\bar{\Pi}_2 \Big(q_1 \exp\left[k/\bar{\Pi}_1^{\alpha-1}\bar{\Pi}_2^{\alpha-1}(1-\theta\bar{\Pi}_1^{\alpha}\bar{\Pi}_2^{\alpha})^{\alpha-1}(1+\theta\bar{\Pi}_1^{\alpha-1}\bar{\Pi}_2^{\alpha+1})^{\alpha-1}\right]$$

$$- q_2 \exp\left[k/\bar{\Pi}_1^{\alpha-1}\bar{\Pi}_2^{\alpha-1}(1+\theta\bar{\Pi}_1^{\alpha+1}\bar{\Pi}_2^{\alpha-1})^{\alpha-1}(1-\theta\bar{\Pi}_1^{\alpha}\bar{\Pi}_2^{\alpha})^{\alpha-1}\right].$$

Thus $D_k(\underline{\Pi}) > 0$ iff $q_1/q_2 > \exp\left[\left(k/\bar{\Pi}_1^{\alpha-1}\bar{\Pi}_2^{\alpha-1}(1-\theta\bar{\Pi}_1^{\alpha}\bar{\Pi}_2^{\alpha})\right) \cdot B(\underline{\Pi})\right].$

where $B(\underline{\Pi}) = (1+\theta\bar{\Pi}_1^{\alpha+1}\bar{\Pi}_2^{\alpha-1})^{1-\alpha} - (1+\theta\bar{\Pi}_1^{\alpha-1}\bar{\Pi}_2^{\alpha+1})^{1-\alpha}.$

Then put $x = (\bar{\Pi}_1\bar{\Pi}_2)^{\alpha-1}$ to obtain $B(\underline{\Pi}) = (1+\theta\bar{\Pi}_1^2 x)^{1-\alpha} - (1+\theta\bar{\Pi}_2^2 x)^{1-\alpha}.$

$$B(\underline{\Pi}) = \theta x(\bar{\Pi}_2 - \bar{\Pi}_1)(\alpha-1) - \theta^2 x^2/_2\, \alpha(\alpha-1)(\bar{\Pi}_2^4 - \bar{\Pi}_1^4) + \ldots\ldots$$

$$> \theta x(\bar{\Pi}_2 - \bar{\Pi}_1)(\alpha-1) - \theta^2 x^2 \alpha(\alpha-1)(\bar{\Pi}_2^4)/_2.$$

by some easy analysis, where $\theta(\underline{\Pi}) = \theta\bar{\Pi}_1^{\alpha}\bar{\Pi}_2^{\alpha}$ w.l.o.g.

Thus $D_k(\underline{\Pi}) > 0$ if $k\theta(\alpha-1) < (1-\theta\bar{\Pi}_1^{\alpha}\bar{\Pi}_2^{\alpha}) / ((\bar{\Pi}_2 - \bar{\Pi}_1) - \theta x\alpha\bar{\Pi}_2^4/_2) \cdot \log(q_1/q_2)$

or $k < g(\bar{\Pi}_1)/_{\theta(\alpha-1)} \log(q_1/q_2)$ for $\bar{\Pi}_1 \in [0,\epsilon)$

But r.h.s. > 0 when $q_1 > q_2$, $\alpha > 1$ and is strictly bounded away from zero. Also in $\lim_{\bar{\Pi}_1\downarrow 0} k < 1/\theta(\alpha-1) \cdot \log(q_1/q_2).$ is resulting constraint.

Hence we can certainly take $k^{**} = \min_{\bar{\Pi}_1} g(\bar{\Pi}_1)/_{\theta(\alpha-1)} \log(q_1/q_2)$ for ϵ sufficiently small.

In fact if we put $\theta(\bar{\Pi}) = \theta \bar{\Pi}_1^{\alpha^*} \bar{\Pi}_2^{\alpha^*}$, $\alpha^* > \alpha$ then we have the far weaker constraint $q_1/q_2 > \exp[k \bar{\Pi}_1^{\alpha^*-\alpha} \bar{\Pi}_2^{\alpha^*-\alpha} \theta(\alpha-1) g(\bar{\Pi}_1)]$ on k and so we can find $k>0$ s.t. $D_k(\bar{\Pi})>0$ on I_1 for $\alpha^*>\alpha$.

iii) Now we do the same calculations for I_3 and by the symmetry of $\bar{\Pi}_1, \bar{\Pi}_2$ we find $k' = \min_{\bar{\Pi}_2} g(\bar{\Pi}_2)/\theta(\alpha-1) \log(q_1/q_2) > 0$ for ϵ^* sufficiently small.

iv) Now take $k^s = \min(k', k^*, k^{**})$ to give $D_{k^s}(\bar{\Pi}) \geq 0$ on I. //

Remarks. 1.5.15.

a) We only required the learning function for I_1 and I_3; the compactness argument is independent of $\theta(\bar{\Pi})$.

b) I shall now show semi-rigorously why we achieve $k>0$ in ii) and iii). Let $\bar{\Pi}_1 \sim 0$ so $\theta(\bar{\Pi}) \sim \theta \bar{\Pi}_1^{\alpha}$

$$D_k(\bar{\Pi})>0 \quad \text{iff} \quad q_1 \exp\left(\frac{k}{(T_1\bar{\Pi}_1)^{\alpha-1}}\right) - q_2 \exp\left(\frac{k}{(T_2\bar{\Pi}_1)^{\alpha-1}}\right) > 0$$

$$\text{or} \quad \exp k\left[\left((T_1\bar{\Pi}_1)^{\alpha-1} - (T_2\bar{\Pi}_1)^{\alpha-1}\right)/(T_1\bar{\Pi}_1 \cdot T_2\bar{\Pi}_1)^{\alpha-1}\right] < q_1/q_2 .$$

and as $(T_1\bar{\Pi}_1)^n > (T_2\bar{\Pi}_1)^n$ we have $0 < k < g(\bar{\Pi}_1) \log(q_1/q_2)$

e.g. $\alpha=2$ then $(T_1-T_2)\bar{\Pi}_1 = \theta(\bar{\Pi}) = \theta\bar{\Pi}_1^2\bar{\Pi}_2^2 \sim \theta\bar{\Pi}_1^2$ and $(T_1\bar{\Pi}_1)(T_2\bar{\Pi}_1) \sim \bar{\Pi}_1^2$.

$$\text{Then} \quad \exp(k\theta) < q_1/q_2 \Rightarrow k < \tfrac{1}{\theta} \log q_1/q_2 .$$

c) The function $D_k(\bar{\Pi}) \not> 0$ for $\alpha^* < \alpha$ ..

Lemma 1.5.16.

The family $\varrho_{k\delta}(\bar{\Pi})$ are sub-regular for k, δ , sufficiently small. Given $\theta(\bar{\Pi})$ $\exists\, k^*$ s.t. for $0<k<k^*$ and $0<\delta<\delta(k^*)$ we have $U_{\theta(\bar{\Pi})} \varrho_{k\delta}(\bar{\Pi}) \geq \varrho_{k\delta}(\bar{\Pi})$

Proof.

We use 1.5.14. and note $\varrho_{k\delta}(\bar{\Pi})$ is sub-regular for $T_1\bar{\Pi}_1 < \delta$ by 1.5.13. We must be careful of the transition at $\bar{\Pi}_1 = \delta$.

a) $T_1\bar{\Pi}_1 > \delta$, $\bar{\Pi}_1 < \delta$ b) $\bar{\Pi}_1 > \delta$, $T_2\bar{\Pi}_1 > \delta$.

Then $\displaystyle\int_{\bar{\Pi}_1}^{T_1\bar{\Pi}_1} f_k(\bar{\Pi}_1)\, d\bar{\Pi}_1 > (T_1\bar{\Pi}_1 - \bar{\Pi}_1) f_k(T_1\bar{\Pi}_1)$.

and $\int_{T_2\bar{\pi}_1}^{\bar{\pi}_1} f_k(\bar{\pi}) < (\bar{\pi}_1 - T_2\bar{\pi}_1)\, f_k(T_2\bar{\pi}_1)$ in both a) and b)

So just apply 1.5.14. ii) which holds for ϵ and hence δ, sufficiently small. Finally, if $T_2\bar{\pi}_1 > \delta$, then $\rho_{k\delta}(\bar{\pi})$ is sub-regular immediately by 1.5.14. since $U_{\theta(\bar{\pi})}\gamma_{k\delta} - \gamma_{k\delta} = D_k(\bar{\pi})$ on $T_2\bar{\pi}_1 > \delta$. Put $k^* = k^s$ which we found in 1.5.14. and just choose $\delta(k^*) < \epsilon$ we used for partitioning I.

Hence $\rho_{k\delta}(\bar{\pi})$ is sub-regular on I for $0 < k < k^*$ and $0 < \delta < \delta(k^*)$.

//

i) Now if $\theta(\bar{\pi}) \sim O(\pi_i^{\alpha})$ as $\pi_i \downarrow 0$ at each boundary, we choose a family $\rho_{k\delta}(\bar{\pi})$ which are sub-regular on I and take $\lim_{\delta\downarrow 0} \rho_{k\delta}(\bar{\pi}) \equiv 1$ for $0 < k < \hat{k}$ and $\bar{\pi} \neq 0$.

Thus $\lim_{\delta\downarrow 0} \rho_{k\delta}(\bar{\pi}) = \gamma_{\theta(\bar{\pi})}(\bar{\pi}) = \gamma_{opt}$ and hence for $\alpha > 1$, we have optimality.

ii) By remark 1.5.15 a), this optimality is only dependent on the behaviour of $\theta(\bar{\pi})$ at the boundaries $[0, \epsilon)$ and $(1-\epsilon^*, 1]$.

//

<u>Theorem</u> 1.5.17.

If $\alpha = 1$ at each boundary and rule is of the form $\theta(\bar{\pi}) = \theta\bar{\pi}(1-\bar{\pi})$, $q_1 > q_2$, then:- a) If $1 > \theta > 6(1 - q_2/q_1)$ then the rule is ϵ-optimal.

(and with more careful analysis we obtain $1 > \theta > (1 - q_2/q_1)/(1 - \log_e 2)$)

b) If $\theta < 1$ and $\theta < (1 - q_2/q_1)$ then the rule is optimal.

(and with more careful analysis we obtain $\theta < (1 - q_2/q_1)\, 2(1 + \theta/16)/(1 + q_2/q_1)$)

<u>Proof</u>.

For the conditionally optimal transition case, learning is due both to boundary and central behaviour so we must assert the form of $\theta(\bar{\pi})$ on $I = [0, 1]$, rather than just at the boundaries, as for $\alpha > 1$.

I shall prove both a) and b) using $\gamma_k(x) = \int_0^x \left(\frac{1-\bar{\pi}}{\bar{\pi}}\right)^k d\bar{\pi}$,

in a similar way to that in which Norman (1968) proves 1.5.4.

First, I consider the simpler to prove, b) for which I construct a sub-regular family $\gamma_k(x)$ with $k\uparrow 1$, and $\lim_{k\uparrow 1} \gamma_k(x)/\gamma_k(1) = \gamma_{opt}(x)$ as in 1.5.12.

i) $U_{\theta(\bar\pi)}\gamma_k(x) - \gamma_k(x) = q_1\bar\pi_1 \int_{\bar\pi_1}^{T_1\bar\pi_1} \left(\frac{1-\bar\pi}{\bar\pi}\right)^k d\bar\pi - q_2\bar\pi_2 \int_{T_2\bar\pi_1}^{\bar\pi_1} \left(\frac{1-\bar\pi}{\bar\pi}\right)^k d\bar\pi$

and since $\left(\frac{1-x}{x}\right)^k$ is monotone$\downarrow$ on $[0,1]$.

$$U_{\theta(\bar\pi)}\gamma_k(x) - \gamma_k(x) \geq q_1\bar\pi_1\theta\bar\pi_1\bar\pi_2^2\left[\left(\frac{(1-\bar\pi-\theta\bar\pi(1-\bar\pi)^2)}{(\bar\pi+\theta\bar\pi(1-\bar\pi)^2)}\right)\right]^k - q_2\bar\pi_2\theta\bar\pi_1^2\bar\pi_2\left[\frac{(1-\bar\pi+\theta\bar\pi^2(1-\bar\pi))}{\bar\pi(1-\theta\bar\pi(1-\bar\pi))}\right]^k$$

$$\geq \theta\bar\pi_1^2\bar\pi_2^2\,\bar\pi_2^k/\bar\pi_1^k\left[q_1\left[\frac{(1-\theta\bar\pi_1\bar\pi_2)}{(1+\theta\bar\pi_2^2)}\right]^k - q_2\left[\frac{(1+\theta\bar\pi_1^2)}{(1-\theta\bar\pi_1\bar\pi_2)}\right]^k\right].$$

$$\geq \theta\bar\pi_1^2\bar\pi_2^2\,\bar\pi_2^k/\bar\pi_1^k\, B(\bar\pi_1) \quad \text{with } \bar\pi_2 = 1-\bar\pi_1.$$

Now we find $\min_{\bar\pi_1} B(\bar\pi_1)$.

<u>Lemma</u>. 1.5.18.

$\frac{\partial}{\partial\bar\pi_1}B(\bar\pi_1)\Big|_{\frac{1}{2}} = 0$ and this is a minimum and is unique.

<u>Proof</u>.

$$\partial/\partial\bar\pi\left[q_1(1-\theta\bar\pi_1\bar\pi_2)^{2k} - q_2(1+\theta\bar\pi_1^2)^k(1+\theta\bar\pi_2^2)^k\right]$$

$$= 2\theta k(\bar\pi_1-\bar\pi_2)\left[q_1(1-\theta\bar\pi_1\bar\pi_2)^{2k-1} + q_2(1-\theta\bar\pi_1\bar\pi_2)\left[(1+\theta\bar\pi_2^2)(1+\theta\bar\pi_1^2)\right]^{k-1}\right]$$

and $\frac{\partial B}{\partial\bar\pi_1} = 0$ iff $\bar\pi_1 = \frac{1}{2}$ and $\partial B/\partial\bar\pi_1\big|_{\frac{1}{2}+\epsilon} > 0$. $\partial B/\partial\bar\pi_1\big|_{\frac{1}{2}-\epsilon} < 0$.

//

Hence we ensure $U_{\theta(\bar\pi)}\gamma_k(\tfrac{1}{2}) - \gamma_k(\tfrac{1}{2}) > 0$ or $q_1(1-\theta/4)^2 > q_2(1+\theta/4)^2$.

Thus $\theta < \zeta(q_1-q_2)/(q_1+q_2)$ with $\zeta = 2(1+\theta^2/16)$.

Then for $\theta < 1$, $2(1+\theta^2/16) > 2$ and so we need satisfy only,

$\theta < (1-q_2/q_1)/(1+q_2/q_1)/2$ But $(q_1+q_2) < 2q_1$ and hence

we get b):- $\theta < 1$ and $\theta < (1-q_2/q_1) \Rightarrow$ rule is optimal.

The more precise inequality will allow us to give certain useful counter-examples in 1.5.20.

<u>Lemma</u>. 1.5.19.

Given any $\theta < 4$, and hence $\theta(\tfrac{1}{2}) < 1$, then $\exists$ q_1, q_2 s.t. the rule is optimal.

Proof.

If $q_2 = 0,\ q_1 \neq 0$ then we need only satisfy $(\theta/_4 - 1)^2 \geq 0$ which holds for all real θ. Clearly if $q_1 \neq 0$ and q_2 is sufficiently small, we shall still satisfy the inequality 1.5.17. b), in its strongest form, by using continuity arguments. Note $(\theta/_4 - 1)^2 = 0$ iff $|\theta| = 4$.

//

ii) For a) we prove that $\exists\, k^* < 1$ s.t. $\forall k : 1 > k > k^*$ then $\gamma_k(\bar{u}) = \int_0^{\bar{u}} \left(\frac{1-\bar{u}}{\bar{u}}\right)^k d\bar{u}$ is super-regular. Thus $\lim_{n \to \infty} U^n_{\theta(\bar{u})} \gamma_k(\bar{u}) = \gamma_{\theta(\bar{u})}(\bar{u}) < \gamma_k(\bar{u})$ and the rule is at best ϵ-optimal.

Again we use the argument based on partitioning. $I_1 = [0, \delta)$. $I_2 = [\delta, 1-\delta^*]$. $I_3 = (1-\delta^*, 1]$.

Consider I_1. We find k_1 s.t. γ_k is super-regular on I_1, for $1 > k > k_1$

$$U_{\theta(\bar{u})} \gamma_k - \gamma_k = q_1 \bar{u}_1 \int_{\bar{u}_1}^{T_1 \bar{u}_1} \left(\frac{1-\bar{u}}{\bar{u}}\right)^k d\bar{u} - q_2 \bar{u}_2 \int_{T_2 \bar{u}_1}^{\bar{u}_1} \left(\frac{1-\bar{u}}{\bar{u}}\right)^k d\bar{u}.$$

$$\leq \bar{u}_1^{1-k}/_{(1-k)} \left[q_1 \bar{u}_1 \left[(1+\theta \bar{u}_2^2)^{1-k} - 1 \right] - q_2 \bar{u}_2^{1+k} \left[1 - (1 - \theta \bar{u}_1 \bar{u}_2)^{1-k} \right] \right]$$

Thus $U\gamma < \gamma$ if

$$q_1 \bar{u}_1 \left[\theta \bar{u}_2^2 (1-k) - k(1-k)\theta^2 \bar{u}_2^4/_2 + k(1-k)(1+k)\theta^3 \bar{u}_2^6/_6 \right] < q_2 \bar{u}_2^{1+k} (\theta(1-k) \bar{u}_1 \bar{u}_2).$$

or $$(q_1 - q_2 \bar{u}_2^k) < k q_1 \theta \left[\bar{u}_2^2/_2 - (1+k) \bar{u}_2^4 \theta/_6 \right]$$

or $$\theta > 2(1 - (q_2/q_1) \bar{u}_2^k) / (k \bar{u}_2^2 (1 - (1+k) \bar{u}_2^2 \theta/_3)).$$

Now since δ can be made arbitrarily close to 0, and k can be arbitrarily close to 1. $\theta > 2(1 - q_2/q_1) / (1 - 2\theta/_3)$ will give $U\gamma < \gamma$

and for $\theta < 1$, $\theta > 6(1 - q_2/q_1)$ for (1-k), δ sufficiently close to 0.

Further, we have $\theta > 2(1 - q_2/q_1) / (1 - 2\theta/_3 + 2\theta^2/_4 - 2\theta^3/_5 + \ldots)$ on taking further terms.

and if $\theta < 1$ we need satisfy only $\theta > (1 - q_2/q_1) / (1 - \log_e 2)$.

or for $\theta < 1$ (since our binomial expansions hold only for $\theta < 1$), we get more precisely $\theta < (\theta - \log_e(1+\theta)) / (1 - q_2/q_1)$ gives ϵ-optimality.

It also seems intuitively clear that if $\theta^* < 1$ gives ϵ-opt, then all θ s.t. $4 > \theta > \theta^*$ give at best only ϵ-opt, but we need the conjectured generalisation to 1.4.6. to prove this naturally.

Consider I_3. Again, on approximating the integrals, we get,

$$U\gamma_k - \gamma_k \leq \bar{\pi}_2^{1+k}/(1+k) \left(q_1 \bar{\pi}_1^{1-k}\left(1 - (1-\theta\bar{\pi}_1\bar{\pi}_2)^{1+k}\right) - q_2\bar{\pi}_2\left((1+\theta\bar{\pi}_1^2)^{1+k} - 1\right)\right)$$

and $U\gamma < \gamma$ if $q_1\bar{\pi}_1^{1-k}(\theta\bar{\pi}_1\bar{\pi}_2(1+k)) < q_2\bar{\pi}_2(\theta\bar{\pi}_1^2(1+k) + \theta^2 k(k+1)\bar{\pi}_1^4/2 + \epsilon)$

where $\epsilon = (1-k) f(\theta, k, \bar{\pi})$ and $\lim_{k\uparrow 1} \epsilon = 0$.

Then as for I_1, we take k sufficiently close to 1, and here also $\bar{\pi}_1$ sufficiently close to 1. This gives $\gamma_k(x)$ super-regular on I_3 if $1 > k > k_2$ and $1 > \theta > 2(q_1/q_2 - 1)$

However, comparing I_1 and I_3 conditions:- $q_1(1 - 2\theta/3) < q_2$ if $\theta > 3/2(1 - q_2/q_1)$

Thus the stricter condition is that at I_1, and so we may disregard that at I_3.

Consider I_2. Here we take k = 1 and use compactness.

$$U_{\theta(\bar{\pi})}\gamma_{k=1} - \gamma_1(x) = (q_2 - q_1)\theta\bar{\pi}_1^2\bar{\pi}_2^2 + q_1\bar{\pi}_1 \log(1+\theta\bar{\pi}_2^2) + q_2\bar{\pi}_2 \log(1 - \theta\bar{\pi}_1\bar{\pi}_2).$$

and $U\gamma < \gamma$ when $\theta > 2(q_1 - q_2) / (q_1\bar{\pi}_2(1-\theta\bar{\pi}_2^2) + q_2\bar{\pi}_1)$.

or more precisely $\theta > (q_1 - q_2) / (q_2\bar{\pi}_1/2 + q_1\bar{\pi}_2(\theta\bar{\pi}_2^2 - \log_e(1+\theta\bar{\pi}_2^2))/\theta^2\bar{\pi}_2^4)$

or $\theta > 2(q_1 - q_2) / (q_2\bar{\pi}_1 + q_1\bar{\pi}_2(1 - 2\theta\bar{\pi}_2^2/3 + 2\theta^2\bar{\pi}_2^4/4 - 2\theta^3\bar{\pi}_2^6/5 + \ldots))$.

Thus $U\gamma < \gamma$ if above inequality holds. We must now use compactness to relate $k=1$ to $k^* < 1$ on $[\delta, 1-\delta^*]$.

Since $\int_x^{\bar{\pi}_1 x} \left(\frac{1-\bar{\pi}}{\bar{\pi}}\right) d\bar{\pi}$ is bounded on $x \in I_2$ (compact), we have $|U\gamma_1 - \gamma_1(\bar{\pi}) - (U\gamma_k(\bar{\pi}) - \gamma_k(\bar{\pi}))| < \epsilon'$ for $1 > k > k_3$ say, and $\bar{\pi} \in I_2$

The ϵ' is chosen s.t. $(\gamma_1 - U\gamma_1) > 2\epsilon'$ and hence $\gamma_k - U\gamma_k > \epsilon'$ and $\gamma_k(\bar{\pi})$ is super-regular on I_2. Now $\min_{\bar{\pi}_1+\bar{\pi}_2=1} (q_2\bar{\pi}_1 + q_1\bar{\pi}_2(1 - 2\theta\bar{\pi}_2^2/3 + \ldots))$ exists by compactness and with some difficulty can be found to lie at the lower end-point $\bar{\pi}_1 = \delta$. (and θ must satisfy the given inequality at all times.)

Finally, we combine the results and obtain the required bounds.

We have $\gamma_k(\bar{u})$ is super-regular on $I=[0,1]$ if $1>\theta>(1-q_2/q_1)/(1-\log_e 2)$.
where $1>k>\max(k_1,k_2,k_3)$. And hence we have that the rule is then ϵ-optimal.

//

To end this rather lengthy section of analysis, we shall consider how # alternations relates to optimality in the transition class. We shall denote this class by $\mathcal{R}_c$ since we have now proved in 1.5.17. that such rules are only conditionally optimal.
And $\mathcal{R}=\mathcal{R}_\epsilon \cup \mathcal{R}_c \cup \mathcal{R}_0$.

Lemma 1.5.20.

i) If $1>\theta>(1-q_2/q_1)/(1-\log_e 2)$ and $\theta(\bar{u})=\theta\bar{u}(1-\bar{u})$ then $\#alts \overset{a.s}{=} \infty$ and the rule is ϵ-optimal.

ii) If $4>\theta>1$ and $\theta q_1>1>\theta q_2$ then we may choose $\epsilon>0$ s.t. $q_2<\epsilon$ and the rule is optimal with $\#alt<\infty$.

Proof.

i) We use 1.5.7. and note that $\theta, q_1, q_2<1$ gives $\#alts \overset{a.s}{=} \infty$, but from 1.5.17, the constraint on θ gives ϵ-optimality.

ii) Using the finer bound of 1.5.17 b) we have $\theta<(1-q_2/q_1)2(1+\theta^2/16)/(1+q_2/q_1)$ gives optimal learning.
But as $q_2\downarrow 0$, we find $\theta=4-\delta$ will satisfy the above for ϵ sufficiently small. Also $\theta q_1>1 \Rightarrow \#alt<\infty$ for the rule,since we have just seen that it is optimal. //

Lamperti and Suppes (1960) have investigated a family of conditionally optimal learning rules, called β-rules. These give $E\#alts<\infty$, but only when their behaviour is actually optimal, for convergence to the boundaries only occurs for certain $\underset{\sim}{q}$ and β_{ij}. Since our conditionally optimal class is U.L. , we do always converge giving at least ϵ-optimality. Optimality is determined by the relation between the learning parameter θ (or β_{ij} for β-rule) and q_1/q_2 (similarity between the actions).

Our bounds on conditional optimality are reasonably tight since the diffusion limit has its optimality transition at $\theta = 2(1 - q_2/q_1)$. In discrete time 1.5.17. asserts that if $\theta < 1$, $\theta < 2(1 - q_2/q_1)/(1 + q_2/q_1)$ gives optimality and $\theta > 3.3(1 - q_2/q_1)$ gives ϵ-optimality. This last inequality becomes arbitrarily close to the diffusion limit as $\theta \downarrow 0$, for consider $\theta > 2(1 - q_2/q_1)/(1 - 2\theta/3)$. Thus it seems reasonable to <u>conjecture</u> $\theta = K(\theta)(1 - q_2/q_1)$ with $K(\theta) > \theta$ and $\lim_{\theta \to 0} K(\theta) = 2$, $\lim_{\theta \to 4} K(\theta) = 4$, $K(\theta) \uparrow$ as $\theta \uparrow$, gives the discrete time transition. Then define $F(\theta) = 1 - \theta/K(\theta)$ so that $(q_2/q_1) = F(\theta)$ gives the transition, with $(q_2/q_1) < F(\theta)$ giving optimality. As in the generalised comparison theorem, our methods of super and sub-regularity are probably not subtle enough to give us it.

1.6. <u>n-Action Extensions</u>.

We define $\gamma^i_{\theta(\underline{\pi})}(\underline{\pi}) = \Pr\left(\lim_{t \to \infty} \pi_i(t) = 1 \mid \underline{\pi}(0) = \underline{\pi}\right)$ and if $(q_1 - q_i) > 0$ $i \neq 1$, we use $\gamma^1_{\theta(\underline{\pi})} = \gamma_{\theta(\underline{\pi})}$. Clearly we must have boundary conditions $\gamma^i_{\theta(\underline{\pi})}(\underline{e}_j) = \delta_{ij}$ where e_j is the unit vector along the j-axis. Also $\gamma^i_{\theta(\underline{\pi})}(\underline{\pi} \text{ s.t. } \pi_i = 0) = 0$.

<u>Lemma</u> 1.6.1.

$$U_{\theta(\underline{\pi})} \gamma^i_{\theta(\underline{\pi})}(\underline{\pi}) = \gamma^i_{\theta(\underline{\pi})}(\underline{\pi}) \quad \text{where} \quad U_{\theta(\underline{\pi})} \Upsilon(\underline{\pi}(t)) = E\left(\Upsilon(\underline{\pi}(t+1)) \mid \underline{\pi}(t)\right).$$

and the solution is unique if $\gamma^i_{\theta(\underline{\pi})}$ is continuous.

<u>Proof</u>.

We have convergence by 1.3.1. $\lim_{t \to \infty} \pi_j(t) \in \{0, 1\}$ $\forall j$.

Now
$$\lim_{n \to \infty} U^n_{\theta(\underline{\pi})} \Upsilon^i(\underline{\pi}) = 0. \Pr(\pi_i(\omega) = 0, \pi_j = \pi_j^* \ j \neq i) + 1. \Pr(\pi_i(\omega) = 1, \pi_j(\omega) = \pi_j^*. \ j \neq i)$$
$$= \gamma^i_{\theta(\underline{\pi})}(\underline{\pi}) \text{ with } \Upsilon^i(\underline{e}_j) = \delta_{ij} \text{ and } \Upsilon^i(\underline{\pi} \text{ s.t. } \pi_i = 0) = 0.$$

And putting $\Upsilon^i(\underline{\pi}) = \gamma^i_{\theta(\underline{\pi})}(\underline{\pi})$ we have the result. Uniqueness for continuous functions is proved as in 1.4.1. Thus $\Delta \gamma^i_{\theta(\underline{\pi})}(\underline{\pi}) \equiv 0$.

//

We shall now use the methods of super and sub-regularity, as in section 4, and we shall extend our results for R_ϵ, R_0 to n-actions. The $\theta(\underline{\pi})$ suffix on $\gamma_{\theta(\underline{\pi})}$ just denotes non-linear U.L., boundary convergent rules, which includes the $\theta_{ij}(\underline{\pi})$ family.

Note that the discontinuous functions $\gamma^i_{\theta(\pi)}(\underset{\sim}{\pi}) \equiv 1$, except for $\gamma^i_{\theta(\pi)}(\pi_i = 0) = 0$, and $\gamma^i_{\theta(\pi)}(\underset{\sim}{\pi}) \equiv 0$ except for $\gamma^i_{\theta(\pi)}(\pi_i = 1) = 1$, always give $\Delta\gamma^i_{\theta(\pi)}(\underset{\sim}{\pi}) \equiv 0$
So we always have the existence of solutions to $\Delta\gamma \equiv 0$ but they will not necessarily be continuous or unique. However, we do have $\lim_{n\to\infty} U^n \gamma^i(\underset{\sim}{\pi}) = \Gamma^i(\underset{\sim}{\pi})$ holding <u>only</u> for the actual absorption probabilities, so we discriminate between solutions using our super and sub-regular $\psi^i(\underset{\sim}{\pi})$. In this way we can determine whether we indeed do have a true discontinuous solution, expressing optimality, as in 1.5.12.
We can actually obtain an infinity of solutions by setting $\sum_i r_i = 1$ $0 \le r_i \le 1$. and $\gamma^i(\underset{\sim}{\pi}) \equiv r_i$, except for the usual boundary conditions of 1.6.1. , so that $\Delta\gamma^i(\underset{\sim}{\pi}) \equiv 0$. But these discontinuous functions do not represent actual absorption probabilities, apart form the optimality case, $r_{i_0} = 1$.

<u>Lemma</u> 1.6.2.

Let $q_1 > q_2 > q_2'$ and environment 1 has parameters q_1 , q_2.
" 2 " " q_1 , q_2'.

And define $\gamma^1_{\theta(\pi),i}(\underset{\sim}{\pi}) = \gamma^1_i(\underset{\sim}{\pi})$ for environment i.

Then if γ^1_1 is monotone, $\gamma^1_2 \geq \gamma^1_1$.

<u>Proof</u>.

We show $U_2\gamma^1_1 \geq \gamma^1_1$ and hence $\lim_{n\to\infty} U_2^n \gamma^1_1 = \gamma_2^1 \geq \gamma^1_1$

Now $U_2\gamma^1_1 - \gamma^1_1 = \pi_1 q_1 [\gamma^1_1(T_1\pi_1) - \gamma^1_1(\underset{\sim}{\pi})] - \pi_2 q_2' [\gamma^1_1(\underset{\sim}{\pi}) - \gamma^1_1(T_2\pi_1)]$.

and $U_1\gamma^1_1 = \gamma^1_1$ so $U_2\gamma^1_1 - \gamma^1_1 = \pi_2(q_2 - q_2')(\gamma^1_1(\underset{\sim}{\pi}) - \gamma^1_1(T_2\pi_1)) \geq 0$

by the monotonicity of $\gamma^1_1(\underset{\sim}{\pi})$.

//

We now extend this result to n-actions and in so doing, we prove an argument used by Narendra and Viswanathan (1972).

<u>Lemma</u> 1.6.3.

Let $q_1 > q_2 > \ldots > \ldots > q_n$ and define $\gamma^1_{\theta(\pi),1}(\underset{\sim}{\pi}) = \gamma^1_1(\underset{\sim}{\pi})$

Now associate $\gamma^1_2(\underset{\sim}{\pi})$ with parameters $q_1 > q_2 = q_3' = \ldots = \ldots = q_n'$.

Then if $\gamma^1_2(\underset{\sim}{\pi}) > \gamma^1_2(\underset{\sim}{\pi}^*)$ when $\pi_1 > \pi_1^*$ (monotone in first argument as in the linear rule θ =const), we have $\gamma^1_1(\underset{\sim}{\pi}) > \gamma^1_2(\underset{\sim}{\pi})$.

<u>Proof</u>.

We prove $U_1\gamma_2 \geq \gamma_2$

$$U_1\gamma_2 - \gamma_2 = \pi_1 q_1 \left(\gamma_2(T_1\underset{\sim}{\pi}) - \gamma_2(\underset{\sim}{\pi})\right) - \sum_{r=2}^{n} \pi_r q_r \left(\gamma_2(\underset{\sim}{\pi}) - \gamma_2(T_1\pi_r)\right).$$

and $U_2\gamma_2 = \gamma_2$

Then $$U_1\gamma_2 - \gamma_2 = \sum_{r=3}^{n} (q_2 - q_r)\pi_r \left(\gamma_2(\underset{\sim}{\pi}) - \gamma_2(T_1\pi_r)\right) \geq 0.$$

since $\gamma_2(\underset{\sim}{\pi}) > \gamma_2(T_1\pi_r)$ by monotonicity in first argument.

In particular $U_1\gamma_2 - \gamma_2 = \left(\gamma_2(\underset{\sim}{\pi}) - \gamma_2(T_2\underset{\sim}{\pi})\right) \sum_{r=3}^{n} (q_2 - q_r)\pi_r \geq 0$ for $\theta_{ij}(\underset{\sim}{\pi}) = \text{const} = \theta$

and hence $\gamma_1(\underset{\sim}{\pi}) \geq \gamma_2(\underset{\sim}{\pi})$ for linear ($\alpha = 0$) learning. This immediately gives the linear rule as n-action, ϵ-optimal, which is used for comparison in 1.6.6. //

<u>Theorem</u> 1.6.4.

In static environment $\mathfrak{M}$ with $q_1 = \ldots = q_m > q_{m+1} \geq \quad \geq q_n$ with learning under $\theta_{ij}(\underset{\sim}{\pi})$ rules with $\theta_{ij}(\underset{\sim}{\pi}) \sim O(\pi_i)^{\alpha_i}, \alpha_i > 1$ as $\pi_i \downarrow 0$

$\sim O(\pi_j)^{\alpha_j}, \alpha_j > 1$ as $\pi_j \downarrow 0$

Then $\exists i \in M$ s.t. $\lim_{t\to\infty} \pi_i(t) = 1$ where $M = \{i : i \leq m\}$

<u>Proof</u>.

Let $M^c = \{m+1, \ldots n\}$. Then under $\theta_{ij}(\underset{\sim}{\pi})$ rules we have:-

i) Convergence to $\pi_i \in \{0,1\}\ \forall i$ by 1.3.1.

ii) U.L. for all $\underset{\sim}{\pi}$.

If $q_m = q_j\ \forall j$ then we have nothing to prove.

Let $\pi_m = \sum_{i\in M} \pi_i$ and similarly $1 - \pi_m = \sum_{i\in M^c} \pi_i$

We have $$\pi_m(t+1) = \pi_m(t) + \sum_{j=m+1}^{n} \theta_{ij}(\underset{\sim}{\pi})\pi_j \quad \text{for } U_i(t) \text{ s.t. } i \in M.$$

$$\pi_m(t+1) = \pi_m(t) - \sum_{j\in M} \theta_{ij}(\underset{\sim}{\pi})\pi_j \quad \text{for } U_i(t) \text{ s.t. } i \in M^c.$$

And $$\sum_{j\in M^c} \theta_{ij}(\underset{\sim}{\pi})\pi_j \leq \max_{\substack{i\in M\\ j\in M^c}} \theta_{ij}(\underset{\sim}{\pi})(1-\pi_m).$$

And $$\sum_{j\in M} \theta_{ij}(\underset{\sim}{\pi})\pi_j \leq \max_{\substack{i\in M^c\\ j\in M}} \theta_{ij}(\underset{\sim}{\pi})\pi_m.$$

We could write $\Pi_m \to \Pi_m + \left(\sum_{j \in M^c} \theta_{ij}(\underline{\Pi}) \Pi_j \Big/ \sum_{j \in M^c} \Pi_j \right)(1 - \Pi_m) \qquad i \in M$

and $\Pi_m \to \Pi_m \left(1 - \left(\sum_{j \in M} \theta_{ij} \Pi_j \Big/ \sum_{j \in M} \Pi_j \right) \right) \qquad i \in M^c.$

Then w.l.o.g. we put $\theta_{ij}(\underline{\Pi}) = \theta(\Pi_i \Pi_j)^\alpha$ and $\theta^*_m = \theta(\Pi_m(1-\Pi_m))^\alpha$.

Clearly $\max_{\substack{i \in M \\ j \in M^c}} \theta_{ij}(\underline{\Pi}) \leq \theta^*_m$.

Thus we reduce n-actions to 2-actions with $q^*_1 = q_1$, $q^*_2 = q_{m+1}$, and learning rule $\theta^*_m(\underline{\Pi}_m)$.

Intuitively, in our reduction, we are always increasing the drift, yet we prove that the final 2-action, θ^*_m process is optimal, and then by comparison, so is the original process.

$$\Delta \Pi_m = \sum_{i \in M} \Pi_i \sum_{j \in M^c} \Pi_j (q_i - q_j) \theta_{ij}(\underline{\Pi}) < (q^*_1 - q^*_2) \Pi_m (1 - \Pi_m) \theta^*_m(\underline{\Pi}_m)$$

Note that the 2-action, θ^*_m process is U.L. and boundary absorbed.

<u>Lemma</u> 1.6.5.

Define $\gamma_1(\underline{\Pi}) = \Pr(\Pi_i \to 1,\ i \in M \mid \underline{\Pi}(0) = \underline{\Pi})$ under $q_1 = \ldots = q_m > q_{m+1} \geq \ldots \geq q_n$.

$\gamma_2(\underline{\Pi}) = \Pr(\Pi_i \to 1,\ i \in M \mid \underline{\Pi}(0) = \underline{\Pi})$. under $q_1 = \ldots = q_m > q_{m+1} = \ldots = q_{m+1}$.

and $\gamma_{2^*}(\underline{\Pi}_m) = \Pr(\Pi_m \to 1$ under $\theta^*_m(\underline{\Pi}_m)$, 2-action rule$)$.

Then for concave γ_{2^*} we have $\gamma_1(\Pi_1 \ldots \Pi_n) \geq \gamma_{2^*}(\Pi_m, 1 - \Pi_m) \quad \forall \underline{\Pi}$.

<u>Proof</u>.

For concave γ_{2^*} we can use the 2-action comparison theorem 1.4.5. , which is easily extended to hold for such concave $\gamma(\underline{\Pi})$. Note that here we have action 1 as optimum which gives concave $\gamma(\underline{\Pi})$ whilst in 1.4.5. action 2 is optimum, giving convex $\gamma(\underline{\Pi})$.

If $\theta_{ij}(\underline{\Pi}) \leq \theta^*_m$ then $U_{\theta_{ij}} \gamma_{2^*} \geq \gamma_{2^*}$, where we have partition $\{1, \ldots m\}$, $\{m+1, \ldots n\}$ and reward probabilities q^*_1 and q^*_2, with learning function $\theta_{ij}(\underline{\Pi})$. So we update as if we had n-actions with $\theta_{ij}(\underline{\Pi})$, yet we compare with 2-action rule $\theta^*_m(\underline{\Pi}_m)$.

But by definition $\theta^*_m(\underset{\sim}{\pi}_m) \geq \max_{\substack{i\in M\\ j\in M^c}} \theta_{ij}(\underset{\sim}{\pi}) \quad \forall \underset{\sim}{\pi}$ and hence if we fix $\underset{\sim}{\pi}(\lambda)$, we have the result $\gamma_2(\pi_1 \ldots \pi_n) \geq \gamma_{2^*}(\pi_m, 1-\pi_m) \quad \forall \underset{\sim}{\pi}.$

with $\pi_m = \sum_{i\in M} \pi_i$. //

Now a) $\theta^*_m \sim O(\pi_m^{\alpha>1})$ as $\pi_m \downarrow 0$

$\sim O(1-\pi_m)^{\alpha>1}$ as $\pi_m \uparrow 1$

Hence by 1.5.12. $\theta^*_m(\underset{\sim}{\pi}_m)$ is optimally 2-action learning, and so is concave.

b) By 1.6.5. we have $\gamma_{\theta_{ij}} \geq \gamma_{\theta^*_m} \quad \forall \underset{\sim}{\pi}$ so that we have n-action optimality over $q_1 = \ldots = q_m > q_{m+1} \ldots = q_{m+1}$.

c) By 1.6.3. for monotone $\gamma_{\theta_{ij}}$ we have $\gamma_1 \geq \gamma_2$ and hence $\gamma_{\theta_{ij}}(\underset{\sim}{\pi}) \equiv 1$ except at $\pi_m = 0$.

So $\exists i \in M$ s.t. $\lim_{t\to\infty} \pi_i(t) = 1$.

This proof may seem artificial since the lemmas are being used in the "degenerate" cases. However, as in 1.5.12., we could consider the discontinuous optimal limit $\gamma_{\theta_{ij}}(\underset{\sim}{\pi})$ as the limit of a sub-regular family and then obtain the n-action optimality.

//

Intuitively, the $\theta_{ij}(\underset{\sim}{\pi})$ rules compare actions in pairs, and no optimal action is allowed to vanish, by a simple extension of asymtotic reflection from sub-optimal boundaries. This is the principle of boundary learning which will be treated conceptually in the next section. Note that $\theta_{ij}(\underset{\sim}{\pi}) < 1$ must always hold.

We now extend our ϵ-optimality result to all n-action boundary absorbed rules.

<u>Corollory</u> 1.6.6.

The family of $\theta_{ij}(\underset{\sim}{\pi})$ rules and the boundary absorbed $\theta(\underset{\sim}{\pi})$ rules are all at least ϵ-optimal.

$$\lim_{\theta \downarrow 0} \lim_{t \to \infty} \pi_m(t) = 1.$$

Proof.

We have $U_{\theta_{ij}} \gamma_{\theta_{ij}} = \gamma_{\theta_{ij}}$ and we prove $U_{\theta_{ij}} \gamma_\theta \geq \gamma_\theta$ where

γ_θ corresponds to $q_1 > q_2 = \dots = q_2$, θ =const, with a 2-action or n-action rule, as identical in result.

$\gamma_{\theta_{ij}}$ corresponds to $q_1 > q_2 \geq \dots \geq q_n$ $\theta_{ij}(\underline{\pi})$ rule.

If there are multiple optima, then the proof proceeds as 1.6.4. , with partitioning $M \vee M^c$.

Now $U_{\theta_{ij}} \gamma_\theta - \gamma_\theta = q_1 \pi_1 \left(\gamma(T_1^* \pi_1) - \gamma(\pi_1) \right) - \sum_{j \geq 2} q_j \pi_j \left(\gamma(\pi_1) - \gamma(T_{2,j}^* \pi_1) \right).$

where $T_{2,j}^* \pi_1 = \pi_1 (1 - \theta_{1j}(\underline{\pi}))$

And since $\gamma_\theta(\pi_1)$ is concave, we proceed as in the 2-action proof, 1.4.6.

$$U_{\theta_{ij}} \gamma_\theta - \gamma_\theta \geq \theta_{11}(\underline{\pi}) q_1 \pi_1 \pi_2^* \left(\frac{\gamma(T_1^* \pi_1) - \gamma(\pi_1)}{T_1^* \pi_1 - \pi_1} \right) - q_2 \pi_1 \left(\sum_{j \geq 2} \theta_{ij} \pi_j \right) \left(\frac{\gamma(\pi_1) - \gamma(T_{2j}^* \pi_1)}{\pi_1 - T_{2j}^* \pi_1} \right).$$

And choose θ s.t. $\theta_{ij}(\underline{\pi}) \leq \theta \ \forall i, j, \underline{\pi}$. with $\pi_2^* = 1 - \pi_1$.

$$\geq \theta_{11}(\underline{\pi}) q_1 \pi_2^* \pi_1 \left(\frac{\gamma(T_1 \pi_1) - \gamma(\pi_1)}{T_1 \pi_1 - \pi_1} \right) - q_2 \pi_1 \left(\sum_{j \geq 2} \theta_{ij} \pi_j \right) \left(\frac{\gamma(\pi_1) - \gamma(T_2 \pi_1)}{\pi_1 - T_2 \pi_1} \right).$$

$$\geq \theta_{11}(\underline{\pi})/\theta \ (U_\theta \gamma_\theta - \gamma_\theta) = 0.$$

Then by 1.6.3. we have $\lim_{\theta \downarrow 0} \gamma_\theta(\underline{\pi}) = 1 \quad \pi_1 \neq 0$ and hence we also have $\theta_{ij}(\underline{\pi})$ rules are all at least ϵ-optimal.

$$\gamma_{\theta_{ij}}(\underline{\pi}) \geq \gamma_\theta(\underline{\pi}).$$

We now have the n-action generalisation of 1.4.5.

Clearly the above goes through with $\theta(\underline{\pi})$ boundary absorbed, even more easily, to give ϵ-optimality.

//

Corollory 1.6.7.

$$\text{If} \quad \left. \begin{array}{ll} \theta_{ij}(\underline{\pi}) \sim O(\pi_i)^\alpha & \pi_i \downarrow 0 \\ \sim O(\pi_j)^\alpha & \pi_j \downarrow 0 \end{array} \right\} \forall i, j.$$

then i) If $\alpha < 1$, the rule is at best ϵ -optimal.

ii) If $\alpha > 1$, the rule is optimal.

<u>Proof</u>.

By 1.6.6., all the rules are at worst ϵ -optimal, but with $i(\underline{\pi}_i) > 0$ where $i(\underline{\pi}_i)$ = prob (just take action i) for $\alpha < 1$ we have $\mathcal{E}(\# a | t)^n < \infty$. Whilst for $\alpha > 1$ we get the result from 1.6.4.

//

Note that it is unresolved if the $\theta(\underline{\pi}) = K(\prod_{i=1}^{n}(1-\pi_i)^{\alpha})$ are optimal for certain α , but 1.6.6 gives them all ϵ-optimal with $\gamma_{\theta(\underline{\pi})} \geq \gamma_{\theta}$.

We have now considered R_ϵ, R_0 extensions, leaving the difficult R_L . Here we can use 1.6.3. to reduce the reward parameters to $q_1 > q_2 = \dots = q_2$ and for optimality we can find K s.t. $\theta < K(1-\frac{q_2}{q_1})$ and proceed in reverse as in 1.6.4. with our concave $\gamma_{opt}(\underline{\pi})$. However, for ϵ -optimality we cannot use 1.6.3. easily for we require monotonicity, and 1.6.5. needs concavity , and fails to work in reverse to give bounds away from optimality. Thus a thorough treatment would seem to require the construction of a n-action sub-regular family $\gamma_R(\underline{\pi})$ in analogy to the 2-action fundamental techniques of R_L analysis.

It is possible that the $\theta_{ij}(\underline{\pi})$ are uniquely optimal family of rules. But further work remains to be done on properties such as the $\mathcal{E}\#alts$, for $K(\prod_j(1-\pi_j)^{\alpha}) = \theta(\underline{\pi})$ rules, if they are to be excluded from consideration. We saw in 1.2 that the $\theta(\underline{\pi})$ and $\theta_{ij}(\underline{\pi})$ rules essentially exhaust U.L. possibilities, and this holds even if we have R - P rules with T_{ij} and S_{ij} in use.

The use of T_{ij}, gives difficulties in normalisation, and the corresponding $\theta_{ij}(\underline{\pi})$ rules can only be defined when $\theta_{ij}(\underline{\pi}) \sim K(\pi_i \pi_j)^{\alpha \geq 1}$.

$\pi_i(t+1) = \pi_i(t) \pm \theta_{ii}(\underline{\pi}) \, (1-\pi_i(t))$. under $u_i(t)$, with sign depending on s(t) = 1,0.

$\pi_j(t+1) = \pi_j(t) \, (1 \mp \theta_{ij}(\underline{\pi}))$ $j \neq i$, with normalisation.

The corresponding diffusion is optimal when $\alpha > 1$,so a discrete time analysis should give similar behaviour to the reward-inaction rules. Difficulties arise in proving the sub regularity of diffusion absorption function w.r.t. the functional equation $\Delta \gamma_{\theta_j}(\underset{\sim}{\pi}) = 0$.

1.7. Comparison between ϵ-optimality and a.s. optimality.

Definition 1.7.1.

A rule is centrally learning at $\underset{\sim}{c} \in (I^n)^\circ$ if $\exists \delta > 0$ s.t. if we construct the regular n-simplex S, in homogenous co-ordinates with faces $(\pi_i - c_i) = \delta_i$ then $\Delta R(\underset{\sim}{\pi}) > 0$ for $\underset{\sim}{\pi} \in S$ (where $\Delta R(\underset{\sim}{\pi}) = \sum_i q_i \Delta \pi_i$ as usual).

Definition 1.7.2.

A rule is boundary learning at $\underset{\sim}{e}_j$ if $\exists \epsilon_j$ s.t. $\Delta R(\underset{\sim}{\pi}) \geq 0 \ \forall \underset{\sim}{\pi}$ s.t. $|\underset{\sim}{\pi} - \underset{\sim}{e}_j| < \epsilon_j$.

We first show that ϵ-optimality corresponds to central learning and so we take the diffusion approximation for the 2-action linear rule and verify that we obtain ϵ-opt without consideration of boundaries.

Example 1.7.3.

Let $0 < a < c < b < 1$, then we show $\lim\limits_{\theta \downarrow 0} P(\,a\,B\,b\,|\,c\,) = 0$, where $P(aBb|c) = \text{Prob}(\text{reach } \pi_1 \leq a \text{ before } \pi_1 \geq b \text{ starting from } c, \text{with } q_1 > q_2)$.

In continuous time $\phi'' + r(\pi)\phi' = 0$ with $r(\pi) = 2k \doteq 2(1 - q_2/q_1)/\theta$

Then subject to $\phi(b) = 0$. $\phi(a) = 1$

We get $\phi(\pi_1) = (e^{-2k\pi_1} - e^{-2kb}) / (e^{-2ka} - e^{-2kb})$ and $\phi(c) = (e^{-2k(c-b)} - 1) / (e^{-2k(a-b)} - 1)$.

$\phi(c) < 2e^{-2k(c-b)} / e^{-2k(a-b)}$ for k large and so $\phi(c) < 2\exp(-2k(c-a))$.

$$\lim_{k \uparrow \infty} \phi(c) = 0 \quad \text{for } c > a, \text{ and we have the result.} \qquad //$$

In 1.11.1 we show how this is linked with Wald's Identity, for the discrete time process.

From this example we see that to from an ϵ-optimal rule, we need only find some point $\underset{\sim}{c}$ at which the rule is centrally learning. Then with $\underset{\sim}{\pi}(0) = \underset{\sim}{c}$ we stop the rule at the first action i s.t. $|\pi_i - c_i| \geq \delta$. An n-dim analogy of 1.7.3 would show $\lim_{\theta \to 0} Pr\left(|\pi_i - c_i| \geq \delta, i \neq opt . \mathcal{B}. |\pi_{opt} - c_{opt}| \geq \delta \,|\, \underset{\sim}{\pi}(0) = \underset{\sim}{c}(0)\right) = 0$, using the methods of section 1.6. to reduce it to a 2-action problem.

In contrast, we now show that optimality corresponds to boundary learning at $\underset{\sim}{e}_j$ $\forall j$ and in 1.7.5. we shall prove that U.L. at boundaries, with $\alpha > 1$, together with inter-boundary communication are sufficient for optimality.

<u>Definition</u> 1.7.4.

i) The boundary point $\underset{\sim}{e}_j$ is stable iff $\Delta \pi_j \geq 0$ for $|\underset{\sim}{\pi} - \underset{\sim}{e}_j| < \epsilon_j$ and some $\epsilon_j > 0$. (The rule must be boundary learning with drift into boundary.)

ii) The boundary point e_j is unstable iff $\exists i \neq j$ s.t. $\Delta \pi_i > 0$ $\forall \underset{\sim}{\pi}$ s.t. $|\underset{\sim}{\pi} - \underset{\sim}{e}_j| < \epsilon_j$, for some $\epsilon_j > 0$, and $\Delta R(\underset{\sim}{\pi}) \geq 0$. (The rule is boundary learning with drift away from boundary).

In our proof of 1.5.12 we saw that we achieved optimality depending on $\Theta(\underset{\sim}{\pi})$ behaviour at the boundary, and learning could be made as rapid as required in finitely many compact subsets of I The U.L. property was only used centrally for communication.

iii) Let the boundary neighbourhood of e_j be $\mathcal{E}_j$. We say that $\underset{\sim}{\pi}(0) \in I \setminus \bigcup_j \mathcal{E}_j$ <u>communicates</u> with $\{\mathcal{E}_i\}$ if a.s. $\exists$ $t > 0$ s.t. $\underset{\sim}{\pi}(t) \in \mathcal{E}_k$ for some k and $\exists$ Prob $p_k(\underset{\sim}{\pi}(0)) > 0$ that we first reach $\bigcup_j \mathcal{E}_j$ at $\mathcal{E}_k$.

<u>Theorem</u> 1.7.5.

For a rule to give a.s. convergence to a stable boundary, it is sufficient that :-

i) Each boundary is either stable or unstable,and at least one is stable.

and ii) The rule is optimally boundary learning ($\mathcal{R}_0$) at $\underset{\sim}{e}_j$ $\forall j$.

and iii) $\underset{\sim}{\Pi} \in I \setminus \bigcup_j \mathcal{E}_j$ communicates with $\{\mathcal{E}_i\}$ $\forall \underset{\sim}{\Pi}$.

Proof.

We prove that as in 1.5.12, we are asymtotically reflected from unstable boundaries, using $\#$ up-crossings $\overset{a.s.}{<} \infty$ across any rational interval. (see Breiman 1968).

Suppose that we are in $\underset{\sim}{\Pi}(t) \in \mathcal{E}_i$, unstable; then if we were to have $\lim_{t\to\infty} \Pi_i(t) = 1$ with $\Pi_i(t^*) \in \mathcal{E}_i$ $\forall t^* > t$ we could extend the optimal rule to I and obtain a contradiction from its sub-optimality. (If $\exists j$ s.t. $\Delta\Pi_j > 0$, $\underset{\sim}{\Pi} \in \mathcal{E}_i$ then $\neg \lim_{t\to\infty} \Pi_j(t) = 0$ if we use optimal rules).

Hence we leave $\mathcal{E}_i$ and next enter $\mathcal{E}_k$ say, and for unstable k we repeat the previous argument for the process is markovian. If $\mathcal{E}_k$ is stable, $\exists$ a finite probability that we never leave. Now apply $\#$ up-crossings $\overset{a.s.}{<} \infty$ and we see that a.s. we do eventually remain in some stable $\mathcal{E}_m$ for all $t > t_0$.

//

This result actually follows easily from 1.5.12 and 1.6.4. , once we notice that the fundamental idea is asymtotic reflection at unstable boundaries, which leaves only the stable boundaries for absorption, if any exist.

Algorithm 1.7.6.

We have n-actions with reward probabilities q_i. To asymtotically converge to the optimum,(and attain its nbd $\mathcal{E}_{opt}$ at t_0 a.s. and remain there for all $t > t_0$) we apply optimal boundary learning to each $\underset{\sim}{e}_j$ and construct nbds $\mathcal{E}_j$. Set $\underset{\sim}{\Pi}(0) \in \mathcal{E}_i$ some i, and if $\exists t$ s.t. $\underset{\sim}{\Pi}(t) \notin \mathcal{E}_i$ set $\underset{\sim}{\Pi}(t+1) \in \mathcal{E}_k$ with probability $1/n$, $1 \leq k \leq n$.

Proof.

Apply 1.7.5. and note,assuming optimum is unique, that $\Delta\Pi_{opt} > 0$ on $\bigcup_i \mathcal{E}_i$, and communication is immediate. Hence $\lim_{t\to\infty} \Pi_{opt}(t) = 1$.
A similar argument follows in the case of multiple optima, when we obtain multiple stable boundaries.

//

The theorem 1.7.5. will be applied in chapter 2 to games between π-cells, in order to prove convergence to pure strategies, and a Nash Point in the case of general sum games. It allows us to investigate the limiting behaviour when the π-cells execute mixed strategies, in attemting to attain the Von Neumann value of the zero-sum game.

Remark 1.7.7.

We can relax the communication with finite probability to all $\mathcal{E}_j$ in 1.7.5., to give a more natural algorithm. Set $\underset{\sim}{\pi}(0) \in \mathcal{E}_1$ and if $\exists t$ s.t. $\underset{\sim}{\pi}(t) \notin \mathcal{E}_1$ set $\underset{\sim}{\pi}(t+1) \in \mathcal{E}_2$ and then visit boundary nbds cyclically until we are absorbed at some stable $\mathcal{E}_m$.

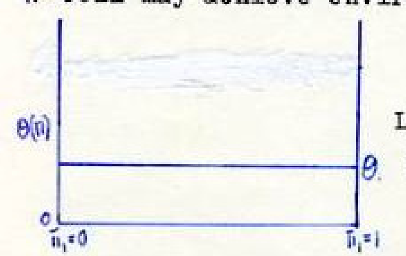
$$\left(\underset{\sim}{\pi}(t^*-1) \in \mathcal{E}_r ,\ \underset{\sim}{\pi}(t^*) \notin \mathcal{E}_r \text{ then } \underset{\sim}{\pi}(t^*+1) \in \mathcal{E}_{(r+1) \bmod n} \right)$$

The proof of sufficiency for a.s. convergence to stable $\mathcal{E}_m$ follows as in 1.7.5, by asymtotic reflection from unstable $\mathcal{E}_r$.

1.8. The Family of π-cell Learning Rules.

We have now derived a form of probabilistic stability theorem for discrete-time stochastic difference equations, and have shown optimality to be a boundary property.

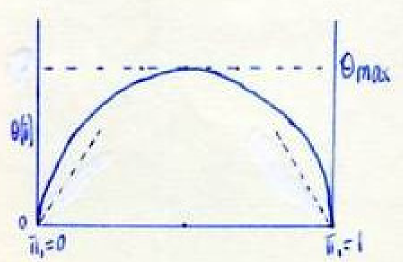
In this short section, we shall list certain U.L. rules which will be used in chapters 2 and 3, in order that the π-cell may achieve environmental adaptation.



Linear Rule

1. θ = constant, for all $\underset{\sim}{\pi}$.

Analysed by Norman (1968).

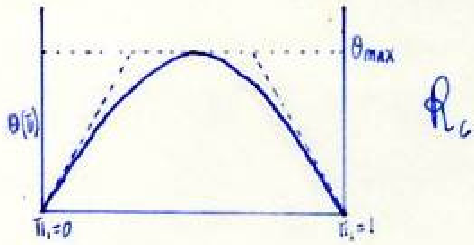


$\mathcal{R}_\epsilon$

2. ϵ-optimal rule with $\theta(\underset{\sim}{\pi}) = K(\pi_1 \pi_2)^{\alpha<1}$.

$$\mathcal{E}(\#\text{alts})^n < \infty, \ \forall n.$$

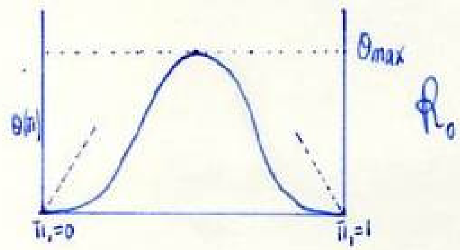
The linear rule is the case $\alpha = 0$.



3. Conditionally optimal rule

$$\theta(\underset{\sim}{\pi}) = K(\pi_1, \pi_2).$$

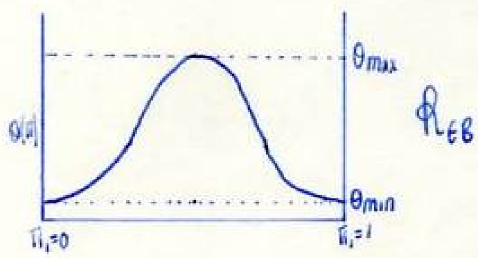
Optimality depends on $q^{s}_{u_i}$.



4. Optimal, with boundary learning

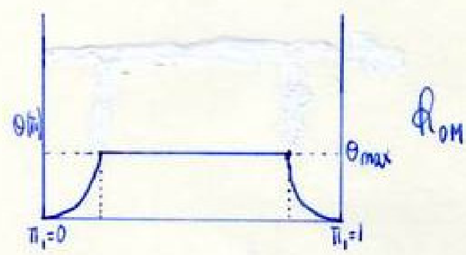
$$\theta(\underset{\sim}{\pi}) = K(\pi_1, \pi_2)^{\alpha>1}$$

$$\#\text{alts} \overset{a.s.}{=} \infty$$

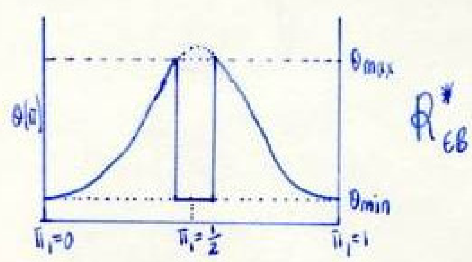


5. Boundary learning, yet ϵ-optimal, always converges with $\mathcal{E}(\#\text{alt})^n < \infty$.

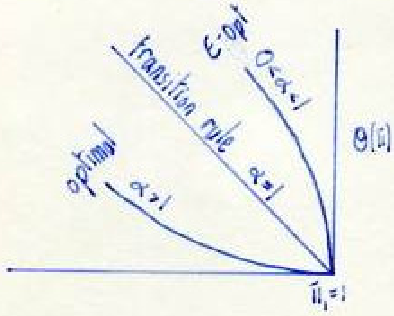
$$\theta(\underset{\sim}{\pi}) = K(\pi_1, \pi_2)^{\alpha>1} + \theta_{min}.$$



6. Optimal, with both boundary and central (middle) learning. Centre is used for mixed strategies, and boundary for Nash Points.



7. The extra central "dip" may help in the cellular differentiation of 3.3.7.



8. An expanded view of the boundary, showing the essential differences between R_0, R_ϵ and R_c .

1.9. Time Dependent Stimulus Probabilities.

Now that we have considered π-cells in static media, it is easy to extend the results to certain time varying $q_i^s(t)$.

Theorem 1.9.1.

If $\exists$ i s.t. $\limsup_{t\to\infty} q_j(t)/q_i(t) < 1 \quad \forall j \neq i$ then

under i) R_0 : $\pi_i \xrightarrow{a.s.} 1$

ii) R_ϵ : $\lim_{\theta \downarrow 0} \gamma(\pi_i) = 1$

and for 2-actions under

iii) If $\limsup_{t\to\infty} q_1(t)/q_2(t) = 1 - \delta < 1$ and $\theta < \delta$ then $\pi_2 \xrightarrow{a.s.} 1$.

iv) If $\limsup_{t\to\infty} q_1(t)/q_2(t) < 1$ and $\liminf_{t\to\infty} q_1(t)/q_2(t) > \log_e 2$

then for $1 > \theta > (1 - \lim q_1(t)/q_2(t))/(1 - \log_e 2)$ we have $\lim_{\theta \downarrow 0} \gamma(\pi_2) = 1$.

Proof.

We first prove a lemma.

Lemma 1.9.2.

i) $\lim_{t^*\to\infty} \prod_{t=0}^{t^*} U^t \psi^i(\underset{\sim}{\pi}) = \gamma^i(\underset{\sim}{\pi})$ with $\psi^i(\underset{\sim}{\pi} : \pi_i = 1) = 1, \ \psi^i(\underset{\sim}{\pi} : \pi_i = 0) = 0.$

ii) $\lim_{t^*\to\infty} \prod_{t=0}^{t^*} U^t \gamma^i(\underset{\sim}{\pi}) = \gamma^i(\underset{\sim}{\pi})$ and $\gamma^i(\underset{\sim}{\pi} : \pi_i = 1) = 1, \ \gamma^i(\underset{\sim}{\pi} : \pi_i = 0) = 0.$

where U^t is the expectation operator related to $q_i^s(t)$.

Proof.

We just have $\lim_{t\to\infty} \prod U^t \psi^i(\underset{\sim}{\pi}) = \mathcal{E}\left(\psi^i(\underset{\sim}{\pi}(\omega)) \mid \underset{\sim}{\pi}(0)\right) = \gamma^i(\underset{\sim}{\pi})$

and the result holds with $\psi^i(\underset{\sim}{\pi}) = \gamma^i(\underset{\sim}{\pi})$. //

Now we stop the process at t' s.t. $q_j(t)/q_i(t) < 1 \quad \forall t > t' \ \forall j \neq i$.

and we find sub-regular ψ^i s.t. $U^t \psi^i \geq \psi^i$ for all $t > t'$

Then $\gamma^i(\underset{\sim}{\pi}(t')) \geq \psi^i(\underset{\sim}{\pi}(t'))$.

For ii), we consider the result 1.5.4. of Norman (1968), for 2-actions.

It is clear that the result is only dependent on $q_1(t)/q_2(t)$ for we just require s to give $f(\pi,s) = \left(\frac{e^{s(1-\pi)}-1}{1-e^{-s\pi}}\right)\frac{\pi}{1-\pi} \geq (\leq)\ q_1/q_2 \quad \forall \pi.$

Hence the same $\gamma_{s,\theta} = e^{s\pi/\theta}$ will give us bounds on $\gamma(\pi)$. Also, we may continue to prove $\gamma_{\theta(\pi)}(\pi(t'))$ is monotone and convex, to give a comparison theorem analogous to 1.4.5. Then we get ii) after noting that the result is unaffected by U^t, $0<t<t'$ when we take the limit $\theta \to 0$. This gives us that all U.L. rules are at worst ϵ-optimal, under the given $q_i^s(t)$.

For i), under R_0 , we can still use $\lim_{\delta \downarrow 0} \xi_{k\delta}$ to give the result, by choosing $k_t < \frac{1}{\theta}\log q_1(t)/q_2(t)$, $\forall t > t'$ with $\inf_{t>t'} k_t > 0$ by $\overline{\lim}_{t\to\infty}\ q_j(t)/q_i(t) < 1 \quad \forall j \neq i.$

Finally, for iii), with R_c rules, we again just note that our super (sub) regular γ_k were functions only of q_1/q_2 so that the results extend naturally.

For n-actions in i) and ii), we mimic the extension theorems of section 1.6. , for $t > t'$. And in the slow learning limit, the trials $0 < t < t'$ do not affect limiting behaviour , for R_ϵ . The result for R_0 is ensured by the asymtotic reflection property.

//

Remarks 1.9.3.

a) Sawaragi and Baba (1975), defined a variable C+δ medium which effectively gave ϵ-optimality when $q_1(t) - q_2(t) > \delta \quad \forall t$ for the linear rule. Our 1.9.1. gives the result for all U.L. rules, and we see that the significant factor is $q_1(t)/q_2(t)$. It is 1.9.2. which enables us to prove 1.9.1. quickly, whilst Sawaragi and Baba actually repeated Norman's analysis in full with the new time-dependence.

b) Tsuji et Al (1973) define a non-static environment which generates $q_1(t) > c > q_2(t)$. and this is said to be completely isolated in the 0^{th} approximation. However, they use a structured ${}_2L_1$ (see3.3.3.) to learn in the environment, whilst a singleton π-cell will give convergence to $\pi_1 = 1$ under R_0 . (1.9.1.)

The paper is based on the work of Yasui and Yajima (1970) who consider 2-state, 2-symbol automata and define k^{th} order isolation. We shall find in chapter 3 when a structured automaton is required rather than a single θ_i . The ${}_2L_1$ is suited to a rapidly switching environment, which contrasts greatly with that of Tsuji et Al.

1.10. Skeletons.

We have noted(near 1.5.20.) that the β-rule is not U.L. We now gain further insight into such additive rules by attempting to construct a uniformly learning rule on a grid which spans $I = [0,1]$.

Definition 1.10.1.

The set $\{\pi_i : i \in Z\}$ forms a grid if

$$\left.\begin{aligned} \pi_{i+1} &= \pi_i + \theta_i(1-\pi_i) \\ \text{and} \quad \pi_{i-1} &= \pi_i(1-\theta_i) \end{aligned}\right\} \forall i \in Z \quad (\text{+ve and -ve integers}).$$

The set spans I if $\pi_\infty = 1,\ \pi_{-\infty} = 0$.

Theorem 1.10.2.

If a U.L. rule is defined on a grid, then it will not span $I=[0,1]$.

Proof.

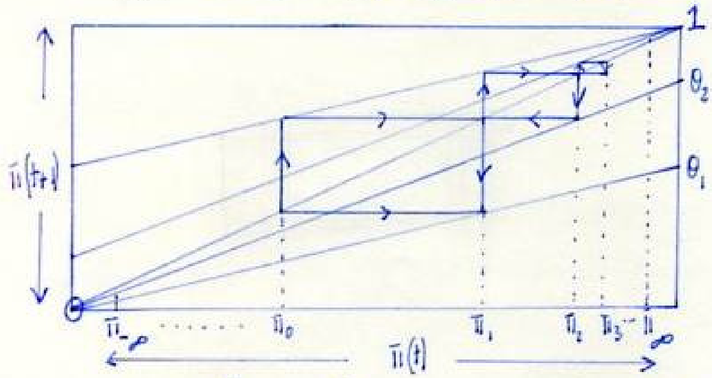
Put $\alpha_n = 1-\theta_n = \pi_{n-1}/\pi_n = (1-\pi_{n+1})/(1-\pi_n) = \pi^c_{n+1}/\pi^c_n$ by "action reversal".

Thus $\pi_1/\pi_\infty = \prod_{n=2}^{\infty} \alpha_n = \pi^c_\infty/\pi_2$

and if $\pi_\infty = 1,\ \pi^c_\infty = 0 \Rightarrow \pi_1 = 0$; absurd.

Hence $\prod_{n=0}^{\infty} \alpha_n$ is > 0.

Suppose $\pi_n \uparrow c$ then $(\pi_0)^2 = c(1-c) = (\alpha/1+\alpha)^2$ say and $c = \frac{1}{2} \pm ((1+3\alpha)(1-\alpha))^{\frac{1}{2}} / 2(1+\alpha)$



e.g. $\alpha = \frac{1}{3}$ gives $c(1-c) = \frac{1}{16}$
$c \sim \frac{1}{15}$

Thus $c \uparrow 1$ as $\alpha \downarrow 0$ with $\alpha = \alpha_0 = 1 - \theta_1$.

The identity $\alpha_n = \frac{\pi_{n-1}}{\pi_n} = \pi^c_{n+1} / \pi^c_n$ arises by rotating the diagram above through 180°; we are reversing actions, with

$$\pi = \text{Prob(take } u_1), \quad 1-\pi = \text{Prob(take } u_2) = \pi^c .$$

//

This result is easily seen to be closely related to the β-rule being non-U.L. The grid process gives $\left(\pi^c_{n+1}/\pi_{n+1}\right) / \left(\pi^c_n/\pi_n\right) = 1/\alpha_n\alpha_{n+1}$, whilst the β-rule has this ratio independent of n, in order for the additive process to span I.

If we could relate U.L. rules to a countable state space, it would help in proving results. So we shall define a spatial skeleton, just as in continuous time markov processes we may extract a τ-skeleton. This spatial skeleton will be used in an alternative method for proving optimality under R_0 .

Let $P(x_i \text{ B } x_{i+2} | x_{i+1})$ = Prob(reach $\pi \leq x_i$ before $\pi \geq x_{i+2}$ starting at $\pi = x_{i+1}$).

Definition 1.10.3.

A skeleton is a set $\{x_i : i \in Z\}$ which spans I =[0,1].

The probabilities $P(x_i B x_{i+2} | x_{i+1})$ can be calculated approximately, assuming no overshoot, using a diffusion, whilst Wald's Identity is useful when $|x_{i+2} - x_i| < \epsilon$ with ϵ sufficiently small for the learning process to be considered as a form of random walk.

In the skeleton process, we only look at the process at the trials at which the π-cell has just crossed a point x_i in the skeleton set.

1.11. <u>Staircases</u>. (An alternative method of proving optimality of R_0).

We first take $[a,b] \subset I$ with $|a - b| < \epsilon$ and $\theta(\pi) < \epsilon^{\gamma > 1}$ for $\pi \in [a,b]$, so that we can approximate the process as a R.W. and use Wald's Identity to obtain the absorption probabilities. Then we construct a skeleton which resembles a staircase, when projected on function $\theta(\pi)$, with successive step lengths of constant ratio β . This β-staircase will approximate and asymtotically converge to the learning function $\theta(\pi)$ at boundaries $\pi \in \{0,1\}$.

In the first lemma, we take $\theta(\pi) = \theta$ and approximate the process in $[a,b]$ as a random walk with increments:-

$\pi_1(t+1) = \pi_1(t) + \theta(1-b)$. if $u_1(t)$ and $s(t) = 1$.

$\pi_1(t+1) = \pi_1(t) - b\theta$. if $u_2(t)$ and $s(t) = 1$

$\pi_1(t+1) = \pi_1(t)$. if $s(t) = 0$.

<u>Lemma</u> 1.11.1.

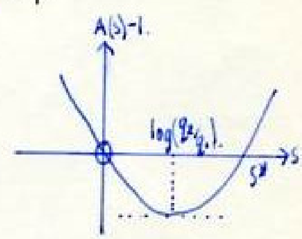
For the R.W. on $[a,b]$, $f(\pi) \geq \left((e^{sb/\theta} - e^{s\pi/\theta}) \big/ (e^{sb/\theta} - e^{s(a-d_-)}) \right)$.

where $f(\pi)$ = Pr(absorbed in $\pi \leq a$ starting from $\pi \in (a,b)$ before $\pi \geq b$).

d_- = max overshoot at $\pi = a$ boundary. And $q_2 > q_1$.

<u>Proof</u>.

Wald's Identity gives $E\left(e^{\sigma(\xi - \pi)} \big/ (A(\sigma))^n \right) = 1.$



and put $A(\sigma) = 1 = bq_1(e^{\sigma(1-b)\theta} - 1) + (1-b)q_2(e^{-\sigma b\theta} - 1) + 1.$

Thus $\sigma = 0$ or $\sigma = s/\theta$ with $s > \log(q_2/q_1) > 0$ for all b.

With $\partial A/\partial\sigma = 0$ if $q_1 e^{\sigma(1-b)\theta} = q_2 e^{-\sigma b\theta}$ or $e^s = q_2/q_1$

Let $\phi(\xi)$ be the stopping density; then W.I. gives

$$\int_{-\infty}^{a} \phi(\xi) e^{\sigma\xi} d\xi + \int_{b}^{\infty} \phi(\xi) e^{\sigma\xi} d\xi = e^{\sigma\pi} \quad \text{with } \sigma = s^*/\theta.$$

Now $q_2 > q_1$ with $s^* > 0$

$$\int_b^{\rho} \phi\, e^{s\varrho/\theta}\, d\varrho \geq (1 - f(\bar{\pi}))\, e^{sb/\theta}$$

$$< (1 - f(\bar{\pi}))\, e^{s(b+d_+)/\theta}.$$

Then $e^{s\bar{\pi}/\theta} \geq e^{sb/\theta}(1-f) + f e^{s(a-d_-)/\theta}$.

or $f(\bar{\pi}) \geq \left(e^{sb/\theta} - e^{s\bar{\pi}/\theta}\right) / \left(e^{sb/\theta} - e^{s(a-d_-)/\theta}\right)$ //

Remark 1.11.2.

i) The diffusion equation $\phi'' + r(\bar{\pi})\,\phi' = 0$ gives:-

$$\phi(\bar{\pi}) = \left(e^{-kb} - e^{-k\bar{\pi}}\right) / \left(e^{-kb} - e^{-ka}\right)$$

with $k = (q_1 - q_2)/q_2\theta$.

and so in W.I. we would expect $s \sim \left((q_2 - q_1)/q_2\right)$ as $\theta \downarrow 0$

i.e. $\theta = \epsilon^{\delta > 1}$ and $|a - b| = \epsilon$ with $\epsilon \downarrow 0$.

And for $\theta(\underset{\sim}{\bar{\pi}}) = \theta(\bar{\pi}_1, \bar{\pi}_2)^{\alpha > 1}$ and $\bar{\pi}_1 \downarrow 0$, the overshoot becomes negligibly small and the diffusion approximation becomes asymtotically, arbitrarily close to the discrete case.

ii) The bounds of Norman, 1.5.4., are just the case with $[a,b] = [0,1]$ since the drift is homogenous throughout the interval.

Theorem 1.11.3.

Under $\mathcal{R}_0$, $\theta'(\bar{\pi}) \equiv 0$ for $\bar{\pi} \neq 1$, where the $\mathcal{R}_0$ family is taken as $\theta(\underset{\sim}{\bar{\pi}}) = \theta(\bar{\pi}_1, \bar{\pi}_2)^{\alpha > 1}$, with $q_1 < q_2$.

Proof.

This is an alternative method of proving 1.5.12. which gives more insight into the process and is easily extended to markovian environments in 1.12.9.

We take $x_1 = 1 - 1/m$ s.t. $m \gg 1$ and so that $\theta(x_1) \sim (1/m)^{\alpha}$.

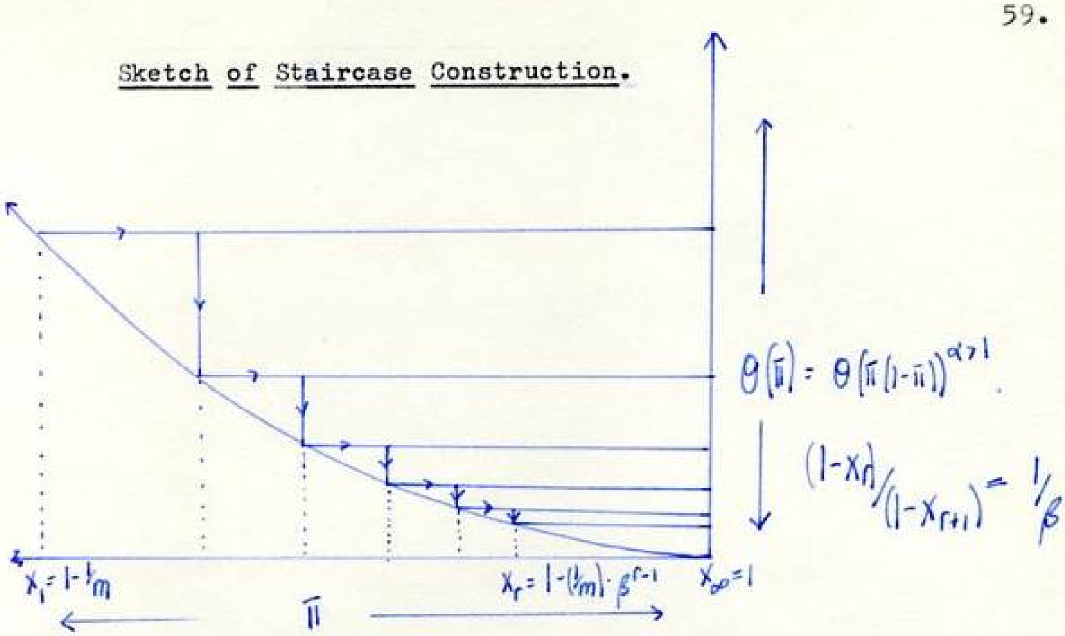
Now we define the β-staircase and estimate $\prod_{i=2}^{n} P(x_{i-1} B x_{i+1} | x_i)$

where $P(aBb|c)$ is defined in 1.10. and

$x_r = 1 - (1/m)\beta^{r-1}$, with $x_2 - x_1 = (1-\beta)/m \gg \theta(x_1)$ · on "first-step".

For $\theta(\bar{\pi}) = K(\bar{\pi}(1-\bar{\pi}))^{\alpha > 1}$ we require $m \gg (K\beta^{\alpha}/(1-\beta))^{1/(\alpha-1)}$

Sketch of Staircase Construction.



By 1.11.1. $Pr(x_{r-1}\ B\ x_{r+1} | x_r) \sim \left(e^{sx_{r+1}/\theta} - e^{sx_r/\theta}\right) / \left(e^{sx_{r+1}/\theta} - e^{sx_{r-1}/\theta}\right)$.

and $= \left(1 - e^{-(s/m\theta)\cdot\beta^r}\right) / \left(1 - e^{-3s\beta^r/m\theta}\right) = \left(1-e^{-\gamma_r}\right) / \left(1-e^{-3\gamma_r}\right)$.

$\sim \left(1 - e^{-\gamma_r}\right)$

neglecting higher order terms and overshoot.

with $\gamma_r = s\beta^r/m\theta$ and $\theta(x_{r-1}) \sim \theta/m^\alpha \left(\beta^{r-2}\right)^\alpha$.

We use $\theta(x_{r-1})$ since this gives smaller (drift/diffusion) $= r(\pi) \propto 1/\theta$. and hence is the worst possible case on $[x_{r-1}, x_{r+1}]$

For $I = [0,1]$, we had the comparison theorem 1.4.5 to prove this result, but as the drift for linear θ is homogenous on I, the result will hold on arbitrary intervals. We just require θ small enough so overshoot can be neglected.

Now $Pr\left(x_{r-1} B x_{r+1} | x_r\right) \sim 1 - \exp\left(- sm^{\alpha-1}/\theta\ \beta^{2\alpha}/\beta^{r(\alpha-1)}\right) = 1 - \exp(-\varrho_r)$.

and $\prod_{r=2}^{n} P\left(x_{r-1} B x_{r+1} | x_r\right) \sim \prod_{r=2}^{n}\left(1 - \exp\left(-sm^{\alpha-1}/\theta \cdot \beta^{2\alpha}/\beta^{r(\alpha-1)}\right)\right)$.

and hence $\prod_{r=2}^{n=\infty} Pr\left(x_{r-1} B x_{r+1} | x_r\right) > 0$ iff $\sum_{r=2}^{n=\infty} \exp(-\varrho_r) < \infty$

But by the ratio test we get $\sum_r \exp(-\varrho_r)$ converges extremely rapidly.

Let $a_r = \exp(-\varrho_r)$

Then $a_r/a_{r+1} = \exp\left(-\delta/\beta^{r(\alpha-1)}\right) / \exp\left(-\delta/\beta^{(r+1)(\alpha-1)}\right)$ with $\delta = (sm^{\alpha-1}/\theta)\beta^{2\alpha}$.

$= \exp\left(\delta/\beta^{r(\alpha-1)}\left(1/\beta^{\alpha-1} - 1\right)\right)$ and $\alpha > 1 \Rightarrow \beta^{\alpha-1} < 1$ if $0 < \beta < 1$.

Hence for $r > r_0$ say $a_r/a_{r+1} > \exp\left(\delta/\beta^{r_0(\alpha-1)}\left(1/\beta^{\alpha-1} - 1\right)\right)$.

and we have convergence by comparison with geometric series.

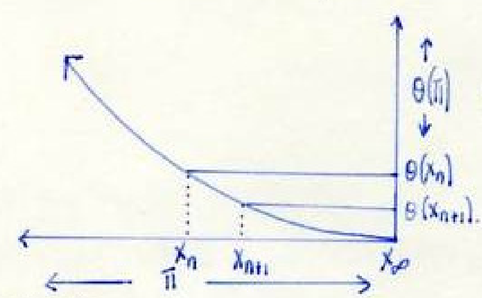
Thus $Pr(x_1 B x_{n+1} | x_n) \geq \prod_{r=2}^{n} Pr(x_{r-1} B x_{r+1} | x_r) > \varrho > 0$ for all $n > 1$.

Remarks 1.11.4.

i) We could use $(1-e^{-\gamma_r})/(1-e^{-3\gamma_r})$ and still get result easily.

ii) We are assuming m is large enough for us to neglect overshoot, which becomes negligibly small as $n \to \infty$. If we included it in $P(x_{r-1} B x_{r+1} | x_r)$, then its effect would be geometrically small when we take $\sum_r a_r$.

Finally we prove that $\varrho = 1$ must hold, in the limit $n \to \infty$.



Let $x_n = 1 - 1/m \cdot \beta^{n-1} = 1 - 1/h$.

$x_{n+1} = 1 - \beta/h$.

Lemma 1.11.5.

$$Pr(x_n B 1 | x_{n+1}) \geq 1 - \exp\left(-z\beta h^{\alpha-1}/\theta\right) \qquad \text{for } z > 0.$$

Proof.

$$P_{\theta(\Pi)}(x_n B 1 | x_{n+1}) \geq P_{\theta'(\Pi)}(0 B 1 | x_{n+1}) \geq P_{\theta(x_n)}(0 B 1 | x_{n+1}) \qquad \text{by 1.4.5.}$$

where $P_{\theta(x_n)}$ denotes probability taken along linear $\theta(x_n)$. Similarly for $P_{\theta(\Pi)}$ along non-linear $\theta(\Pi)$.

And with $\theta'(\Pi) = \theta(\Pi)$ for $\Pi \geq x_n$.
$= \theta(x_n)$ for $\Pi \leq x_n$.

And $P_{\theta'(\Pi)}$ denotes the corresponding probability along $\theta'(\Pi)$.

Thus $P_{\theta(\pi)}(x_n B 1 | x_{n+1}) \geq (1 - \exp(-z(1-x_{n+1})/\theta(1/n))$ by 1.5.4. with super-regular $e^{-z\pi/\theta}$.

and $Pr_{\theta(\pi)}(1 B x_n | x_n) \leq \exp(-z(1-x_{n+1})/\theta(1/n))$. //

Now denote $P(x_n B 1 | x_{n+1}) = b_n$.

We start from x_{n+1} and with probability $1-b_n$ we are absorbed.

Else with probability b_n we reach x_n, and with probability $\geq \xi > 0$ we escape to x_1 before x_{n+1}.

We show ultimate escape is certain.

$\lim_{n\to\infty}$ Pr(absorbed in $\pi=1$ starting at $\pi = 1 - 1/2n$ before reaching x_1).

$= \lim_{n\to\infty} Pr(1 \text{ B } x_1 | x_{n+1}) \leq (1-b_n) + b_n(1-\xi)(1-b_n) + \dots + (1-b_n)(b_n(1-\xi))^r$.

$= \lim_{n\to\infty} (1-b_n)/(1-(1-\xi)b_n)$.

But $\lim_{n\to\infty} b_n = 1$ for $\alpha > 1$ whilst for $\alpha = 1$ this fails and we still need the technique of 1.5.17.

Hence $\lim_{n\to\infty} Pr(x_1 \text{ B } 1 | x_{n+1}) = 1$.

Now the process is semi-martingale with $\Delta\pi < 0$ so # upcrossings $\overset{a.s.}{<} \infty$ and $\pi \to v \in \{0,1\}$ a.s.

To prove absorption in $\pi = 0$ independent of s/mg theorem, we just reverse the above and show $\lim_{n\to\infty} P(\pi = 1/m \text{ B } \pi = 0 | \pi(0) = 1/n) = 0$ for any fixed m, (w.r.t. n) with $m \gg 1$. Then the result follows since the boundaries communicate, as defined in 1.7.4. //

Remark 1.11.6.

The reason for constructing a staircase is that for a dynamic environment we can still obtain an exponentially small drift towards the sub-optimal boundary, for intervals $[x_r, x_{r+1}]$ with $1-\beta$ fixed, yet arbitrarily small. (The optimal boundary will be shown to maximize the average reward.) Then with β fixed and $x_n \uparrow 1$, we have $\theta(x_{r-1}) \ll |x_{r-1} - x_{r+1}|$ for large r and we can use the theory developed by Miller (1962), and subsequently by Keilson and Wishart (1964), for processes defined on markov chains.

Intuitively, letting $1-\beta$ be arbitrarily small yet $\beta<1$ gives $\theta(\pi) \sim$ constant on each "step" and we can ensure that the process asymtotically becomes that of a R.W. , defined on a markov chain, on each "step" as we approach $\bar{\pi} = 1$.

1.12. Dynamic Environments.

We now take $\mathfrak{M}\left(\Delta_{\alpha\beta}, q_{u_i}^{\alpha,s}\right)$ as the environment, with $\underset{\sim}{e}\Delta = \underset{\sim}{e}$ equilibrium vector, and we show that the limiting behaviour of the reinforcement process is only dependent on the ratios $\left\{ \underset{\sim}{e}.\underset{\sim}{q}_i^{s=1} / \underset{\sim}{e}.\underset{\sim}{q}_j^{s=1} \right\}$. The distribution over environment states is denoted by $\underset{\sim}{\omega}(t)$.

Theorem 1.12.1.

i) $\lim_{n\to\infty} U^n \Upsilon^i(\underset{\sim}{\pi}(0), \underset{\sim}{\omega}(0)) = \gamma^i(\underset{\sim}{\pi}(0), \underset{\sim}{\omega}(0))$ where $\Upsilon^i(\pi_i=1, \underset{\sim}{\omega}) = 1$, $\Upsilon^i(\pi_i=0, \underset{\sim}{\omega}) = 0$

ii) $U\gamma^i(\underset{\sim}{\pi}, \underset{\sim}{\omega}) = \gamma^i(\underset{\sim}{\pi}, \underset{\sim}{\omega})$ where $\gamma^i(\pi_i=1, \underset{\sim}{\omega}) = 1$, $\gamma^i(\pi_i=0, \underset{\sim}{\omega}) = 0$.

with U defined as usual:- $U\Upsilon^i(\underset{\sim}{\pi}(t), \underset{\sim}{\omega}(t)) = E\left(\Upsilon^i(\underset{\sim}{\pi}(t+1), \underset{\sim}{\omega}(t+1)) \mid \underset{\sim}{\pi}(t), \underset{\sim}{\omega}(t)\right)$.

Proof.

$$\lim_{n\to\infty} U^n \Upsilon^i(\underset{\sim}{\pi}, \underset{\sim}{\omega}) = 1.\gamma^i(\underset{\sim}{\pi}, \underset{\sim}{\omega}) + 0.(1-\gamma^i) = \gamma^i(\underset{\sim}{\pi}, \underset{\sim}{\omega})$$

$$\text{and } \gamma^i(\underset{\sim}{\pi}, \underset{\sim}{\omega}) = \gamma^i = \lim_{n\to\infty} U^n\gamma^i = \lim_{n\to\infty} U(U^{n-1}\gamma^i) = U\lim_{n\to\infty} U^{n-1}\gamma^i = U\gamma^i.$$

where we can interchange U operator and limit, since $U^n\gamma^i$ converges uniformly to γ^i . //

We now actually write out the U operator explicitly, and so consider the short term behaviour of the process.

Let $\phi(\underset{\sim}{\pi}, \underset{\sim}{\omega})$ be linear in ω_α , $\phi = \sum_{\alpha=1}^{m} \phi_\alpha \omega_\alpha$ and put $\underset{\sim}{\Phi} = (\phi_\alpha)$.

Lemma 1.12.2.

$$\underline{U\underset{\sim}{\Phi}(\underset{\sim}{\pi}) = \sum_{s,i} \pi_i \left(Q_i^s \Delta \underset{\sim}{\Phi}(T_i^s \underset{\sim}{\pi})\right).}$$

with $T_i^0 \underset{\sim}{\pi} = \underset{\sim}{\pi}$

$T_i^1 \underset{\sim}{\pi} = (T_2\pi_1, \ldots T_2\pi_{i-1}, T_1\pi_i, T_2\pi_{i+1}, \ldots T_2\pi_n)$.

With T_i defined as:- $T_2\pi_i = \pi_i(1-\theta(\underset{\sim}{\pi}))$ and $Q_i^s = \begin{pmatrix} q_i^{1,s} & & & 0 \\ & q_i^{2,s} & & \\ & & \ddots & \\ 0 & & & q_i^{m,s} \end{pmatrix}$

$T_1\pi_i = \pi_i + \theta(\underset{\sim}{\pi})(1-\pi_i)$.

<u>Proof.</u>

$$U\phi(\underset{\sim}{\pi}, \underset{\sim}{w}) = \sum_i \pi_i\left((\underset{\sim}{w}.\underset{\sim}{p}_i)\,\phi(\underset{\sim}{\pi}.\ L_i^{s=0}\underset{\sim}{w}) + (\underset{\sim}{w}.\underset{\sim}{q}_i)\,\phi(T_i^1\underset{\sim}{\pi}\ L_i^{s=1}\underset{\sim}{w})\right).$$
$$= \sum_{s,i} \pi_i\left((\underset{\sim}{w}.\underset{\sim}{q}_i^s)\,\phi(T_i^s\underset{\sim}{\pi}\ L_i^s\underset{\sim}{w})\right).$$

with $(L_i^s\underset{\sim}{w})_\alpha = \sum_\beta w_\beta\, q_i^{\beta,s}\Delta_{\beta\alpha} \Big/ \left(\sum_\beta w_\beta q_i^{\beta,s}\right).$

Now linearity gives $L_i^s\underset{\sim}{w} = L^{*s}\underset{\sim}{w} \Big/ \underset{\sim}{q}_i^s\underset{\sim}{w} = \sum_\beta L^{*s}_{\alpha\beta} w_\beta \Big/ \underset{\sim}{q}_i^s\underset{\sim}{w}$

Thus $\underset{\sim}{\Phi}(\underset{\sim}{\pi}) \to \sum_i \pi_i\left(P_i\Delta\underset{\sim}{\Phi} + Q_i\Delta\underset{\sim}{\Phi}(T_i^1\underset{\sim}{\pi})\right).$

$$\to \sum_i \pi_i\, L^{*s}\,\underset{\sim}{\Phi}(T_i^s\underset{\sim}{\pi}).$$

and $U\underset{\sim}{\Phi} = \sum_{s,i,\alpha} \pi_i\left(q_i^{\alpha,s}\Delta_{\alpha\beta}\,\phi_\beta(T_i^s\underset{\sim}{\pi})\right)$ as required. //

Now if we fix the increments $T_i^s\underset{\sim}{\pi}$ we have the process as a R.W. defined on the environmental markov chain. So for 2-actions, mimicing 1.11.1. we define $M(s) \triangleq bQ_1\Delta(e^{s(1-b)}-1) + (1-b)Q_2(e^{-sb}-1) + \Delta - I$ on interval $[a,b]$ with $|a-b| = \epsilon$ say.

Also with $\phi = \sum w_\alpha\phi_\alpha$ and $\underset{\sim}{\Phi} = (\phi_\alpha)$ we let

$$W\underset{\sim}{\Phi}(\underset{\sim}{\pi}) \triangleq b\left(P_1\Delta\underset{\sim}{\Phi} + Q_1\Delta\underset{\sim}{\Phi}(\pi+\theta(1-b))\right) + (1-a)\left(P_2\Delta\underset{\sim}{\Phi} + Q_2\Delta\underset{\sim}{\Phi}(\pi-\theta b)\right).$$

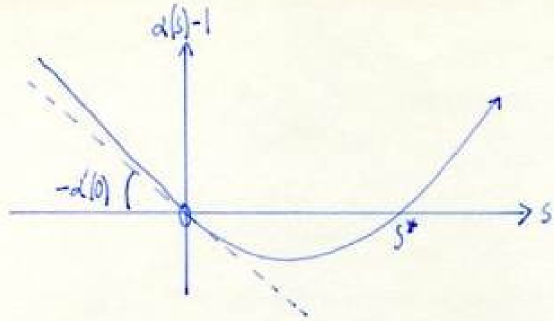
and we try to solve $W\underset{\sim}{\Phi} = \underset{\sim}{\Phi}$ using $\underset{\sim}{\Phi} = \underset{\sim}{R}\,e^{s\pi/\theta}$.

This gives $M(s)\underset{\sim}{R} = 0$ or that $|M(s)| = 0$ gives the relevant solutions. So we have obtained the correspondence between M(s) and the approximation W to the U operator.

Actually $M(s) = P(-s) - I$ where $(P(s))_{ij} = \Delta_{ij}\, m_{ij}(s)$ with θ divided out for convenience, and with $m_{ij}(s)$ = m.g.f. for R.W. increments if environment makes transition $i \to j$.

Miller (1962) uses P(s) as the fundamental matrix and on writing $|P(s) - \alpha(s)I| = 0$ to define $\alpha(s)$, shows that the drift constant, corresponding to A(s) = 1 in scalar case, is given by $\alpha(-s) = 1$ for markov processes.

We have $-m'_{ij}(0)$ is the mean i,j, increment and $m_{jk}(t) = \int_{-\infty}^{\infty} e^{-tx}\,dg_{jk}(x)$ generally, where $dg_{jk}(x)$ gives the increment distribution.



The Solutions of

$\alpha(s) = 1$.

Lemma 1.12.3.

a) $M(0) = \Delta - I$.

b) $d/_{ds} |M(s)| \big|_0 < 0$ iff $\underset{\sim}{e} \cdot \underset{\sim}{q}_1^{s=1} > \underset{\sim}{e} \cdot \underset{\sim}{q}_2^{s=1}$.

c) $\alpha'(0) > 0$ iff $d/_{ds} |M(s)| \big|_0 > 0$.

Proof.

a) is just observation, whilst for b) we have:-

$$d/_{ds} |M(s)| \big|_0 = \operatorname{tr} (\operatorname{adj} (M(s)) \, dM/_{ds}) \big|_0 = b(1-b) \operatorname{tr} (\operatorname{adj} (\Delta - I) N \Delta)$$

where $N = Q_1 - Q_2$.

Thus $d/_{ds} |M(s)| \big|_0 < 0$ iff $\prod_{j \neq 1} (\lambda_j - 1) \operatorname{tr} (\xi_1 \eta_1' N) < 0$

where $\xi_1 \eta_1' = S_1$ the first spectral matrix.

Thus $-\operatorname{tr}(S_1 N) < 0$ iff $\underset{\sim}{e} \cdot \underset{\sim}{q}_1^{s=1} > \underset{\sim}{e} \cdot \underset{\sim}{q}_2^{s=1}$

For c) $d/_{dt} |P(t) - \alpha(t) I| \big|_0 = \operatorname{tr} (\operatorname{adj} (P(0) - \alpha(0) I) \; (P'(t) - \alpha'(t) I))$

and so $\operatorname{tr}(S_1 (P'(0) - \alpha'(0) I)) = 0$ and $\alpha'(0) = \sum_{i,j} e_i \Delta_{ij} m_{ij}'(0)$.

$= -b(1-b) \operatorname{tr}(S_1 N) = \operatorname{tr}(S_1 P'(0))$.

So $\alpha'(0) = -b(1-b) \sum_{\alpha} e_{\alpha} (q_1^{\alpha} - q_2^{\alpha}) = -\sum \text{drift}$ and hence c)

This is essentially the same calculation as for b) with P(s) - I replacing M(s). Note also $P'(0) = M'(0)$ and $\alpha(s) = 1 \iff |M(s)| = 0$.

//

By transferring from U operator to M(s) techniques, we have a close approximation to large iterates of U, when $\overline{\Pi}$ is kept within the small interval $[a,b]$. If we let $\theta(\overline{\Pi}) = O(\epsilon^{(2)})$ when $|a-b| = \epsilon$ then as $\epsilon \downarrow 0$, this approximation becomes exact.

On using $\mathcal{R}_0$ rules with the β-staircase, we have precisely these conditions holding as $\pi \to \nu \in \{0,1\}$ so we need only consider $\alpha(s)=1$ and sign($\alpha'(0)$) to determine the limiting behaviour as in Miller(1962).

We shall now state a central limit theorem of Keilson and Wishart (1964) which gives us the asymtotic independence of environment and π-increment process.

<u>Theorem</u> 1.12.4. (Keilson and Wishart)

If $F_r(x,k) = Pr(R(k)=r,\ X(k) \le x)$ then

$$\lim_{k\to\infty} F(x\sqrt{k}+km,\ k) = \Phi(x\sigma^{-1})\underline{e}$$

where the environment has equilibrium vector $\underline{e}$ and states r. The mean drift $m=-\alpha'(0)$ and variance $\sigma^2 = \alpha''(0) - (\alpha'(0))^2$ which are expressed in terms of P(s) matrix, Keilson and Wishart (1967). In their notation, k is the disrete time and X is the value of the increment process variable.

<u>Proof</u>.

Keilson and Wishart (1964). //

This result allows us to prove that as $\theta\downarrow 0$ we asymtotically follow the mean drift.

<u>Lemma</u> 1.12.5.

i). If $\theta\downarrow 0$ and $k\to\infty$ s.t. $\theta k = K$ = constant, then

$$\lim_{k\to\infty} Pr(|X(k)-km|>y)\to 0 \qquad \text{for all } y>0$$

ii) This limit remains true as $km=Z\downarrow 0$ if $\theta=O(Z^{c>1})$, $y=O(Z)$, $c=$const, and with $\theta=o(y)$ and so K$\downarrow$ 0 also.

<u>Proof</u>.

$$Pr(X(k)-km>y) = Pr((X(k)-km)/\sqrt{k} > y/\sqrt{k}) \sim \int_{y/\sigma\sqrt{k}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz \quad \text{asymtotically.}$$

Now put $m=m'\theta$ and $km=Z=km'\theta$ with $k\theta=K=Z/m'$

Also $\sigma^2=O(\theta^2)$ for $\theta\downarrow 0$ can be veified, say $\sigma=\theta\sigma'$

Then $y/\sigma\sqrt{k} = y/\theta\sigma'\sqrt{k}^{\sqrt{\theta}} = y/\sigma'\sqrt{k\theta}.$

Hence $\lim\limits_{\substack{k\to\infty \\ \theta\to 0}} \Big\} k\theta = K \quad Pr\left(X(k) - km > y\right) = 0$ for $y > 0.$

and result follows by a similar argument for $Pr(X(k) - km < -y).$

Also if $y = O(Z)$ and $\theta = O(Z^{(>1)})$ then $\lim\limits_{Z \downarrow 0} y/\sqrt{Z\theta} = \infty$

and we have justification for the remark after 1.12.3. We asymtotically follow the mean drift even as the interval width $\downarrow$ 0, so long as $\theta \downarrow 0$ more rapidly, as defined above. Z is effectively the distance travelled by the increment process. //

We now just need to assert that $\exists\, s^* < 0$ s.t.

$f(\Pi) > \left(e^{-s^* b/\theta} - e^{-s^* \Pi/\theta}\right) \Big/ \left(e^{-s^* b/\theta} - e^{-s^*(a-b-1)/\theta}\right)$ in the notation of 1.11.1.

Such an s^* is related to the real non-zero root of $|M(s)| = 0$.

Lemma 1.12.6.

$\exists\, s^* < 0$ s.t. $|M(s)| = 0 \Rightarrow s < s^* < 0$ for all b when $\alpha'(0) > 0$.

Proof.

Consider the 2nd deivative of $|M(s)|$. This has an upper bound for all b by observation. Then by Taylor's Theorem, all solutions to $|M(s)| = 0$ with $s \neq 0$ must be bounded strictly away from the origin by some $s^* < 0$. Similarly when $\alpha'(0) < 0$ and $\underline{e}.\underline{q}^1_{-1} > \underline{e}.\underline{q}^1_{1}$ we obtain $\exists\, s^* > 0$, with f(Π) re-defined for "b" absorption. //

Freedman (1973) proves a useful pair of inequalities which give us the same asymtotic mean drift following and his results also hold for finite stopping times. However it has not been possible to actually construct absorption probabilities in the manner of Miller (1962) as this would require Wiener-Hopf factorizations. But since we are not concerned with environment states at absorption, such a full analysis is unnecessary and it is sufficient to just apply 1.12.3. $\rightarrow$ 1.12.6. for reinforcement in a dynamic environment.

Theorem 1.12.7.

a) If $\theta \downarrow 0$ with $\theta = O(\varepsilon^{c>1})$ when $|b-a| = O(\varepsilon)$, c = const.

then $P(a\ B\ b \mid c = (a+b)/2) = O(e^{-s(b-a)/\theta})$ with $s > s^*$ for $a < b$

and s^* as in 1.12.6. where $\alpha'(0) < 0$.

b) $\lim_{\theta \downarrow 0} \gamma(\pi_1) = 1$, $\pi_1 \neq 0$ iff $\underline{\varrho} \cdot \underline{q}_1^{s=1} > \underline{\varrho} \cdot \underline{q}_2^{s=1}$.

Proof.

The bound on s follows from 1.12.6. as $|M(s)| = 0 \iff \alpha(s) = 1$.
The drift towards the sub-optimal boundary becomes exponentially small by 1.12.4. and 1.12.5. , since asymtotically, as $\theta \downarrow 0$ the mean dynamic process becomes indistinguishable from a static process. Miller's paper (1962) is also relevant, except that he considered just one absorbing barrier. However, in this case we obtain P(absorbed at the barrier) $\sim O(e^{-ys/\theta})$ where $\alpha(s) = 1$ and $y \triangleq$ distance from barrier, with drift away from it, and $y/\theta \gg 1$, $s > 0$.

For b) we apply a) in each interval $[a,b]$, and then piecing intervals together as in the staircase theorem.

Thus $P(1\,B\,0 \mid \underset{\sim}{\pi}(0) = \pi_1 \in (0,1)) \geq \lim_{n \to \infty} \prod_{r=1}^{n} (1 - e^{-sn^{c-1}}) = 1.$

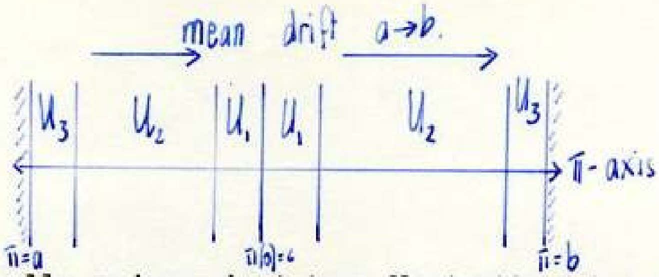
when $(b-a) = 1/n$ and $\theta = (1/n)^{c>1}$.

//

Here we are interested only in the asymtotic behaviour when the environment and π-process become independent as $\theta \downarrow 0$, so we ignore the short term fluctuations of 1.12.2.

Note that in 1.12.7. a) that we will actually have $P(aBb \mid c)$ exponentially small for all $c \in [a,b]$ as $\theta \downarrow 0$ and the absorption probabilities asymtotically reduce to the form obtained for the static medium in 1.11.1. This is required in our β-staircase construction for dynamic environments.

Remark 1.12.8.



Diagrammatically we have short term fluctuations in region U_1, whilst in U_2 we follow mean drift with arbirarily high probability as $\theta \downarrow 0$. Then in U_3 we could consider the environmental state for the full analysis, whereas we just require $P(a\ B\ b|c)$ for our reinforcement rules.

Theorem 1.12.9.

Under R_0, $\pi_i \to 1$ iff $\underline{e}\cdot\underline{q}_i > \underline{e}\cdot\underline{q}_j \quad \forall j \neq i$ with $\pi_i(0) > 0$

Proof.

First the result for 2-actions follows from the β-staircase construction 1.11.3. , since the approximation for $P(x_{i-1} B x_{i+1} | x_i)$ still holds with drift constant $s > 0 \ \forall b \in [0,1]$.

So $\lim_{n\to\infty} P(x_1\ B\ x_{n+1} | x_n) > \zeta > 0$ holds with $1-\beta$ arbitrarily small since 1.12.5. ii) is found to be satisfied, and also 1.12.7. a) holds for each "step" as we approach any boundary.

Then as before we get $\lim_{n\to\infty} P(x_1 B\ 1\ | x_n) = 1$, when drift is away from the boundary, and hence the result.

For n-actions we just use the $\theta_{ij}(\underline{\pi})$ rules to give asymtotic reflection from sub-optimal boundaries, using a comparison argument analogous to 1.6.4. to take us from n-actions to a 2-action model, when we can then apply the staircase result above. //

Remark 1.12.10.

i) A π-cell learning under R_0 asymtotically takes that action which maximizes the average payoff in a dynamic $\mathfrak{M}(\Delta_{\alpha\beta}, q^{a.s}_{u_i})$.

ii) We use 1.12.4. and 1.12.9. in chapter 3. There we find that a network of π-cells asymtotically maximizes its payoff at the next trial w.r.t. the equilibrium environmental distribution in its present state.

1.13. Learning Barriers.

We know from 1.7. that $\mathcal{R}_0$ is boundary learning and $\mathcal{R}_C$ is centrally learning, yet in this section we show that there is still a unification in the underlying learning mechanism.

Definition 1.13.1.

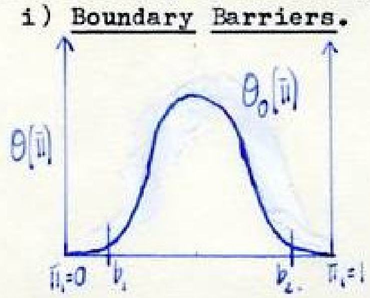
A learning barrier $\underline{\underline{a}}$ $[a,b] \subset I$ s.t. $|a-b| = \epsilon$ and $\theta(\pi) = O(\epsilon^{c>1})$, $\pi \in [a,b]$.

This section will consist of numbered remarks relating to this concept.

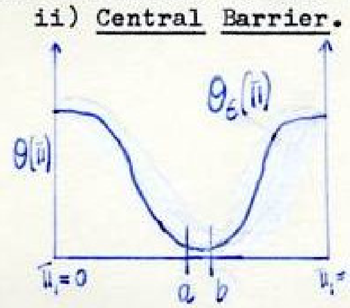
1.13.2. For boundary learning we put a barrier around each $\pi_i = 0$, whilst for central rules we place a barrier away from all boundaries. From our previous theory, when the equilibrium drift is from $a \to b$, $P(aBb \mid c = \frac{1}{2}(a+b)) < e^{-z(b-a)/\theta}$, for $z > 0$, and independent of θ and $[a,b]$. This follows from weak convergence to a diffusion as $\theta \downarrow 0$ and 1.12.5. , 1.12.7. for dynamic $\mathcal{M}$. For optimality we have a learning barrier in any arbitrarily small interval of all boundaries, and this provides the mechanism for the asymtotic reflection from sub-optimal boundaries.

1.13.3. The barrier is a potential step and acts as a form of "diode", in that we can only pass through the barrier with high probability if we are travelling with the equilibrium drift. The effective "field strength" is $\partial J(\pi)/\partial \pi \, \hat{\underline{x}} = \underline{E}$ and at barriers placed centrally or at sub-optimal boundaries we have $|\underline{E}| \uparrow \infty$.

i) Boundary Barriers.



ii) Central Barrier.



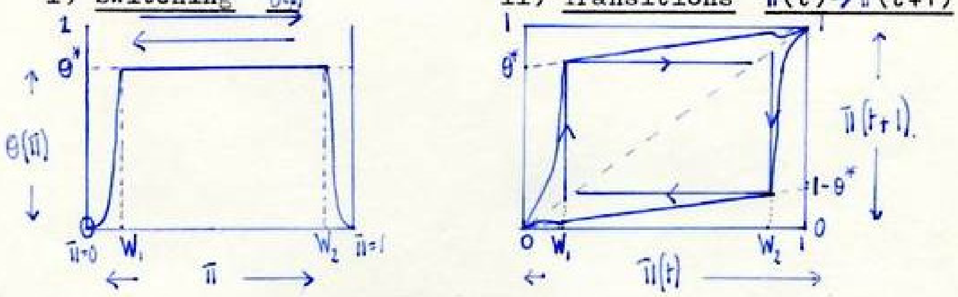
1.13.4. We could put $\theta_0(\pi) + \theta_C(\pi) =$ constant, to see that $\mathcal{R}_0$ and $\mathcal{R}_C$ may be viewed as complementary ways of learning, yet they both operate using the learning barrier mechanism.

1.13.5. For optimal boundary learning rules, we effectively just have a deterministic stability theorem to satisfy. So we could actually reformulate the theory in terms of control theoretic terminology to give switching between attractors through a catasrophe until we find a stable attractor.

When we consider networks of π-cells, this would entail an automaton increasing its environmental adaptation through structural catasrophes as in the work of Thom (1975). The automaton could then increase both its memory depth and number of actions used. (see chapter 3).

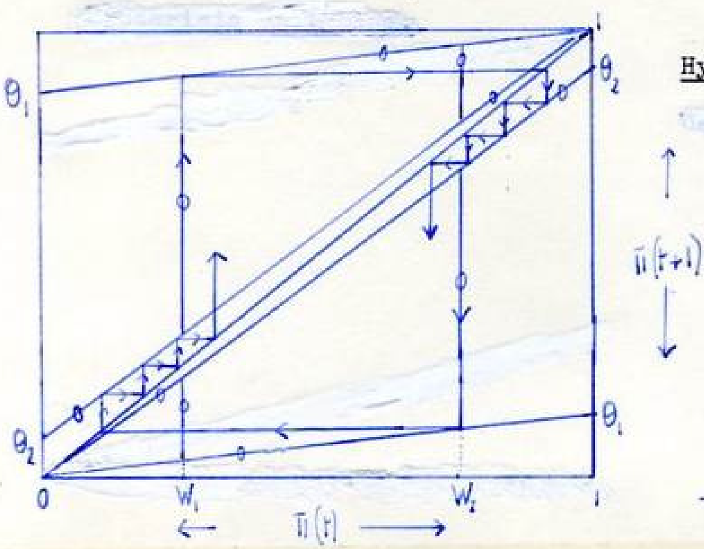
1.13.6. The advantage of stochastic automata is that we have an explicit mechanism for incorporating environmental information into the structure. Thus a π-cell incorporates both a learning mechanism and boundary switching.

i) Switching $\theta(\pi)$ ii) Transitions $\pi(t) \to \pi(t+1)$



We can choose θ^* for the switching using the transition diagram. Now, if W_1 and W_2 are arbitrarily close to their respective boundaries the π-cell just has the boundary learning barriers which are stable only if they give maximum average payoff.

1.13.7. We can also easily define $\theta(\pi)$ to give a hysteresis effect in its switching, still retaining the U.L. property, thus mimicing the switching of the cusp catasrophe. This effect is closely related to the grid 1.10.1. , and with $W_1+W_2 = 1$, we obtain an "overshoot" on boundary switching if and only if $W_2 < 1/(2-\theta^*)$.



Hysteresis Switching.

Use θ_1 and θ_2 with θ_1 near 1 and θ_2 near 0, s.t. $\theta_2 << 1-\theta_1$. This gives R_{EB} learning rule.

We have:-

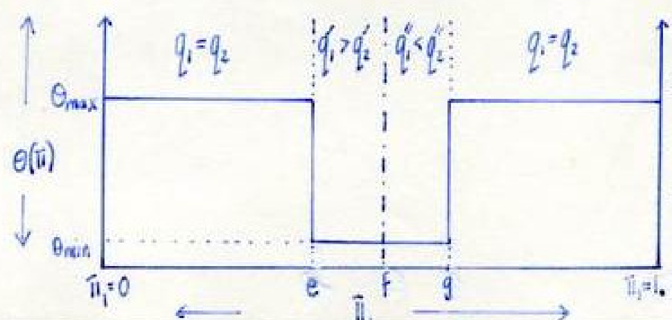
a) $\theta(\pi) = \theta_1$, $\pi \in [W_1 W_2]$.

b) $\theta(\pi) = \theta_2$, $\pi \in [0, W_1) \cup (W_2, 1]$.

—o— denotes transition line $\pi(t) \to \pi(t+1)$.

1.13.8. In π-cell games and π-cell networks, we find that the effective stimulus probabilities are dependent on the action distribution, so that $\underset{\sim}{q} = \underset{\sim}{q}(\underset{\sim}{\pi}(t))$. This can give rise to a potential barrier between strictly stable boundaries. Then, π-cell switching behaviour between stable limiting structures can be likened to the "tunnelling" of charged particles in quantum theory. In 2.4.5. we briefly examine one of the simplest cases of 2 strictly stable boundaries, which represent Nash Points of π-cell games.

A potential barrier is formed on placing two learning barriers, (potential steps) "back to back" as shown in a particular $\underset{\sim}{q}(\underset{\sim}{\pi})$ process below.



Potential Barrier.

$LB_1 = \{\pi_1 : \pi_1 \in [e, f]\}$

$LB_2 = \{\pi_1 : \pi_1 \in [f, g]\}$.

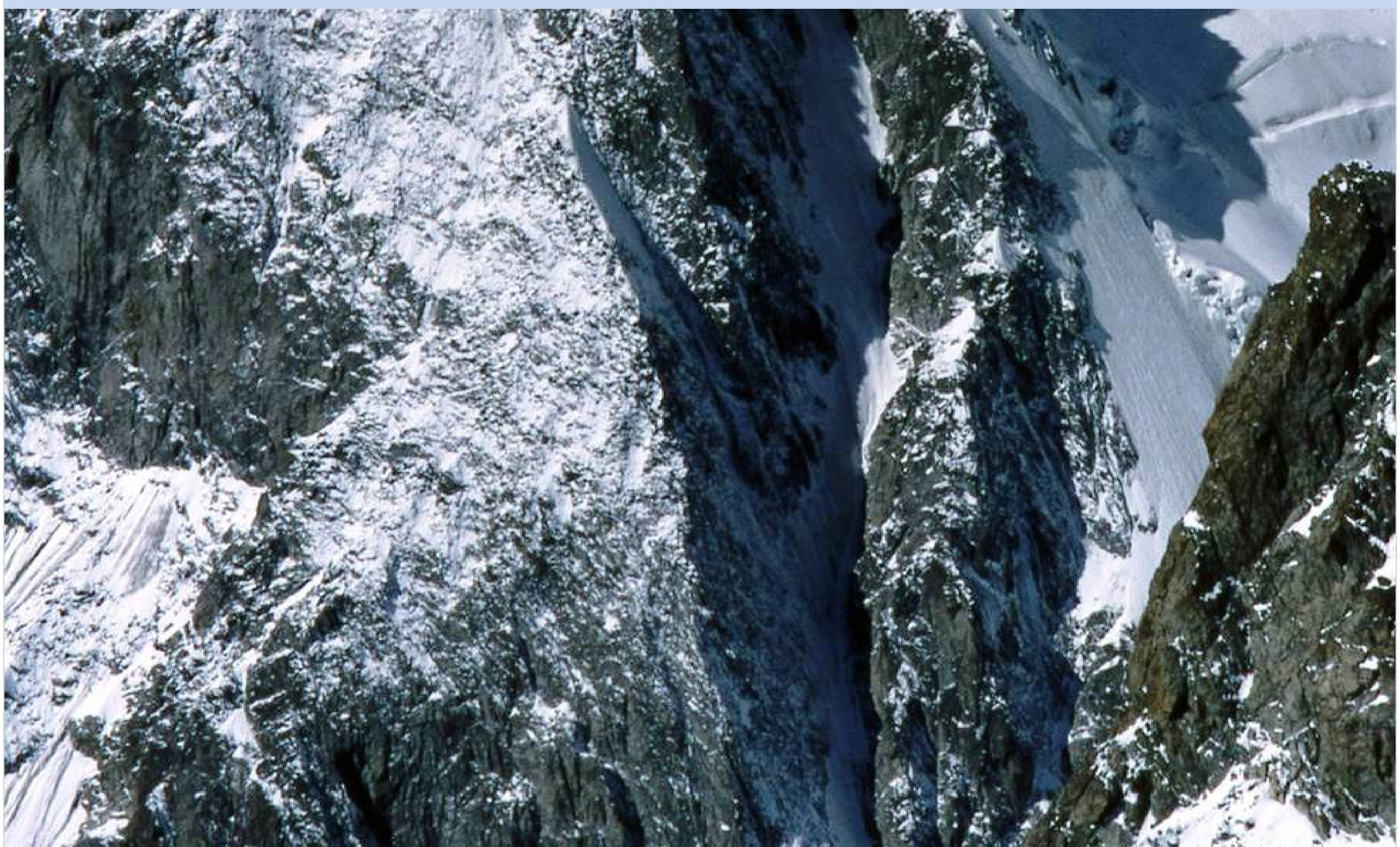
where LB_i denotes learning barrier.

$PB = \{\pi_1 : \pi_1 \in [e, g]\}$.

1.13.9. Our reinforcement rules R_0 depend only on the boundary learning barriers for their optimality, which could perhaps be thought of as stochastic attractors and repellers. This concept of learning barriers and boundary switching is important when we discuss structured automata, and is the basis for the main characterization theorems.

Chapter Two

Games between Automata



Chapter 2.

There is no remembrance of former things, nor will there be any remembrance of later things yet to happen among those who come after.

Ecclesiastes 1 v 11.

2. Games between Unstructured Automata.

2.1. The Model for Π-cell Games.

In the previous chapter we considered the Π-cell playing against an environment $\mathcal{M}$. It is now natural to formulate a game between Π-cells so that for a particular Π-cell $\otimes_i$ the environment is the remaining $\{\otimes_{j\neq i}\}$. These games were introduced by Chandrasekaren and Shen (1968) and developed to a limited extent by Narendra and Viswanathan (1974), who first demonstrated, through computer simulation, the possibility of cyclic behaviour.

We shall first develop the theory for zero-sum 2-automata games to indicate the main features, before considering general sum n-automata games.

Definitions 2.1.1.

i) Let g_{ij} = expected winnings to automaton 1 taking u_i when automaton 2 takes u_j.

and Π_i^α = Pr(automaton α takes u_i).

ii) Let p_{ij} = Pr(under actions (i,j), automaton 1 receives penalty -1).

= Pr(" " " " 2 " reward +1).

$$\text{with } q_{ij} + p_{ij} = 1 \text{ and } g_{ij} = q_{ij} - p_{ij} = 1 - 2p_{ij}.$$

iii) Let p_i^1 = Pr(automaton 1 receives s=0, using u_i) $= \sum_j p_{ij}\Pi_j^2$.

$$= \tfrac{1}{2}(1 - \sum_j g_{ij}\Pi_j^2).$$

and similarly $p_j^2 = \sum_k \Pi_k^1(1 - p_{kj}) = \tfrac{1}{2}(1 + \sum_k \Pi_k^1 g_{kj})$.

The above definitions hold for an arbitrary automaton, so now we restrict the study to Π-cells, $\otimes_i$, with environment determined by $p_j^i(\{\Pi^{k\neq i}(t)\})$. The game is zero-sum in expectation rather than deterministically.

2.2. Pure Saddles.

We shall use the usual game theoretic terminology and first consider the case of a pure saddle point. Using the optimal boundary learning theory of 1.7. , the following result is almost immediate.

Theorem 2.2.1.

For 2 π-cells acting under R_0 and if

$$\max_m \min_n g_{mn} = \min_n \max_m g_{mn}$$

then $\lim_{t\to\infty} \pi_i^1(t) = 1$ and $\lim_{t\to\infty} \pi_j^2(t) = 1$ only if (i,j) is a ~~unique~~ pure saddle of g_{ij}.

Proof.

We assume that we have a strict saddle, in that $g_{rj} < g_{ij} < g_{ik}$ with $r \neq i$, $k \neq j$, and hence that the pure saddle is unique. However, the proof is easily modified to take care of non-uniqueness, as in 1.6.4. and also in the general Nash Theorem 2.5.1.

As before we use $\Delta\pi_i^k(t) = \mathcal{E}\left(\pi_i^k(t+1) \mid \underset{\sim}{\pi}^\alpha(t) \text{ for each } \alpha\right) - \pi_i^k(t)$ to express the conditional increment.

Then $\Delta\pi_i^1(t) = \pi_i^1/2 \sum_{j,k} \theta_{ij}(\underset{\sim}{\pi}^1(t))\, \pi_j^1(t)\, (g_{ik} - g_{jk})\, \pi_k^2(t)$ 1)

and $\Delta\pi_j^2(t) = \pi_j^2/2 \sum_{i,k} \theta_{ij}(\underset{\sim}{\pi}^2(t))\, \pi_i^2(t)\, (g_{ki} - g_{kj})\, \pi_k^1(t)$ 2)

a) We shall first prove convergence. The process can only be absorbed at a boundary $\pi_i^k \in \{0,1\}\ \forall i,k$ where $\theta_{ij}(\underset{\sim}{\pi}^k) = 0$. Now we apply 1.7.5. and assume w.l.o.g. that (1,1) is the saddle.

In a small neighbourhood N_{11} of (1,1), using 1) we obtain:-

$$\Delta\pi_1^1 = \tfrac{1}{2}\pi_1^1 \sum_j \theta_{ij}(\underset{\sim}{\pi}^1)\, \pi_j^1\, (g_{11} - g_{j1})\, \pi_1^2 + O\left(\max_{k\neq 1} \sum_j \theta_{ij}(\underset{\sim}{\pi}^1)\, \pi_j^1\, (g_{ik} - g_{jk})\, \pi_k^2\right) \quad 3)$$

Then we take ϵ s.t. $\left.\begin{array}{l}\pi_1^1 > 1-\epsilon \\ \text{and } \pi_1^2 > 1-\epsilon\end{array}\right\}$ gives $\begin{array}{l}\Delta\pi_1^1 > 0 \\ \Delta\pi_1^2 > 0\end{array}$ in N_{11}.

Since $\theta_{ij}(\underset{\sim}{\pi}^k) > 0$ away from boundaries, we are either absorbed at $(i,j) \neq (1,1)$, or else we eventually enter N_{11} where π_1^1 and π_1^2 are sub-martingales. We modify the argument of 1.7.4. so that the process is stopped if either $\underset{\sim}{\pi}^1$ or $\underset{\sim}{\pi}^2$ leaves N_{11}.

If the process were not convergent, we would enter N_{11} i.o. , but for a s/mg, # upcrossings $\overset{a.s.}{<} \infty$, and we get a contradiction.

Hence if the process cannot be absorbed at any $(i,j) \neq (1,1)$, we converge to $(1,1)$.

b) Suppose $\lim_{t\to\infty} \pi^1_{j\neq 1} = 1$ and $\lim_{t\to\infty} \pi^2_k = 1$ with $(j,k) \neq$ saddle.

Then either $\exists$ n s.t. $g_{nk} > g_{jk}$ or m s.t. $g_{jm} < g_{jk}$.
Assume w.l.o.g. that $\exists$ $g_{nk} > g_{jk}$ then in neighbourhood N_{jk} of boundary point (j,k) we have:-

$$\Delta\pi^1_n = \tfrac{1}{2}\pi^1_n \theta_{nj}(\underline{\pi}^1)\pi^1_j (g_{nk}-g_{jk})\pi^2_k + O\Big(\max_{m\neq k} \sum_j \theta_{nj}\pi^1_j \pi^2_m (g_{nm}-g_{jm})\Big). \qquad 4)$$

and take $\pi^1_j > 1-\epsilon'$, $\pi^2_k > 1-\epsilon'$ s.t. $\Delta\pi^1_n > 0$ in N_{jk}.

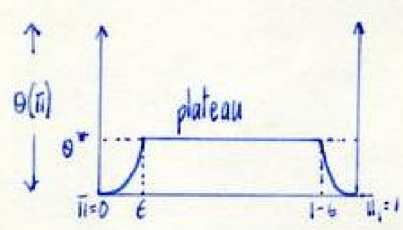
We have $q^1_i(\underline{\pi}^2(t)) = (1+g_{ik})/2 + O(\epsilon')$ in N_{jk}, and $q^1_n(\underline{\pi}^2(t)) > q^1_j(\underline{\pi}^2(t))$.

But the learning function $\theta_{ij}(\underline{\pi}^k)$ is optimal and hence boundary N_{jk} is probabilistically reflecting by 1.7.5., since all boundaries communicate.

Hence by a) $\lim_{t\to\infty} \pi^1_i = 1$ and $\lim_{t\to\infty} \pi^2_j = 1$ only if (i,j) is a pure saddle. //

For π-cell games it is useful to define boundary learning rules with $\theta_{ij}(\underline{\pi}) = \theta =$ constant away from boundaries, giving a centrally learning plateau for mixed strategy trajectories.
Consider the case of 2-actions and R_θ rule $\theta(\underline{\pi})$.

Define $R_{\theta M}$ rule s.t. $\theta(\underline{\pi}) = \theta(\pi_1\pi_2)^{\alpha>1}$ for $\pi_i < \epsilon$ else $\theta(\pi) = \theta^* =$ constant so $\theta^* = \theta(\epsilon(1-\epsilon))^\alpha$



Similarly for n-actions with $\theta_{ij}(\underline{\pi})$

We then obtain
$$\left.\begin{aligned} \Delta\pi^1_i &= \theta^*\pi^1_i/2\Big(\sum_j g_{ij}\pi^2_j - V(\underline{\pi}^k)\Big) \\ \text{and}\quad \Delta\pi^2_j &= -\theta^*\pi^2_j/2\Big(\sum_i \pi^1_i g_{ij} - V(\underline{\pi}^k)\Big) \end{aligned}\right\}$$ away from boundaries.

where $V(\underline{\pi}^k) = \sum_{i,j} \pi^1_i g_{ij} \pi^2_j$ = value of the game.

2.3. Mixed Strategies.

The analysis of games without pure saddles is both more difficult and yet more interesting, in that analogies arise naturally with population processes. Indeed, we can now reformulate such processes as games between species in an attempt to give insight to operating mechanisms. Biologists have recently been searching for such a formulation, as in the paper of Maynard-Smith (1973).

Also from the Hardy-Weinberg equations of mathematical genetics we have $p' = (p^2 w_{11} + 2pqw_{12})/W$ where p = frequency of genotype a_1

$q = 1-p$ = " " " a_2

w_{ij} = selective viability of $a_i a_j$

and $W = p^2 w_{11} + 2pqw_{12} + q^2 w_{22}$.

and $p=p'$ when $u = p'/q' = (w_{12}-w_{22})/(w_{12}-w_{11}) = p/q$.

This equilibrium point is precisely that obtained if the process is viewed as a game with game matrix w_{ij}. Although here each "player" is constrained to have the same distribution over the genotype strategies a_1 and a_2.

In this thesis we shall restrct ourselves to the theoretical π-cell framework. We shall examine 2x2 games, although, as in chapter 1, many results can be extended to n-actions. The next lemma indicates the reason for cyclic behaviour, under any U.L. rule.

Lemma 2.3.1.

If there is no pure saddle, then $\Delta\pi^1_i = \Delta\pi^2_j = 0$ iff $(\pi^1_i, \pi^2_j) = \lambda^i_j$ is the Von Neumann value of the game.

Proof.

$\Delta\pi^2_i = \Delta\pi^1_i = 0 \ \forall i \Rightarrow V(\underset{\sim}{\pi}^k) = \sum_i \pi^1_i g_{ir} = \sum_k g_{jk} \pi^2_k, \forall r, j$, but this has 5.)

solution λ^i_j, the V-N optimal mixed strategy.

We find $\lambda_1^1 = 1 / (1 + (g_{11} - g_{12}) / (g_{22} - g_{21}))$ and $\lambda_1^1 / \lambda_2 = \left(\frac{g_{21} - g_{22}}{g_{12} - g_{11}} \right)$

$$\lambda_1^2 = 1 / (1 + (g_{21} - g_{11}) / (g_{12} - g_{22})).$$

with $\lambda_j^i \in (0,1)$ //

We shall see that the λ_j^i also generate the deterministic trajectories for n-actions for a linear U.L. rule, where λ_j^i satisfy the equations 5), of 2.3.1. In practice, given an arbitrary g_{ij}, we should use θ_{0M} with ϵ sufficiently small to give λ_j^i lying on the central plateau of $\theta_{ij}(\underline{\Pi})$. The $\theta_{ij}(\underline{\Pi})$ were chosen to give boundary optimality at pure saddles, whilst $\theta(\underline{\Pi})$ rules give the homogenous central learning for mixed strategies. We shall consider the form of deterministic trajectories around and we find a naturally arising diffusion of a constant. Such processes are analysed by Barbour (1973).

For the 2x2 game we use $\Pi_1^1 = x$, $\Pi_1^2 = y$. $\underline{\omega} = (\Delta x, \Delta y)$, $\underline{\lambda} = (\lambda_x, \lambda_y)$ with $\lambda_x = \lambda_1^1$, $\lambda_y = \lambda_1^2$.

Lemma 2.3.2.

For $\underline{\lambda} \in [0,1]^2$ and $\underline{x} = (x,y) \in [0,1]^2$ then $(\underline{x} - \underline{\lambda}).\underline{\omega} = 0$

iff $x = \lambda_x$ or $y = \lambda_y$ or $x = y$ or $x + y = 1$.

Proof.

$$\left.\begin{aligned} \Delta x &= \theta_x x (g_{11} y + g_{12}(1-y) - V(\underline{x})) \\ \Delta y &= \theta_y y (V(\underline{x}) - x g_{11} - (1-x) g_{21}) \end{aligned}\right\} \text{ with } \begin{aligned} \theta_x &= \theta x^\alpha (1-x)^\alpha \\ \theta_y &= \theta y^\alpha (1-y)^\alpha. \end{aligned}$$

and putting in λ_x and λ_y we obtain:-

$$(x - \lambda_x)\Delta x = \beta \theta_x x (1-x)$$
$$(y - \lambda_y)\Delta y = -\beta \theta_y y (1-y).$$

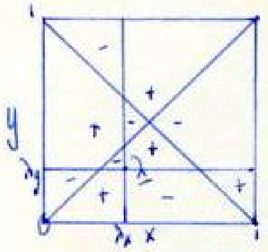
where $\beta(\underline{x}, g_{ij}) = (x - \lambda_x)(y - \lambda_y)((g_{11} + g_{22}) - (g_{12} + g_{21})) = \gamma (x - \lambda_x)(y - \lambda_y)$

$$(\underline{x} - \underline{\lambda}).\underline{\omega} = \theta\beta \left(x^{\alpha+1}(1-x)^{\alpha+1} - y^{\alpha+1}(1-y)^{\alpha+1} \right) = \theta\beta(x-y)(1-(x+y))(f(x,y)).$$

with $f(x,y) > 0$ for $(x,y) \in [0,1]^2$.

Now if $g_{11} + g_{22} = g_{12} + g_{21}$ we get $1/\lambda_x = 1/\lambda_y = 0$.

Hence the only roots are those stated.



It is interesting to note that on the boundary ∂I^2, symmetry gives us sgn$((\underline{x}-\underline{\lambda})\cdot\underline{\omega})$ is +ve precisely half its length.

We also have $V(\underline{x}) - V_{opt} = \beta$. and $\underline{\nabla} V = \gamma((y-\lambda_y), (x-\lambda_x))$

so $\beta = (V_x V_y/\gamma)$

Then $(\underline{x}-\underline{\lambda}).\underline{\omega} = \theta\,(V_x V_y/\gamma)\,(x-y)(1-(x+y))\,f(x,y)$. //

With our intuition strengthened by the previous lemma, we shall consider simple properties of the deterministic trajectories.

<u>Theorem</u> 2.3.3.

If deterministically we set

$$dx/dt = \gamma\theta_x(y-\lambda_y)x(1-x) = \theta_x V_x\, x(1-x).$$
$$dy/dt = -\gamma\theta_y(x-\lambda_x)y(1-y) = -\theta_y V_y\, y(1-y).$$

with $\theta_x = \theta x^\alpha(1-x)^\alpha$ and $k = \alpha+1$

Then:- a) The trajectories are $\Phi^* = f(x)\,f(y)$ =constant

where $\log f(x) = \int (x-\lambda_x)/(x(1-x))^k\theta \; dx$.

b) The trajectories are periodic, and also if $(2\lambda-1)^2 > (2k-1)/k$ for $\lambda = \lambda_x$ and $\lambda = \lambda_y$

then all Φ^* are convex.

c) $\int_{\Phi^*=c} V(\underline{x})\,dt \Big/ \int_{\Phi^*=c} dt = \overline{V(\underline{x})} = V_{opt}$.

<u>Proof</u>.

a) Suppose $\exists\ \Phi(x,y)$ s.t. $d\Phi(x,y) = 0$ expressing "energy conservation".

On $\underline{\omega}$ trajectories $d\Phi = \Phi_x \dot{x} + \Phi_y \dot{y} = 0$.

and so $\Phi_x(y-\lambda_y)/\theta_y y(1-y) = \Phi_y(x-\lambda_x)/\theta_x x(1-x)$

and put $\Phi_y = (y-\lambda_y)/\theta y^k(1-y)^k$ and similarly for x.

Then let $\Phi^* = \exp \Phi = f(x)f(y)$ =constant, as required

If we put $\theta_x^* = \theta x^{\alpha+1}(1-x)^{\alpha+1}$ then $\dot{x} = \theta_x^* V_x$
$\dot{y} = -\theta_y^* V_y$

and $f(x) = \exp \int V_y / \theta\theta_x^* \, dx$

In particular for k=1 $f(x) = x^{\lambda x}(1-x)^{1-\lambda_x}$.

and k=2 $\log f(x) = \lambda x/x + (1-\lambda_x)/1-x + (1-2\lambda x)\log(x/1-x)$.

b) For periodicity the result follows as in Volterra's equations as $f(x)$ is bounded. Goel et Al (1971) give a proof.

For convexity we examine 2nd derivatives.

$$\dot{y} = -(x-\lambda_x)/(y-\lambda_y) \quad (y(1-y)/x(1-x))^k.$$

and $$y'' = (x-\lambda_x)^2/(y-\lambda_y) \quad (y(1-y)/x^2(1-x)^2)^k \, (h(x)+h(y)).$$

where $$h(x) = g(x)(x(1-x))^k/(x-\lambda_x)^2.$$

and $$g(x) = (x-\lambda_x)k(1-2x) - x(1-x)$$

Now $g(x) > 0$ for some x iff $(1-2\lambda_x)^2 < (2k-1)/k$

Hence for $g(x)<0, \; g(y)<0$, y'' has the appropriate sign, under given condition.

In particular for k=1 we have convex Φ^*, whilst for k=3, say, all Φ^* are convex if $0{\cdot}067 < \lambda_i < 0{\cdot}933$.

Since $\underline{\omega}$ is tangential to Φ^*, we have expected outward drift, stochastically, for convex Φ^*, for all $\underline{x} \in I^2$.

c) We have $\int_{\Phi^*=c} \Delta\Pi_i^1 = 0$ using our old notation.

Thus $$\int_{\Phi^*=c} \Delta\Pi_i^1/\Pi_i^1\theta(\Pi_i^1) = -\int_0^T \left(\sum_j \Pi_i^1 g_{ij}\Pi_j^2 - \sum_k g_{ik}\Pi_k^2\right)dt = 0.$$

and hence $$\sum_k g_{jk}\bar{\Pi}_k^2 = \sum_{i,j}\overline{\Pi_i^1 g_{ij}\Pi_j^2} = \bar{V} = \sum \bar{\Pi}_h^1 g_{hj}$$

But solutions to this are uniquely $\bar{\pi}_j^i = \lambda_j^i$, where $\bar{\pi}_j^i$ denotes time-averaged π_j^i on the deterministic, continuous time trajectories.

Hence $\bar{V} = \sum_{i,j} \lambda_i^1 g_{ij} \lambda_j^2 = V_{opt}$. //

Corollory 2.3.4.

For n-actions and k=1, the trajectories are given by:-

a) $\Phi^* = \prod_{i,j} (\pi_j^i)^{\lambda_j^i} = constant.$

If the optimal strategy is completely mixed, $\lambda_j^i \in (0,1)\ \forall i,j$, then:-

b) $\lim_{T\to\infty} \int_{0\ \text{with}\ \Phi^*=c}^{T} V(\underline{\pi}^k)\,dt \Big/ \int_0^T dt = V_{opt}$

Proof.

a) We verify that $d\Phi^*=0$ on the trajectories.

$$d\Phi = \sum \lambda_j^i\, d\pi_j^i / \pi_j^i \qquad \Phi = \log \Phi^*$$

$$\Delta\pi_j^1 = -\,\theta/2\ \pi_j^1 \Big(\sum_{i,j} \pi_i^1 g_{ij} \pi_j^2 - \sum_k g_{jk} \pi_k^2 \Big)$$

$$\Delta\pi_j^2 = \theta/2\ \pi_j^2 \Big(\sum \pi_k^1 g_{kj} - \sum_{i,j} \pi_i^1 g_{ij} \pi_j^2 \Big)$$

Hence deterministically

$$d\Phi = -\tfrac{1}{2} \sum_j \lambda_j^1 \Big(V(\underline{\pi}^k) - \sum_k g_{jk} \pi_k^2 \Big) + \tfrac{1}{2} \sum_j \lambda_j^2 \Big(V(\underline{\pi}^k) - \sum_k \pi_k^1 g_{kj} \Big) = 0$$

since $\sum \lambda_j^1 = \sum \lambda_j^2 = 1$ and by definition $\sum_j \lambda_j^1 g_{jk} = \sum_j g_{kj} \lambda_j^2 = V_{opt}$.

b) If $\lambda_j^i \in (0,1)\ \forall i,j$ then $\Phi^* = const$ will be closed hypersurfaces around $\underline{\lambda}$. Again see Goel et Al (1971) for the treatment of similar trajectories, using analogies with statistical mechanics.

Now $$\tfrac{1}{T} \int_{0\ \Phi^*=c}^{T} 2\,d\pi_j^1 / \pi_j^1 \theta = K/T \big(\log \pi_j^1(T) - \log \pi_j^1(0) \big) \qquad K = const$$

$$= \tfrac{1}{T} \int_0^T \Big(\sum_k g_{jk} \pi_k^2 - V(\underline{\pi}) \Big).\,dt$$

But for $\Phi^* = c$, $\big| \log (\pi_j^i(T)) \big| / T \to 0$ as $T \to \infty$

Thus by the same argument as before $\bar{\pi}_j^i = \lambda_j^i$

and $\bar{V}(\underline{\pi}^k) = V_{opt}$ where time average is taken as $T \to \infty$. //

Lemma 2.3.5.

For small oscillations about λ with $\theta_x = \theta x^\alpha (1-x)^\alpha$

a) $\int_{\Phi^*} dt \sim 2\pi / \theta \left(\prod_{i,j} \lambda_j^i\right)^{k/2}$

b) We have ellipses to first order in $\epsilon_x = (x-\lambda_x)$, $\epsilon_y = (y-\lambda_y)$

with ratio (major/minor) axis $\sim \left(\prod_i \lambda_i^1 / \prod_i \lambda_i^2\right)^{k/2}$

Proof.

$d\epsilon_x/dt = \theta_x' \epsilon_y$ with $\theta_x' = \theta \lambda_x^k (1-\lambda_x)^k$.

and similarly for y, giving $\epsilon_x'' = -\theta_x' \theta_y' \epsilon_x$.

So period $= 2\pi / (\theta_x' \theta_y')^{1/2}$ and $\epsilon_x = A\cos\left((\theta_x'\theta_y')^{1/2} t\right)$
$\epsilon_y = B\sin\left((\theta_x'\theta_y')^{1/2} t\right)$

with $A/B = (\theta_x'/\theta_y')^{1/2}$.

For $\lambda_x = \lambda_y$ or $\lambda_x = 1-\lambda_y$ we have circles.
We also see that $\int_{\Phi=c} dt \propto 1/\theta$ for all trajectories and k. //

The pursual of the analysis of mixed strategy games leads to many difficulties, unless we are to take "large population approximations" about the Φ =const. However, I shall sketch why I believe superfluous strategies may vanish deterministically.

Sketch 2.3.6.

Let $\bar{\pi}_j^i = \lambda_j^i$ be optimal mixed strategies with $\lambda_s^r = 0$ for some r,s including, say, $\lambda_1^1 = 0$. Then if $\pi_s^r < \epsilon$ $\forall r,s$ s.t. $\lambda_s^r = 0$ then for large T we obtain $\frac{1}{T}\int_{\text{trajectory}} 2\, d\pi_1^1 / \pi_1^1 \theta = -\left(\bar{V}(\pi) - \sum_j g_{1j} \bar{\pi}_j^2\right) < 0.$

By definition, a superfluous strategy has $\sum_j g_{ij} \pi_j^2 < V$ at the saddle and $\bar{\pi}_j^i \doteqdot \lambda_j^i$ $\bar{V}(\pi) \doteqdot V_{opt}$ near Φ^* =const. Thus $\pi_1^1(T) < \pi_1^1(0)$. and $\pi_1^1 = 0$ is stable. A more rigorous treatment may show this naive argument to be false. //

Although we may begin with $\Phi^* = \prod (\pi_j^i)^{(\lambda_j^i)^*}$ as trajectories with some i,j, s.t. $(\lambda_j^i)^* \notin [0,1]$ we hope asymtotically to obtain the optimal

(λ_j^i) of the reduced square matrix. It seems plausible that π-cells may automatically proceed through an algorithm of the type described by Karlin (1959 p50-51), in their attempt to maximize payoff at each trial.

It is precisely the U.L. property which enables π-cells to discrminate between arbitrarily small differences in payoff (as $\theta \downarrow 0$) and hence give such natural trajectories around $\underline{\lambda}_{opt}$.

Lemma 2.3.7.

In a pure saddle with 2-strategies, at least one of $\pi_1^1(t)$ and $\pi_1^2(t)$ are semi-martingales, throughout their domain.

Proof.

We have $\underline{\lambda} \notin I^2$ so at least one of $(y-\lambda_y)$ or $(x-\lambda_x)$ has constant sign. //

Corollory 2.3.8.

Under R_t with saddle (i,j) s.t. $\lambda_j^2 \notin [0,1]$.

then $\lim_{\theta \downarrow 0} \lim_{t \to \infty} \pi_i^1(t) = 1$.

Proof.

This result is due to the domination of strategy i.

$\lambda_j^2 \notin [0,1] \Rightarrow \Delta\pi_i^1(t) > 0$ for all $\underset{\sim}{\pi}^k$ away from the boundary

and now apply 1.4.5. for the result.

//

This corollory indicates the type of difficulty we encounter for n >2 actions, using R_t when there are usually no dominating strategies. Boundary learning is then essential to achieve optimality.

Lemma 2.3.9.

Under R_t $\exists$ (i,j)s.t. $\lim_{t \to \infty} (\pi_i^1, \pi_j^2) = (1,1)$

Proof.

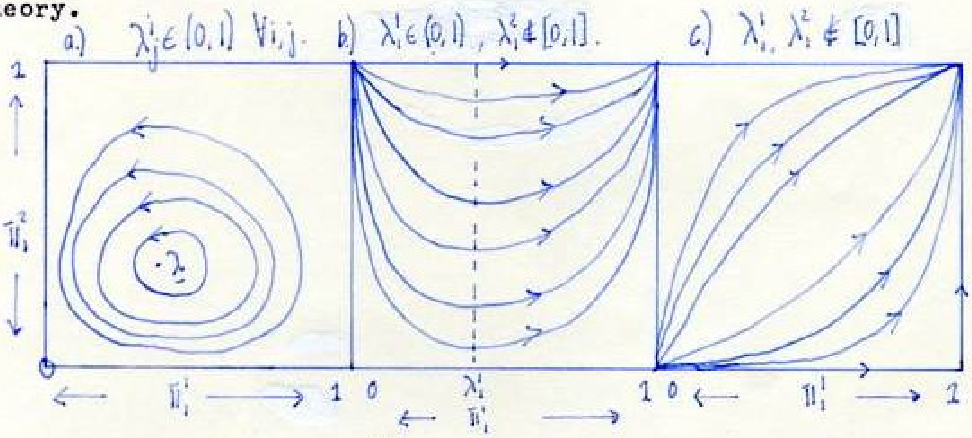
This is an application of 1.5.7. since $i(\pi_i^k) > 0$ so eventually each $\mathcal{A}_k$ will take some strategy infinitely often.

//

We know that 2.3.9. is false for optimal rules by 2.2.1. , but this does not prevent $\lim_{t\to\infty} \Phi^* \xrightarrow{a.s.} 0$, so that we cycle arbitrarily close to the boundary. We could prove this by showing that Φ^* is a super-martingale, which certainly seems to be true for convex Φ^* However, a rigorous treatment would involve much manipulation, so we shall leave it as a conjecture.

Whilst the mixed strategy behaviour gives insight into the operation of U.L. rules, the π -cells are more suitable when optimal strategies are pure. To preserve cyclic trajectories as a deterministic structure we could use networks of π-cells, which are considered in the next chapter, or the structured automata of Krinskii (1963, 1966).

We can actually avoid dealing with superfluous strategies by choosing a class of games which admits a unique completely mixed saddle, such as the Minkowski-Leontief form of economic theory.



The basic forms of Φ^* =constant for zero-sum games are sketched above. It is to be emphasised that the π-cell achieves optimal deterministic time-averaged behaviour, without any knowledge of the opponents strategy; only if its own strategy is successful. For convergence to λ Brown (see Karlin (1959)) required both players to possess complete knowledge of the past, and even then convergence is by no means easy to prove. Note that the theory in this section is easily extended to cover the cases with $\lambda^i_j \in \{0,1\}$ for some i,j, which occurs when $g_{rt}=g_{mn}$ for some $(r,t) \neq (m,n)$.

2.4. General Sum π-cell Games.

Definition 2.4.1.

i) Let g^k_{ij} = expected reward for k^{th} π-cell if $\otimes_1$ uses u_i
$\otimes_2$ uses u_j

as before $g^k_{ij} = q^k_{ij} - p^k_{ij}$ and $q^1_i = \frac{1}{2}(1 + \sum_j g^1_{ij}\pi^2_j)$.

ii) A Nash Point is defined as (i,j) s.t. $g^1_{ij} \geq g^1_{mj}$ $m \neq i$
$g^2_{ij} \geq g^2_{in}$ $n \neq j$

We proceed as before, omitting the convergence proof until we treat n-automata games, with the most general Nash convergence theorem.

Lemma 2.4.2.

For 2x2 matrix games the deterministic, continuous time trajectories are given by $\Phi^* = \prod_{i,j}(\pi^i_j)^{\mu^i_j}$ =constant, for linear U.L. rules.

where $\mu^i_j = (-1)^i \lambda^i_j / \gamma^i$ $\qquad \gamma^i = \theta_i\left(\sum_j g^i_{jj} - \sum_{j \neq k} g^i_{jk}\right)$.

and $\sum_j \lambda^1_j g^2_{jj'} = V^2 \qquad \sum g^1_{ij}\lambda^2_j = V^1 \quad \forall i$

Proof.

Put $\Delta x = \Delta \pi^1_1 = \gamma^x x(1-x)(y-\lambda_y)$

$\Delta y = \Delta \pi^2_1 = \gamma^y y(1-y)(x-\lambda_x)$

Thus find Φ s.t. $\partial\Phi$ on trajectories.

$\Phi_x \dot{x} + \Phi_y \dot{y} = 0$ if $\Phi_x = (x-\lambda_x)/\gamma^x x(1-x)$

$\Phi_y = -(y-\lambda_y)/\gamma^y y(1-y)$

or $\Phi = -1/\gamma^x\left(\lambda_x \log x + (1-\lambda_x)\log(1-x)\right) + 1/\gamma^y\left(\lambda_y \log y + (1-\lambda_y)\log(1-y)\right)$

and hence result. //

Corollory 2.4.3.

Under non-linear U.L. rules $\theta(\pi) = \theta\pi^\alpha(1-\pi)^\alpha$

$\Phi^* = f^x(x)/f^y(y)$ = constant with $f^x(x) = \exp\left(\int (x-\lambda_x)\,dx / \gamma^x (x(1-x))^{k=1+\alpha}\right)$.

Proof.

We have $\Phi_x = (x-\lambda_x) / \gamma^x (x(1-x))^k$ and similarly for Φ_y

and now integrate.

For zero-sum games $\gamma^x + \gamma^y = 0$ so we obtain f(x) f(y) = const as in 2.3.3. a). //

Lemma 2.4.4.

For 2x2 games Φ^* is periodic iff:-

a) $\lambda^i_j \in (0,1)\ \forall i,j$ and $(\gamma^1 > 0$ iff $\gamma^2 < 0)$.

or b) there are no Nash points.

Proof.

We shall assume a linear U.L. rule so that we can see the form of Φ^* explicitly. However the result is true for all U.L. rules.

a) If $\lambda^i_j \in (0,1)$ and $(\gamma^1 > 0$ iff $\gamma^2 < 0)$ then

$$\Phi^* = \left(x^{\lambda_x}(1-x)^{1-\lambda_x}\right)^{1/\gamma_x} \left(y^{\lambda_y}(1-y)^{1-\lambda_y}\right)^{-1/\gamma_y} = \text{const.}$$

or g(x)h(y) = constant and is periodic since g(x), h(y) are bounded on [0,1], assuming $\gamma^x > 0$ w.l.o.g.

If $\gamma^x > 0$ and $\gamma^y > 0$ say, then either $g(x)$ or $h(y)$ will be unbounded, giving hyperbolic trajectories.

Similarly if $\lambda^i_j \notin (0,1)$ for some i,j.

Now using a) we prove b).

Suppose $\exists$ a Nash point at (1,1) $g^1_{21} < g^1_{11}$, $g^2_{11} > g^2_{12}$

Then $\gamma^1 > 0$, $\gamma^2 > 0$ if $\lambda^i_j \in (0,1)$ and we have a contradiction.

And if Φ^* is periodic we get $g^1_{11} < g^1_{21}$ as contradiction. //

Corollory 2.4.5.

i) If $\exists$ 2 Nash Points then $\Phi^* = c$ forms a saddle.

ii) If $\exists$ only one Nash Point then $\exists\ \lambda^i_j \notin (0,1)$.

iii) $\int_{\Phi^*=c} V^k(\underset{\sim}{\Pi}^r : \gamma r)\,dt \Big/ \int_{\Phi^*=c} dt = V^k$, the equilibrium winnings.

Proof.

i) We take (1,1) and (2,2) as Nash points w.l.o.g.

Then $g^1_{21} < g^1_{11}$, $g^2_{11} > g^2_{12}$, $g^1_{22} > g^1_{12}$, $g^2_{22} > g^2_{21}$.

and hence $\gamma^1 > 0$, $\gamma^2 > 0$ and $\lambda^i_j \in (0,1)$

We get the result for (1,2) and (2,1) by symmetry.

Now construct the trajectories around $\underset{\sim}{\lambda}$ as in the sketch below.

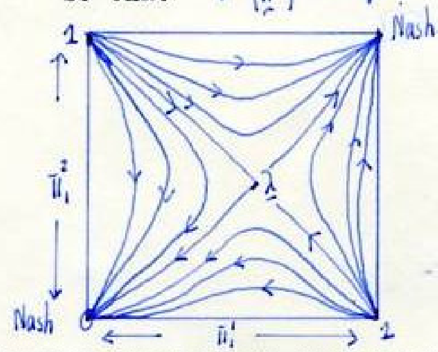
ii) This is immediate. Again, as in zero-sum games, we have

either a) $\lambda^i_j \notin (0,1)$ $\forall i, j$ with Π^1_1 , Π^2_2 s/mg, giving process at least ϵ-optimal.

or b) $\lambda^1_j \in (0,1)$, $\lambda^2_i \notin (0,1)$ say , and again a result analogous to 2.3.8 will hold.

iii) Just index V by superfix k in 2.3.3. c) for the result.

So that $\overline{V^k}(\underset{\sim}{\Pi}^r) = V^k$



All $\Phi : c$ end at $\Pi^k_i \in \{0,1\}$ and are essentially hyperbolic around the saddle.

This is the only other basic form we need to consider. Again the theory is easily extended to cover the special cases of $\lambda^i_j \in \{0,1\}$ for some i,j.

//

2.5. n-Automata Games.

Definition 2.5.1.

i) Let $g^k_{\underset{\sim}{\beta}}$ = expected reward to $\mathfrak{A}_k$ if $\mathfrak{A}_i$ takes β_i .

ii) Let $\Pi^k_{\underset{\sim}{\beta}} = \prod_{i \neq k} Pr(\mathfrak{A}_i \text{ takes } \beta_i)$.

iii) We put $g^k_{\underset{\sim}{\beta}} = q^k_{\underset{\sim}{\beta}} - p^k_{\underset{\sim}{\beta}}$ and to emphasise β_k we shall use ${}_i g^k_{\underset{\sim}{\beta}'}$ if $\beta_k = i$, similarly for ${}_r\Pi^k_{\underset{\sim}{\beta}'}$. (Note, β_k is put only as a left suffix in this notation with $\underset{\sim}{\beta}' = (\beta_1 \ldots \beta_{k-1}, \beta_{k+1} \ldots \beta_n)$.)

iv) $p_i^k = \frac{1}{2}\left(1 - \sum_{\beta} {}_i g^k_{\beta:i} \bar{\Pi}^k_{\beta}\right)$ as $\sum_{\beta} {}_r \bar{\Pi}^k_{\beta'} = 1 \;\; \forall r.$

Then $\Delta \bar{\Pi}_i^k = -\bar{\Pi}_i^k/2 \sum_{j,\beta'} \bar{\Pi}_j^k \, \theta_{ij}(\underline{\Pi}^k) \left({}_j g^k_{\beta'} - {}_i g^k_{\beta'}\right) \bar{\Pi}^k_{\beta}.$

with $\bar{\Pi}_i^k$ as before and $\Delta \bar{\Pi}_i^k(t) = \mathcal{E}\left(\bar{\Pi}_i^k(t+1) \mid \underline{\Pi}^r(t), \forall r\right) - \bar{\Pi}_i^k(t).$

<u>Theorem</u> 2.5.2.

If $\exists$ a Nash Point then we converge a.s. to some Nash β, under R_0 .

<u>Proof</u>.

Again using our boundary learning theory, the result is just a generalization of 2.2.1.

Assume w.l.o.g. that $\bar{\Pi}_1^k = (\underline{1})_k$ is Nash.

so $\quad {}_j g^k_{\beta:\underline{1}'} \leq {}_1 g^k_{\underline{1}'} \quad \forall j, k.$

a) We first prove convergence by examining the probabilistic boundary stability. In a small ϵ-neighbourhood of $\underline{1}$ we have:-

$$\Delta \bar{\Pi}_1^k = \tfrac{1}{2} \bar{\Pi}_1^k \sum_j \theta_{1j}(\underline{\Pi}^k) \bar{\Pi}_j^k \left({}_1 g^k_{\underline{1}'} - {}_j g^k_{\underline{1}'}\right) \bar{\Pi}^k_{\underline{1}} + O\left(\max_{\beta \neq \underline{1}} \sum_j \theta_{1j}(\underline{\Pi}^k) \bar{\Pi}_j^k \left({}_1 g^k_{\beta'} - {}_j g^k_{\beta'}\right) \Pi^k_{\beta}\right).$$

and we can take ϵ s.t. for $\bar{\Pi}_1^k > 1-\epsilon \;\; \forall k$, $\Delta \bar{\Pi}_1^k > 0$.

We effectively have a boundary learning process with time dependent g^k_{β} .

We "stop" the process if any $\bar{\Pi}_1^k$ leaves the ϵ-nbd $N_{\underline{1}}$. Then we obtain convergence by 1.7.5. as in 2.2.1.

Martingale theory is not essential here, for we can always use a staircase form of argument (1.11.3.) at boundaries to give Nash convergence, and also to give more intuitive insight.

b) Suppose $\bar{\Pi}_j^1 \uparrow 1, {}_j\bar{\Pi}^1_{\beta} \uparrow 1$ with β not Nash.

Then as in 2.2.1. $\exists\, n, k$ s.t. $\Delta \bar{\Pi}^k_{n \neq \beta_k} > 0$ in N_{β} ϵ-nbd and then under R_0 we have instability and the boundary learning theorem 1.7.5. gives $\bar{\Pi}^k_{\beta} \not\to 0$ so β is not an absobing boundary.

Hence, since boundaries communicate,

$\lim_{t \to \infty} \bar{\Pi}^k_i(t) = 1$ only if $\beta_k = i$ is component of Nash Point, β.

So $\bar{\Pi}^k_i \uparrow 1, {}_i\bar{\Pi}^k_{\beta} \uparrow 1$, only if β is Nash Point. //

In this most general case we have the difficulty of the non-uniqueness of Nash Point , but in many simple games we actually find the Nash Point is unique, as in 2.5.4.,2.5.5.below.

We apply 2.5.2 to games considered by Russian authors, initiated by Tsetlin (1963), in the context of structured automata. I believe that it is more natural to use unstructured automata, in particular, Π-cells, whenever we wish autonomous tracking to probabilistically stable boundaries. We only require structure to fix optimal non-boundary behaviour (as in cyclic $\Phi^* = c$) , in a new boundary formulation.

Tsetlin called the following homogenous games and his papers give computer simulations rather than convergence proofs.

Example 2.5.3. The Investment Game.

Let $1 \geq \alpha_1 \geq \dots \geq \alpha_\nu \geq 0$ and $\nu > n$ = no of players

and α_i/m_i =Pr(reward using i if m_i automata use it),

so $_i q^{k}_{\beta} = \alpha_i/m_i$ if $\beta_j = i$ occurs m_i times.

Then a Nash Point occurs if $\alpha_i/m_i \geq \left(\alpha_j/m_j+1\right) \quad \forall i,j.$

We can prove this exists by induction; if $\alpha_1/2 < \alpha_\nu$ then result, else put $\alpha_r < \alpha_1/2 < \alpha_{r-1}$ and continue re-ordering until stable. It is unclear whether the result is unique w.r.t. $\underset{\sim}{m} = (m_1, \dots m_\nu)$ at a Nash Point.

Now if we allow Π-cells to play this game under R_r , then 2.5.2. gives Nash convergence with $_i g^{k}_{\beta} = 2\alpha_i/m_i - 1$.

//

Example 2.5.4. Investment Game with Common Bank.

Now put $q^{k}_{\beta} = \frac{1}{n} \sum_j \left(\alpha_j/m_j\right)$ to give the investment game with common bank - we share out winnings.

Then $\underset{\sim}{\beta}^* = (\alpha_1, \alpha_2, \dots \dots \alpha_n)$ or a permutation gives a unique Nash

Point in $\underset{\sim}{m}=\underset{\sim}{1}$. Each $\bar{\mu}$ -cell asymtotically takes a different action from the set of the n best actions. If a $\mathfrak{G}_i$ fails then the remaining $\mathfrak{G}_{j\neq i}$ converge to Nash Point $(\alpha_1 \ldots \alpha_{n-1})$. Tsetlin calls this feature of behaviour "reliability".

So put $g^k_{\underset{\sim}{\beta}} = 2\left(\frac{1}{n}\sum_j \alpha_j/m_j\right)-1$ where m_j $\bar{\mu}$-cells take action j, as given by $\underset{\sim}{\beta}$.

Besides being a Nash Point, this solution $\underset{\sim}{\beta}^*$ is also a More Point in that we achieve maximum payoff.

//

Example 2.5.5. The Gur Game.

Let there be 2 strategies 0, and 1 say, and reward probability = $p\left(\frac{m}{n}\right)$ for all players.

m = number of $\bar{\mu}$-cells using 1.

n = " " " .

And let $p\left(\frac{m}{n}\right)$ have a unique maximum at ξ_0 say.

This is the Gur game described by Borovikov and Bryzgalov (1965).

Hence $\exists$ m^* s.t. $p\left(\frac{m^*-1}{n}\right) < p\left(\frac{m^*}{n}\right) > p\left(\frac{m^*+1}{n}\right)$ and $p\left(\frac{m^*}{n}\right) > p\left(\frac{i}{n}\right)$ $i \neq m^*$

and $g^k_{\underset{\sim}{\beta}} = 2\left(p\left(\frac{m}{n}\right)\right)-1$ $\forall k$ iff $\underset{\sim}{\beta}$ s.t. m $\bar{\mu}$-cells take 1.

We see that this is really a simple case of the previous example in that winnings are shared, with a choice of just 2 actions. However, it is a fundamental form and more recently Schmukler (1970) has considered it in more depth.

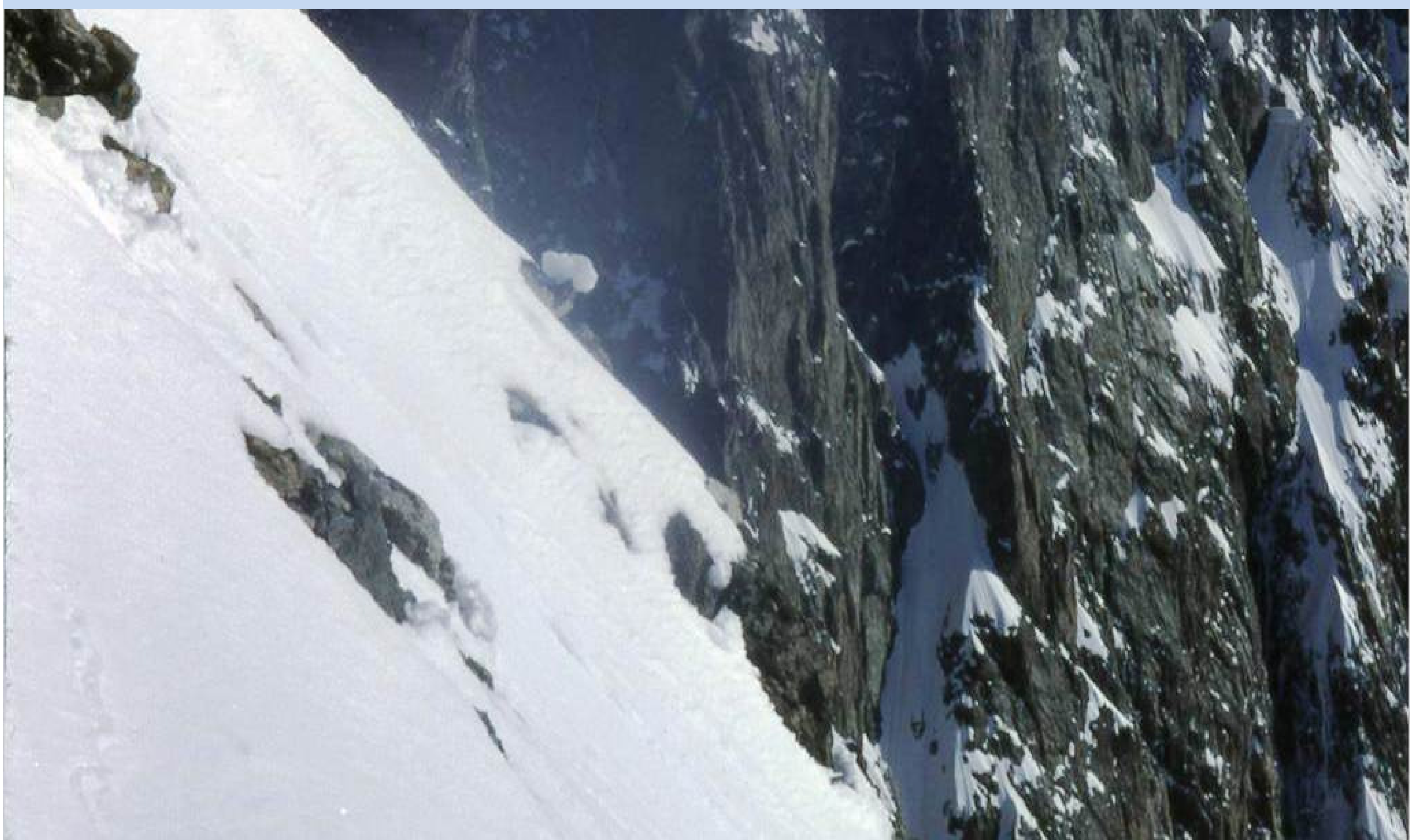
Using 2.5.2. we achieve a.s. convergence to the Nash Point.

//

$\bar{\mu}$ -cells can only achieve competitive solutions (Nash) since U.L. ensures maximum reward at the next trial.(asymtotically) So a $\bar{\mu}$-cell will always try to "double cross" its opponents unless we reformulate the game as in 2.5.5. , to ensure that this would feed back on to its payoff, converting perhaps a Pareto optimum to a Nash Point.

Chapter Three

Structured Learning Automata



Chapter 3.

" Organic life develops away from the concentric unicellular phase as the evolution of the species develops, and progresses along an axis, taking a direction and discovering aims."

Le Corbusier.

(La Ville Radieuse, 1935)

3. Structured Automata.

3.1. The Model for Π-cell Networks.

We have seen that the Π-cell responds asymtotically only to the environmental equilibrium probabilities. (1.12.9. and 2.3.3.) We shall now introduce a structure which responds to environmental fluctuations through utilizing optimal boundary learning (1.5.12.) to ensure, asymtotically a deterministic graph. Tsetlin (1961) introduced non-evolving structured automata in $\mathfrak{M}$ and Vorontsova and Varshavskii (1964) conducted computer simulations on structures evolving under the β-rule of Luce (1959). These simulations indicated that an initially random graph will converge to a quasi-linear graph resembling Tsetlin's automata, with near optimal payoff.

Since then, both Fu, in Mendel and Fu(1969), and Feichtinger (1970) have carried out an analysis similar to that of Vorontsova and Varshavskii, for static $\mathfrak{M}$, repeating the same errors when considering increments in transition probabilities. This is almost certainly due to their being unaware of the work of Norman (1966 - 1974). Yet even then there is a considerable amount of additional analysis required if we wish to apply uniform learning rules to Π-cell networks. The theory developed in chapter 1 now enables us to provide such an analytical basis for the evolution of stochastic automata with structure.

Definitions 3.1.1.

i) We define environment $\mathfrak{M}\left(\Delta_{\alpha\beta},\ q_{u_i}^{\alpha, s_a}\right)$ with

$\Delta_{\alpha\beta}$ = Pr($E_\alpha(t) \rightarrow E_\beta(t+1)$). where E_α is the environment state α .

$q_{u_i}^{\alpha, s_a}(t)$ = Pr(stimulus $s_a(t) = s \in \{0,1\} \mid u_i(t)$ and $E_\alpha(t)$).

and $\underset{\sim}{s}(t) = (s_1(t), s_2(t))$, each s_a independent with probability $q_i^{\alpha, s_a}(t)$, as above.

This vector stimulus $\underset{\sim}{s}(t)$ is used to enable the structure characterization theorems to be proved as a generalization of the method used in 1.12.9. We shall see that this definition prevents the process from "overlapping", just as for π-cell games we had a stochastic zero-sum, rather than deterministic zero-sum game.

ii) Let the automaton transition probabilities be given by:-

$$\sigma_{ij}^{s}(t) = Pr\left(x_i(t) \to x_j(t+1) \mid s_2(t) = s \in \{0,1\}\right).$$ where x_i denotes state i.

iii) Let $\Gamma_i = \{x_k : \pi\text{-cell } \otimes_i$ is always used$\}$

and $\pi_j^i(t) =$ Pr(for $\otimes_i$ we use $u_j(t)$ at time t)

iv) The reinforcement rules are uniformly learning for both σ_{ij}^{s} and π_j^i .

$$\sigma_{ij}^{s}(t+1) = \sigma_{ij}^{s}(t) + \theta_{jj}(\sigma_{ij}^{s}(t))\,(1-\sigma_{ij}^{s}(t))$$ if $s_2(t)=s$ in $x_i(t)$ and $x_i \to x_j$ with $s_1(t+1) = 1$.

$$\sigma_{ij}^{s}(t+1) = \sigma_{ij}^{s}(t)\,(1-\theta_{jr}(\sigma_{ij}^{s}(t))).$$ if $s_2(t)=s$ in $x_i(t)$ and $x_i \to x_{r\neq j}$ with $s_1(t+1) = 1$.

$$\sigma_{ij}^{s}(t+1) = \sigma_{ij}^{s}(t)$$ if $s_1(t+1) = 0$.

and with normalization $\sum_j \sigma_{ij}^{s}(t+1) = 1 \quad \forall i, s.$

v) For $\otimes_i$ we have $s_a(t) = s$ with probability $\sum_j \pi_j^i q_j^{\alpha, s_a}$, if E_α, so that each pair $s_1(t)$, $s_2(t)$ are independent once we are in a certain $x_j(t)$.

vi) If $i \in \Gamma_k$ in definition iv) above then we reinforce $\pi_j^k(t)$ as below.

$$\pi_j^k(t+1) = \pi_j^k(t) + \theta'_{jj}(\underset{\sim}{\pi}^k(t))\,(1-\pi_j^k(t)).$$ if $s_1(t)=1$, $u_j(t)$.

$$\pi_j^k(t+1) = \pi_j^k(t)\,(1-\theta'_{jm}(\underset{\sim}{\pi}^k(t)))$$ if $s_1(t)=1$, $u_{m\neq j}(t)$.

$$\pi_j^k(t+1) = \pi_j^k(t)$$ if $s_1(t) = 0$.

with normalization $\sum_j \pi_j^k(t+1) = 1 \quad \forall k$

The state transition is made independent of $s_1(t)$, even though this would give us more information on $\mathfrak{M}$. However, if we just used $s_1(t)=s_2(t)=s(t)$, then although it seems likely that we would obtain similar limiting structures, we could not split the process with respect to conditional expectation of increments in σ_{ij}^s . Thus the the vector stimulus $\underset{\sim}{s}(t)$ is more a technical device for proving theorems rather than an intrinsic feature of the learning process.

The same π-cell $\circledast_k$ can be reinforced from any state $i \in \Gamma_k$. Only through this formulation can we obtain evolving π-cell networks with "memory", and we shall discuss the partitioning of actions in section 3.7.

We denote this network of π-cells by $\mathcal{G}'(\circledast)$. We shall first consider the case in which we have $\pi_j^i \in \{0,1\} \ \forall i,j$ so that only σ_{ij}^s evolves.

$$\longrightarrow x_i(t)\ s_1(t) \mid s_2(t) \xrightarrow{\sigma_{ij}^{s_2}(t)} x_j(t+1)\ s_1(t+1) \mid s_2(t+1) \longrightarrow$$

where $s_a(t)$ are received with probability $\sum_j \pi_j^i(t)\, q_{ij}^{\alpha, s_a}$ if in E_α .

This does impose some initial structure, but we shall consider the complete case of simultaneous σ_{ij}^s and π_k^h reinforcement later. It will then be seen that this adds very little to the analysis of limiting structures, and yet we can then immediately generalise to hierarchies of π-cells, $\mathcal{G}^n(\circledast)$, in 3.9.

If $\pi_j^i \in \{0,1\}$ then we have a π_1-cell which is just a single action taken with probability 1. Rather than refer to this as a π_1-cell network we shall use the term <u>structured</u> <u>automaton</u>. In the next section we begin by showing the relationship between structures and π-cells. The structured automaton is denoted $\mathcal{G}'(\circledcirc)$

with θ as a Π_1-cell, or equivalently, an action.

3.2. Static Environments.

We shall prove that an evolving structured automaton asymtotically maximizes the payoff in a static environment $\mathfrak{M}$. Thus it performs the same function as a singleton Π-cell.

Definitions 3.2.1.

i) Let $\xi_i(t+1) = \sigma^s_{ji}(t)$ for $s_2(t)=s$ and $x_j(t)$.

Indeed, here we could put $s_1=s_2=s$, but we shall keep to s(t) in order to avoid confusion.

ii) Let $\Delta\sigma^s_{ij}(t) = \mathcal{E}\left(\sigma^s_{ij}(t+1) \mid \mathfrak{F}_t\right) - \sigma^s_{ij}(t)$.

where $\mathfrak{F}_t$ is the field of events for $0 \leq t' < t$ and also includes $\sigma^s_{ij}(t)$, $s_1(t)$.

We could define $s_1(t)$ and $s_2(t+\varepsilon)$ say, so that $\mathfrak{F}_t$ actually contains all events at times $t' \leq t$, but since we are more concerned with the concepts rather than complete rigour, we shall just split the process at $s_1(t)$, as in the diagram of 3.1.

For static environments:- $\Delta\sigma^s_{ij}(t) = \mathcal{E}\left(\sigma^s_{ij}(t+1) \mid \xi_i(t), \sigma^s_{ij}(t), s_1(t)\right) - \sigma^s_{ij}(t)$.

iii) A state i will be called +ve recurrent if

$$\underline{\text{not}} \lim_{t\to\infty} \xi_i(t) = 0.$$

Theorem 3.2.2.

In a static environment under $\mathfrak{R}_0$:-

a) $\lim_{t\to\infty} \sigma^s_{ij}(t) = v^s_{ij} \in [0,1] \quad \forall i,j.$

b) If i is +ve recurrent then $\lim_{t\to\infty} \sigma^s_{ij}(t) = 1$ for some j s.t. $q^1_j \geq q^1_k \quad k$.

Proof.

We have $\Delta\sigma^s_{ij}(t) = q^s_i\, \xi_i(t)\, \sigma^s_{ij}(t) \sum_k \sigma^s_{ik}(t)\, \theta_{jk}(\sigma^s_{ij})\,(q_j - q_k)$.

and for $q^1_j \geq q^1_k$ we find $\Delta\sigma^s_{ij}(t) \geq 0 \quad \forall t$

a) Now by semi-martingale convergence theorem, $\sigma^s_{ij} \xrightarrow{a.s.} v^s_{ij}$.

Now proceed by induction on q_i^1 in order of magnitude as in 1.3.1. to obtain the result for a).

b) If $\Delta\sigma_{ij}^s > 0$ except at $\sigma_{ij}^s \in \{0,1\}$ then $v_{ij}^s \in \{0,1\}$ and by n-action optimality theorem for R_0, 1.6.4. we obtain $r_{ij}^s = 1$ only if $q_j^1 \geq q_k^1$, $\forall$ k .

Similarly, if $\Delta r_{ij}^s \equiv 0$ with $q_j = q_k$ $\forall$ k, then we obtain the result since the conditional variance must vanish at v_{ij}^s, as for the singleton π-cell.

It is easier to put:-

$$\Delta^* \sigma_{ij}^s(t) = \mathcal{E}\left(\sigma_{ij}^s(t+1) \,|\, \sigma_{ij}^s(t) \text{ and } x_i(t)\right) - \sigma_{ij}^s(t).$$

for we now obtain a π-cell if i is +ve recurrent. But for dynamic environments we gain more insight by using the definition of 3.2.1. and so we have adopted it here also.

//

So the +ve recurrent states form a deterministic, connected structure with $q_i = q_j = \max\limits_k q_k$ $\forall$ i,j s.t. $\gamma_i > 0$ where

$$\gamma_i = \sum_{j,s} \gamma_j \, q_j^s \, v_{ji}^s = \lim_{t\to\infty} \overline{\rho_i(t)} = \text{limiting time-averaged } \rho_i(t).$$

Thus a structured automaton maximizes its payoff in a static and on considering q_i as a π_1-cell, 0, we have that the absorbing class is a subset of Γ_{opt}.

Also for each (i,s), σ_{ij}^s acts as a π-cell with links $i \to j$ as actions.

<u>Corollory</u> 3.2.3.

For $\rho'(\bullet)$ in M under R_0 :-

a) $\lim\limits_{t\to\infty} \pi_j^i = v_j^i \in [0,1]$ $\forall i,j$ and $\lim\limits_{t\to\infty} \sigma_{ij}^s = v_{ij}^s$ $\forall i,j,s.$

b) If i is +ve recurrent, $i \in \Gamma_k$ then

i) $\lim\limits_{t\to\infty} \pi_j^k = 1$ for some j s.t. $q_j \geq q_m$ $\forall m.$

ii) $\lim\limits_{t\to\infty} \sigma_{ij}^s = 1$ for some j s.t. $q_j \geq q_m$ $\forall m.$

Proof.

Now $\Delta\bar{\pi}^i_j \triangleq \mathcal{E}\left(\bar{\pi}^i_j(t+1) \mid \mathcal{F}_t\right) - \bar{\pi}^i_j(t)$

$$\Delta\bar{\pi}^k_j(t) = \sum_{i\in\Gamma_k} \varrho_i(t)\, \bar{\pi}^k_j(t) \sum_m \bar{\pi}^k_m (q_j - q_m)\, \theta_{jm}\left(\underset{\sim}{\bar{\pi}}^k(t)\right).$$

and as in 3.2.2. $\bar{\pi}^k_j \to v^k_j$ as $t\to\infty$ and then by induction.

And for $\Delta\sigma^s_{ij}(t)$ we must consider boundary behaviour, so let N^i_j be a neighbourhood of v^i_j s.t. $|\bar{\pi}^i_j - v^i_j| < \epsilon^i_j$ say.

$$\Delta\sigma^s_{ij}(t) = \varrho_i(t)\, \sigma^s_{ij}(t) \left(\sum_h \bar{\pi}^k_h q^s_h\right) \sum_k \sigma^s_{ik}\, \theta_{jk}\left(\sum_a v^m_a q^i_a - \sum_b v^n_b q^i_b\right) + O\left(F(\epsilon^i_j)\right).$$

with $j \in \Gamma_m$ and $k \in \Gamma_n$, and $i \in \Gamma_k$

and $F(\epsilon^i_j)$ is arbitrarily small, compared with the leading terms, as in 2.5.2.

Now order $\sum_a v^m_a q^i_a$ and take m* s.t. $\sum v^{m^*}_a q^i_a \geq \sum v^m_a q^i_a \quad \forall m.$

If $\sum_a v^m_a q^i_a = \sum_a v^n_a q^i_a \;\forall n$ then $\Delta\sigma^s_{ij} \to 0$ and we approach a martingale form, from which we can obtain convergence $\sigma^s_{ij} \to v^s_{ij}$.

If the limiting $\bar{\pi}$-cell payoffs are unequal, then we successively form semi-martingales for the result, as in 1.3.1.

If i is +ve recurrent, then the conditional variance > 0 gives $v^n_b, v^n_a \in \{0,1\}$ since <u>not</u> $\lim_{t\to\infty} \varrho_i(t) = 0$. (Note that this limit will only exist in the non +ve recurrent case).

In neighbourhood N^i_j of v^i_j we can apply optimal boundary learning theory for sem-martingale $\sigma^s_{ij}(t)$ as in 1.7.4. , since all boundaries communicate. (Note that the +ve recurrent set of states is non-empty for all finite automata).

Again, as in 2.5.2. , we can use a staircase construction instead on martingale theory to give us convergence only to stable boundaries. //

A $\bar{\pi}$-cell network thus has the same limiting payoff as a singleton $\bar{\pi}$-cell in static $\mathcal{M}$, so we certainly loose nothing (apart from simplicity) by this extension.

Our aim is to allow $\mathcal{G}^i(\theta)$ to adapt to a dynamic $\mathcal{M}$ through use of the equilibrium distribution for σ^s_{ij} . This adaptation will characterise a certain family of structures related to the work of Tsetlin (1961) and Stratonovich (1964) in that we discretize bayesian updating in the environmental likelihood simplex. First we strengthen our intuition by considering some simple properties of deterministic structures in dynamic $\mathfrak{M}$.

3.3. Evolution in a 2-Medium.

Definitions 3.3.1.

i) Let ρ^α_i = prob(E_α , in equilibrium | in state i).

so $\rho^\alpha_i = \gamma^\alpha_i / \gamma_i$ with γ^α_i=prob(E_α and x_i, in equilibrium).

ii) The symmetric 2-medium $\mathfrak{M}_2(\Delta_{\alpha\beta}, q^{\alpha,s}_{\mu_i})$ is defined as:-

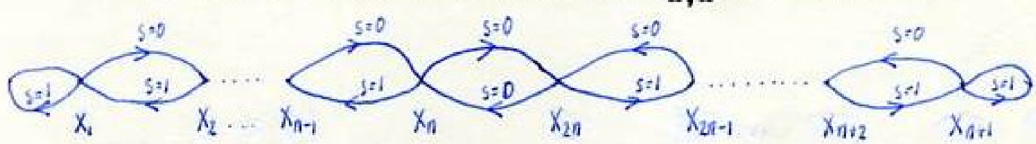
$$\Delta_{\alpha\beta} = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}$$

and $q_1^{11}=q_0^{01}=q_0^{10}=q_1^{00}=q$.

$q_1^{10}=q_0^{00}=q_0^{11}=q_1^{01}=p=1-q$.

The results from the analysis here have natural extensions to many other $\mathfrak{M}$ and in the following section 3.4. we consider the symmetric n-medium $\mathfrak{M}_n$.

iii) Let the structured automaton $L_{n,n}$ be defined as:-



and so that $L_{1,1}$ is just:-

and for $1 \leq i \leq n$ $i \in \Gamma_1$.

$n+1 \leq i \leq 2n$ $i \in \Gamma_0$.

This $L_{n,n}$ is the basic linear symmetric automaton.

Lemma 3.3.2.

a) $\rho^0_1 = q - \delta(q-p)$ for $L_{1,1}$.

b) $\rho^0_{r\geq n+1}, \rho^1_{r<n} \uparrow$ as $r \downarrow$ for $L_{n,n}$.

c) $\rho^0_{n-1} = q$ iff $f(ny)/f(y) = 1/\delta$ for $L_{n,n}$ where

$f(y) = \cosh y - 1$ and $\cosh y = \frac{1}{2pq}\left(\frac{1-\delta}{1-2\delta}\right) - 1.$

d) $\bar{R}(L_{2,2}) = \bar{R}(L_{1,1})$ iff $\rho^0_{n-1} = q$ with n=2, where

$\bar{R}$ = average payoff = $\sum_{\alpha,i} \pi^\alpha_i q^\alpha_i$.

Proof.

a) Solve $\pi^\beta_j = \sum_{\alpha,i,s} \pi^\alpha_i q^{\alpha s}_i \delta^s_{ij} \Delta_{\alpha\beta}$ 1)

b) Solve 1) for $L_{n,n}$ as in the paper of Tsetlin (1961) and then show that $\pi^0_r/\pi^1_r > \pi^0_{r+1}/\pi^1_{r+1}$ and hence that $\rho^0_{r<n} \uparrow$ as $r\downarrow$ and similarly for $\rho^1_{r\geq n+1} \uparrow$ as $r\downarrow$.

c) Again, as in b), the result arises from manipulation and its truth is closely linked with our next result d).

d) It can be shown that $\bar{R}(L_{n,n}) = \dfrac{\tanh(ny/2)}{\coth(y/2)+n\gamma\coth(ny)}$ 2)

with $\gamma = 2\delta/(1-2\delta)(q-p)^2$ and cosh(y) as before.

Using c) we obtain $\rho^0_2 = q$ if $(\cosh 2y - 1)\delta = \cosh y - 1$

or $\cosh(y) = \left(\frac{1}{2\delta} - 1\right)$

and substituting for cosh(y) gives $\delta(1-\delta)/(1-2\delta) = pq$.

But Tsetlin (1963) gives this as the condition for $\bar{R}(L_{2,2}) = \bar{R}(L_{1,1})$ and indeed it is not too difficult to verify this using 2).

//

The result d) is almost certainly false for the general case when $\bar{R}(L_{n,n}) = \bar{R}(L_{n-1,n-1})$, and this is related to the optimal automaton not possessing the SOSA property (self-one-step-ahead) except for $L_{1,1}$. This property is defined in 3.3.8. and considered in detail in subsequent sections.

The manipulation required to achieve the above results is most tedious and so apart from a), they have not been extended to $\mathcal{M}_n$ even though they almost certainly exist.

We shall now see the relevance of 3.3.2. in giving rise to the family of limiting structures denoted by $\mathcal{A}_0$. First we consider the fine structure of the process before taking the equilibrium values which $\mathcal{R}_0$ utilizes.

<u>Definitions</u> 3.3.3.

i) Let $\Theta_j^\beta(t)$ =Pr(in E_β | in state j at time t).

ii) Let $\omega_i^\alpha(t)$ =Pr(in E_α and state i at t).

iii) Let $\rho_i(t)$ =Pr(in state i at time t).

iv) We define $\Delta\sigma_{ij}^s(t) = \mathcal{E}\left(\sigma_{ij}^s(t+1) \mid s_1(t), \{\sigma_{ij}^s(t)\}, \mathcal{F}_t\right) - \sigma_{ij}^s(t)$.

Now by bayesian rules let $(\omega_i^\alpha)^*(t+1) = \sum_\beta \Theta_j^\beta(t)\, q_j^{\beta,s_2}\, \sigma_{ji}^{s_2}\, \Delta_{\beta\alpha} \Big/ \left(\sum_\beta \Theta_j^\beta q_j^{\beta,s_2}\right)$ 3)

and where $s_2(t)$ and $u_j(t)$ are used in $x_j(t)$.

Now we receive $s_1(t+1)$ and $\omega_i^\alpha(t+1) = q_i^{\alpha,s_1}(\omega_i^\alpha)^*(t+1) \Big/ \left(\sum_\beta (\omega_i^\beta)^*(t+1)\, q_i^{\beta,s_1}\right)$ 4)

So using 3) and 4) we have $\Theta_i^\alpha \xrightarrow{s_2(t),\ s_1(t+1)} \omega_i^\alpha(t+1)$.

Now $\rho_i(t+1) = \sum_\alpha \Theta_j^\alpha(t)\, q_j^{\alpha,s_2}\, \sigma_{ji}^{s_2} \Big/ \left(\sum_\alpha q_j^{\alpha,s_2}\Theta_j^\alpha\right)$. 5)

so $\Theta_i^\alpha(t+1) = \omega_i^\alpha(t+1) / \rho_i(t+1)$

Hence we generate our conditional probabilities by inductively using 3) $\rightarrow$ 5), and $\Delta\sigma_{ij}^s(t) = \mathcal{E}\left(\sigma_{ij}^s(t+1) \mid \Theta_i^\alpha(t), \{\sigma_{ij}^s(t)\}, \rho_i(t)\right) - \sigma_{ij}^s(t)$.

Note that we can take $(\Theta_i^\alpha)^* = \Theta_i^\alpha q_i^{\alpha,s_1} q_i^{\alpha,s_2} \Big/ \sum_\beta \Theta_i^\beta q_i^{\beta,s_2} q_i^{\beta,s_1}$ 6)

for the $\underset{\sim}{\rho}$ updating and the result is the same as if we had two successive updatings, for the $s_2(t)$ and $s_1(t+1)$ taken individually.

Lemma 3.3.4.

i) If $\{\sigma^s_{ij}\}$ is constant then $\bar{\omega}^\alpha_i = \Upsilon^\alpha_i$, $\bar{\Theta}^\alpha_i = \varrho^\alpha_i$, $\bar{\varrho}_i = \Upsilon_i$

Where $\Upsilon^\alpha_i = \sum_{\beta,j,s} \Upsilon^\beta_j \Delta_{\beta\alpha} q^{\beta,s}_j . \sigma^s_{ji}$, $\Upsilon_i = \sum_\alpha \Upsilon^\alpha_i$ and $\varrho^\alpha_i = \Upsilon^\alpha_i / \Upsilon_i$

ii) $\Delta\sigma^s_{ij}(t) = C \sum_{\alpha,k} (\omega^\alpha_k)^*(t+1) R^\alpha_{jk}$ where $C = \sum_\beta q^{\beta,s}_i \omega^\beta_i(t)$

and $R^\alpha_{jk} = \Theta_{jk}(\{\sigma^s_{ij}\})(q^{\alpha,1}_j - q^{\alpha,1}_k)$ = reward comparison function.

Proof.

i) This is a consequence of bayes' rule and we indeed also have $\bar{\omega}^\alpha_i = (\bar{\omega}^\alpha_i)^* = \Upsilon^\alpha_i$ and this is proved by writing out all possibilities.

ii) $\Delta\sigma^s_{ij}(t) = \sum_\alpha \omega^\alpha_i(t) q^{\alpha,s}_i \sum_{k,\beta} \sigma^s_{ik} \Delta_{\alpha\beta} \Theta_{jk}[\{\sigma^s_{ij}(t)\}] (q^{\beta,1}_j - q^{\beta,1}_k)$.

but by i) we have the result immediately since we just update w.r.t. the stimulus $s_2(t)$ to determine our expected payoffs on taking an action at the next trial.

//

Lemma 3.3.5.

The maximum average payoff is achieved by the bayesian rule:-

Let a) $\omega_\alpha(t+1) = \sum_\beta \omega_\beta(t) q^{\beta,s}_i \Delta_{\beta\alpha} / \sum_\beta \omega_\beta q^{\beta,s}_i$ if $s(t)=s$ and use $u_i(t)$.

where $\omega_\alpha(t)$=Pr(in E_α at time t).

Then b) At time t+1 take u_j only if $\sum_\alpha \omega_\alpha (q^{\alpha,1}_j - q^{\alpha,1}_k) \geq 0 \quad \forall k$.

(Here we have the simplified model with single scalar stimulus s, and we assume that the environmental parameters $\{\Delta_{\alpha\beta}, q^{\alpha,s}_{u_i}\}$ are known, so that only E is unknown.) For vector stimulus we modify a) in a similar way to eqn 6).

Proof.

Stratonovich (1964) proves this.

//

Thus the O.S.A. policy is optimal here since environmental information is "gained equally from all states". It is for this reason that our discrete automaton will achieve environmental adaptation, using reinforcement only over successive trials.

We now consider evolving automata in $\mathfrak{M}_2$ and our next theorem gives both existence and characterization of the limiting family of structures.

Theorem 3.3.6.

In $\mathfrak{M}_2$ under $\mathcal{R}_0$ with q, $1-\delta > \frac{1}{2}$.

a) $\sigma_{ij}^{s=0} \to 1$ only if $\rho_i^{\alpha}(\nu_{ij}^{s}) > q$ with $i \in \Gamma_\alpha$, $j \in \Gamma_i$.

b) $\sigma_{ij}^{s=1} \to 1$ only if $\rho_i^{\alpha}(\nu_{ij}^{s}) > p$ with $j \in \Gamma_i$.

where $q_i^{11} = q$ for $i \in \Gamma_1$ say $1 \leq i \leq n$

and $q_i^{01} = q$ for $i \in \Gamma_0$ say $n+1 \leq i \leq 2n$

Thus we have $q_i^{\alpha 1} = q$, and $q_i^{\alpha 0} = p$ for $i \in \Gamma_\alpha$.

c) $\sigma_{ij}^{s} \to \nu_{ij}^{s} \in \{0, 1\}$. for +ve recurrent i.

Proof.

We write out the expected increments:-

$$\Delta\sigma_{ij}^{0}(t) = s_{\nu\tau} K \rho_i(t)\, \sigma_{ij}^{0}(t) \,(\Theta_i^{\nu} - q) \sum_{k \in \Gamma_{1-\tau}} \sigma_{ik}^{0}\, \theta_{jk}(\{\sigma_{ij}^{0}\}). \qquad 7)$$

$$\Delta\sigma_{ij}^{1}(t) = s_{\nu\tau} K \rho_i(t)\, \sigma_{ij}^{1}(t) \,(\Theta_i^{\nu} - p) \sum_{k \in \Gamma_{1-\tau}} \sigma_{ik}^{1}\, \theta_{jk}(\{\sigma_{ij}^{1}\}).$$

where $s_{\nu\tau} = +1$ if $\nu = \tau$ and $i \in \Gamma_\nu$, $j \in \Gamma_\tau$.

$s_{\nu\tau} = -1$ if $\nu \neq \tau$ $K = (1-2p)(1-2\delta)$.

Consider now $\{\sigma_{ij}^{s}\}$ fixed so that we can consider the process for each state as a random walk defined on the environment markov chain. Thus near the boundary absorbing set we can apply the techniques of section 1.12. which were used to prove 1.12.9. Here for each state i, we have an underlying markov process with equilibrium distribution $\Upsilon_i^{\alpha}(\sigma_{ij}^{s})$, which is an example of the type of process considered by Miller (1962).

Then $$\overline{\Delta\sigma_{ij}^{0}} = s_{\nu\tau} K \Upsilon_i\, \sigma_{ij}^{0} \,(\rho_i^{\nu} - q) \sum_{h \in \Gamma_{1-\tau}} \sigma_{ih}^{0}\, \theta_{jh}(\{\sigma_{ij}^{s}\}). \qquad 8)$$

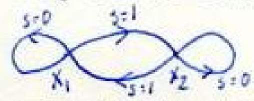
$$\overline{\Delta\sigma_{ij}^{s}} = s_{vz} K \gamma_i \sigma_{ij}^{s} \left(\rho_i^{v} - p\right) \sum_{k \in I - \tau} \sigma_{ik}^{1}\left(\theta_{jk}\left(\{\sigma_{ij}^{s}\}\right)\right). \qquad 9)$$

where $\overline{\Delta\sigma_{ij}^{s}}$ $\triangleq$ mean increment in σ_{ij}^{s} w.r.t. the equilibrium process. We can do this since the process is split up by the vector stimulus $\underset{\sim}{s}(t)$ so that successive increments do not "overlap", as shown in 3.1.

We now apply optimal boundary learning in a dynamic medium for the result (1.7.5. 1.12.9.) which is now immediate from 7) and 8), noting that:-

i) A limiting structure always exists since ${}_2\underset{\sim}{L}_1$ is stable with $\rho_1^{v} = q - \delta(q-p) > \frac{1}{2}$. Convergence to stable boundaries follows by a β-staircase construction as in 1.12.9. for each +ve recurrent state. This gives us c) and a non-empty limiting set of structures.

ii) If $\delta > \frac{1}{2}$ then the only stable limit is ${}_2\underset{\sim}{L}_1^{c}$ as below



iii) If $q < \frac{1}{2}$ then the only stable structure is ${}_2\underset{\sim}{L}_1$, assuming also $\delta < \frac{1}{2}$.

iv) Although $\overline{\Delta\sigma_{ij}^{s}} > 0$ say for all $j \in P_i$, $i \in P_\alpha$ if $\rho_i^{\alpha} > q$. we get convergence to just one $j_0 \in P_i$, as in the n-action theorem with multiple optima 1.6.4. The dynamic medium adds nothing for if we have 2-actions with $q_1^{\alpha,s} = q_2^{\alpha,s}\ \forall \alpha$ then $\gamma(\Pi) = \Pi$ for all $\mathcal{M}$, as we have a martingale.

Thus in contrast with bayesian rules which oscillate if we try to decide between 2 identical hypotheses, our U.L. rules still give boundary convergence. This property is important in their application to Π-cell networks.

v) The +ve recurrent states i are those with $\gamma_i(v_{ij}^{s}) > 0$.

//

<u>Corollory</u> 3.3.7.

In $\mathcal{M}_2$ a) ${}_2\underset{\sim}{L}_r$ are stable if $\rho_{r-1}^{1} > q$.

b) If $\delta(1-\delta)/q(1-q) \geq (1-2\delta)$

then ${}_2L_1$ is the unique and optimal limit, when $\delta < \frac{1}{2}$.

Proof.

a) We have $\rho_i^1 > q$ gives $i \to j \in \Gamma_i$ and by 3.3.2. $\rho_r^1 \uparrow$ as $r \downarrow$ with $\rho_{r-1}^1 > q$ iff $f(ny)/f(y) < 1/y$.

b) This is just case d) of 3.3.2. where ${}_2L_1$ is stable iff it is optimal, and uniqueness follows easily.

//

Theorem 3.3.6. actually gives $i \xrightarrow{s} j$ if and only if we maximize the expected reward at the next trial w.r.t. the equilibrium distribution γ_i^α in our present state. So we essentially have each σ_{ij}^s as a Π-cell $\otimes_{(i,s)}$ in medium with equilibrium γ_i^α and transitions $(\alpha, i) \to (\beta, j)$ with probability $\Delta_{\alpha\beta}\, q_i^{\alpha,s}\, \sigma_{ij}^s$, and with actions u_j. Thus we have σ_{ij}^s as a process defined on a markov chain in the manner of Miller (1962) and Keilson and Wishart (1965), in the slow learning limit $\theta \downarrow 0$. The σ_{ij}^s processes then interact through $\rho_i^\alpha(\sigma_{ij}^s)$ to give a certain family of stable solutions which all have the SOSA property. Thus we just need apply 1.12.9. to each $\otimes_{(i,s)}$ for 3.3.6.

Definition 3.3.8.

i) A deterministic automaton has the SOSA property if for every state i, we have $\nu_{ij}^s = 1$ only if u_j maximizes $\sum_{\beta,\alpha} \rho_i^\alpha \Delta_{\alpha\beta}\, q_i^{\alpha,s}\, q_{u_j}^{\beta,1}$ where $\rho_i^\alpha = \gamma_i^\alpha/\gamma_i$ =Pr(in E_α | in x_i). (SOSA ~ self-one-step-ahead).

ii) We denote the family of automata with the SOSA property by $\mathcal{A}_0$.

Thus the automaton maximizes the payoff at the next trial w.r.t. its own equilibrium distribution over states and environments.

3.4. Evolution in an n-Medium.

Definition 3.4.1.

i) We define the symmetric n-medium $\mathcal{M}_n(\Delta_{\alpha\beta}, q_{\forall i}^{\alpha,s})$ as:-

a) $\Delta_{\alpha\beta} = \begin{pmatrix} 1-(n-1)\delta, & \delta, & \cdots & \delta \\ \delta & 1-(n-1)\delta & & \\ \vdots & & \ddots & \delta \\ \delta & & \delta & (1-(n-1)\delta) \end{pmatrix}$ b) $q_i^{i1} = q = q_{j\neq i}^{i0}$, else $1-q=p$,

and $i \in \Gamma_\alpha$ if $q_i^{\alpha,1} = q$.

So $\Delta_{\alpha\beta} = \delta \quad \alpha \neq \beta$ with $\sum_\beta \Delta_{\alpha\beta} = 1$

$\Delta_{\alpha\beta} = (1-(n-1)\delta) \quad \alpha = \beta$

ii) The automaton ${}_nL_1$ has transitions $V_{ii}^{s=1} = 1 \ \forall i, \ V_{i,i+1}^{s=0} = 1$ with

$V_{n1}^{0} = 1$ so that we have 

$\sim {}_nL_1$. with

n indicating the number of states.

We now prove a lemma similar in nature to 3.3.2.

Lemma 3.4.2.

If $\delta < 1/n$ and $q > \frac{1}{2}$ then:-

a) $\psi_i^r \downarrow$ in r for ${}_nL_1$.

b) $\bar{R}({}_nL_1) > p+((q-p)/n)$, which is the n-cell limit in $\mathcal{M}_n$ under R_0 .

c) ${}_nL_1$ has the SOSA property whilst ${}_{m<n}L_1$ is not SOSA in $\mathcal{M}_n$.

d) $\bar{R}({}_nL_1) > \bar{R}(\lim_{\theta \downarrow 0} L_{R-P})$ in $\mathcal{M}_n$.

e) $\bar{R}(L_{R-P}) = \bar{R}({}_nL_1)$, in any static environment.

Proof.

a) Solve $\psi_j^\beta = \sum_{\alpha,i,s} \psi_i^\alpha q_i^{\alpha,s} \sigma_{ij}^s \Delta_{\alpha\beta}$ to get

$$\psi_1^{r>1} = 1/n^2 \left(1 - (1/p - 1/q)(1/n\delta - 1)^{-1}\left(\gamma^n\left(\frac{1+nq\,\delta/p}{1+np\,\delta/q}\right) - 1\right)^{-1}\gamma^{r-2}\right) < 1/n^2.$$

with $\gamma = \left(\dfrac{q+np\delta}{q(1-n\delta)}\right) > 1.$

and $\Upsilon_1^1 = 1/n^2(p+nq\delta)\left(\left(q+np\delta\right) - \left(q/p - 1\right)n\delta\left(\delta^n\theta^{-1} - 1\right)^{-1}\right) > 1/n^2$

with $\theta = (1+np\delta/q)/(1+nq\delta/p)$ and $\sum_1^n \Upsilon_1^n = 1/n.$

and for $\delta = 0,\ \Upsilon_1^{r\neq 1} = p/q\,\Upsilon_1^1\ ,\ \delta = 1/n,\ \Upsilon_1^r = \Upsilon_1^s\ \forall r,s.$

Clearly then $\Upsilon_1^r \downarrow r$ by combining the results above.

b) $\bar{R}({}_nL_1) = p + \Upsilon_1^1(q-p)n.$ with $\Upsilon_1^1 > 1/n^2$ and hence result.

c) We need to check $\omega_1^{s=1}/\omega_{r>1}^{s=1} > 1$ and $\omega_2^{s=0}/\omega_1^{s=0} > 1.$

where ω_r^s = Pr(in E_r | use u_1 and receive stimulus s, $\underset{\sim}{\omega}$).

and ω_r = Pr(in E_r when in state x_1). $= \xi_1^r.$

$$\omega_1^1/\omega_r^1 = \left(\omega_1 q(1-(n-1)\delta) + (1-\omega_1)p\delta\right)/\left(\omega_1 q\delta + \omega_r(1-(n-1)\delta)p + (1-(\omega_1+\omega_r))\delta p\right) = \omega_1 q/\omega_r p$$

and $\omega_1^1/\omega_r^1 > 1$ iff $\omega_1/\omega_r > p/q.$

Similarly $\omega_2^0/\omega_1^0 > 1$ iff $\omega_2/\omega_1 > p/q.$

and as $\omega_2/\omega_r > 1$ we need only consider ω_2.

But as $\delta \downarrow 0\quad \Upsilon_1^{r\neq 1}/\Upsilon_1^1 \rightarrow p/q\qquad \delta \uparrow 1/n\quad \Upsilon_1^{r\neq 1}/\Upsilon_1^1 \rightarrow 1$

and by monotonicity we have $p/q < \Upsilon_1^{r\neq 1}/\Upsilon_1^1 < 1.$

Hence ${}_nL_1$ is SOSA.

To show that ${}_{m<n}L_1$ is <u>not</u> SOSA let us take actions $u_1 \ldots u_m$ with states $x_1 \ldots x_m$ and actions $u_{r>m}$ with no states.

Then $\xi_r^{n>1, s>m} = 1/n$ by symmetry for all states $r \leq m$.

yet $\xi_r^{t\neq r}{}_{t\leq m} < 1/n$. So $\xi_r^{s>m}/\xi_r^{t\neq r}{}_{t\leq m} > 1$

So if we wish to maximize payoff at the next trial given a penalty we should take an action $u_{r>m}$ which the structure will not allow.

Hence $_{m<n}L_1$ is not SOSA in $\mathcal{M}_n$.

d) Here, L_{R-P} is the only known reinforcement rule for unstructured automata "tracking" in a dynamic environment.

$$\left.\begin{aligned} \pi_i(t+1) &= \pi_i(t) + \theta(1-\pi_i(t)) \\ \pi_j(t+1) &= \pi_j(t)(1-\theta) \end{aligned}\right\} \quad u_i(t) \text{ and } s(t) = 1. \quad j \neq i.$$

and

$$\left.\begin{aligned} \pi_i(t+1) &= \pi_i(t)(1-\theta) \\ \pi_j(t+1) &= \pi_j(t) + \theta\pi_i(t)/(n-1) \end{aligned}\right\} \quad u_i(t) \text{ and } s(t) = 0. \quad j \neq i.$$

Norman (1972) obtains a limiting normal distribution for sufficiently small θ about the point $\mathcal{E}(\pi_i(\infty)) = 1/\rho_i \big/ \sum_j 1/\rho_j$, with variance ~ $O(\theta^2)$.

Now from b) we have $\bar{R}(_nL_1) > p+((q-p)/n)$, but with L_{R-P} we have $\bar{R}(L_{R-P}) > p+((q-p)/(n+(q-p)/p))$, which is the case when the environment has just switched. But in the limit $\theta \to 0$ of slow learning we shall asymtotically have $\bar{R}(L_{R-P}) = p+((q-p)/n)$, by letting $\theta \ll \delta$.

normal limits. Thus we have the result d).

e) For static environments we have:-

$\bar{R}(_nL_1) = \sum_i \gamma_i q_i$. where $\gamma_i \rho_i = \gamma_{i-1}\rho_{i-1}$.

Thus $\gamma_i = 1/\rho_i \big/ \sum_j 1/\rho_j$ So $\bar{R}(L_{R-P}) = \sum_i \mathcal{E}(\pi_i(\infty))q_i = \sum_i \gamma_i q_i = \bar{R}(_nL_1)$.

Now by d) as $\theta \downarrow 0$ $\bar{R}(_nL_1) > \bar{R}(L_{R-P})$

also as $\delta \downarrow 0$ $\bar{R}(L_{R-P}) \to \bar{R}(_nL_1)$

yet $\bar{R}(_nL_r) > \bar{R}(_nL_1) \to \bar{R}(L_{R-P})$ which gives justification for a structural strategy rather than unstructured tracking as used by Chandrasekaren (1967). In addition, an evolving $\mathcal{G}$ in $\mathcal{M}$ (static) achieves optimal payoff.

//

The result e) has the interpretation that the L_{R-P} rule has the same average payoff as the strategy ; reward - stay put, penalty- move on. This strategy gives precisely the probability matching of Norman and Yellott (1966) and in the later work of Norman (1972).

<u>Remarks</u> 3.4.3.

i) It is possible that for some θ and δ, $\bar{R}({}_nL_1) < \bar{R}(L_{R-P})$ yet in many cases we could use an alternative structure ${}_nL_r$ with larger memory. The L_{R-P} rule in dynamic $\mathfrak{M}$ has not yet been analysed in detail.

ii) Just as the ${}_2L_1$ was the kernal structure for $\mathfrak{M}_2$ we shall now show that ${}_nL_1$ is a stable limit in $\mathfrak{M}_n$ by considering $\overline{\Delta\sigma}^s_{ij}$.

iii) We shall assume that $\rho^\alpha_i\uparrow$ on "arms" $i \in \Gamma_\alpha$ as in 3.3.2. b)

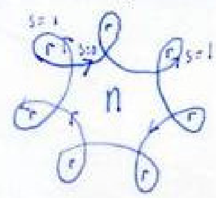
Certainly as $\delta\downarrow 0$ $\rho^\alpha_r/\rho^\alpha_{r+1} \to q/p$ with $r\Downarrow$ as we proceed down the "arm".

e.g. $1 \leq i \leq 6$ gives $i \in \Gamma_\alpha$,

and $\rho^\alpha_i\uparrow$ as $i\downarrow$.

We then obtain ${}_nL_r$ as SOSA , and for given δ we shall obtain some r_{max} with r indicating the memory depth. (Or "arm" length.)

${}_nL_r \sim$



iv) Exact calculations for ${}_nL_r$ are prohibitively complicated, when we consider how difficult it is even to obtain the results of 3.3.2. for ${}_2L_r$.

<u>Theorem</u> 3.4.4.

A structured automaton evolving under R_0 in medium $\mathfrak{M}_n$ has ${}_nL_1$ as a stable limiting structure, if $n\delta < 1$, $2p < 1$, and ${}_{m<n}L_1$ is unstable.

<u>Proof.</u>

We write out:-

$$\overline{\Delta\sigma}^{s=0}_{ij} = K\sigma^s_{ij}\gamma_i\Big(\sum_{k\notin\Gamma_i}\sigma^s_{ik}\theta_{jk}(\{\sigma^{s=0}_{ij}\})(\rho^j_i p - \rho^k_i q)\Big) \quad \text{for } j \in \Gamma_i \qquad 10.)$$

$$\overline{\Delta\sigma}^{s=0}_{ij} = K\sigma^s_{ij}\gamma_i\Big(\sum_{k\in\Gamma_i}\sigma^s_{ik}\theta_{jk}(\rho^j_i q - \rho^k_i p) + \sum_{k\notin\Gamma_i}\sigma^s_{ik}\theta_{jk}\, q\,(\rho^j_i - \rho^k_i)\Big). \qquad 11.)$$

for $j \notin \Gamma_i$.

and $K = (1 - n\delta)(q-p)$ and $\rho_i^j = \rho_i^j(\{\sigma_{ij}^s\})$

and similarly for s=1, interchanging p and q; and for M_2 we just obtain our equations in 3.3.6.

The derivations of 10) and 11) are not of great interest, but we shall see in a later theorem why $\overline{\Delta\sigma}_{ij}^s$ must have the above form.

The $\overline{\Delta\sigma}_{ij}^s$ are taken with $\{\sigma_{ij}^s\}$ held fixed, as in 3.3.6. , then the optimality property of $\theta_{jk}(\{\sigma_{ij}^s\})$ will give convergence to that action which maximizes expected payoff w.r.t. the equilibrium distribution. Let x_i be a +ve recurrent state.

a) Now for s=0, $j \in \Gamma_i$, $\overline{\Delta\sigma}_{ij}^{s=0} > 0$ if $\rho_i^j/\rho_i^k > q/p \quad \forall k \notin \Gamma_i$

and so by boundary learning $\sigma_{ij}^s \to 1$ for some $j \in \Gamma_i$

iff $\rho_i^j/\rho_i^k > q/p \quad \forall k \notin \Gamma_i$

For if $\rho_i^j/\rho_i^k < q/p$ for some k then $\overline{\Delta\sigma}_{ik}^s > 0$ with

$\sigma_{ij}^s = 1 - \epsilon_{ij}^s$ say in a sufficiently small nbd N_{ij}^s of the boundary V_{ij}^s

b) For s=0, $j \notin \Gamma_i$

$\overline{\Delta\sigma}_{ij}^{s=0} > 0$ if $\rho_i^j/\rho_i^k > p/q$, $k \in \Gamma_i$ and $\rho_i^j/\rho_i^m > 1 \quad \forall m \notin \Gamma_i$.

And by boundary learning $\sigma_{ij}^s \to 1$ for some $j \notin \Gamma_i$

iff the above holds, else we get a contradiction as in a).

It is easy to see that this exhausts all possibilities for if $\exists$ k s.t. $\rho_i^j/\rho_i^k < q/p$, $j \in \Gamma_i$, $k \notin \Gamma_i$ then take $\alpha : \rho_i^\alpha \geq \rho_i^\beta$, $\alpha, \beta \notin \Gamma_i$

so that we have $\overline{\Delta\sigma}_{il}^s > 0$ with $l \in \Gamma_\alpha$.

Similarly for s=1 we get

c) $\overline{\Delta\sigma}_{ij}^{s=1} > 0$ with $j \in \Gamma_i$ if $\rho_i^j/\rho_i^k > p/q \quad k \notin \Gamma_i$.

so that $\sigma_{ij}^s \to 1$ for some $j \in \Gamma_i$ iff the above holds.

d) $\bar{\Delta}\sigma_{ij}^{s=1} > 0$ with $j \notin \Gamma_i$ if $\rho_i^j/\rho_i^k > q/p$ $k \in \Gamma_i$ and $\rho^j_{i/\rho_i^m} > 1, \forall m \notin \Gamma_i$

and $\sigma_{ij}^s \to 1$ for some $j \notin \Gamma_i$ iff the above holds.

We must note, as in 3.3.6. that we have $\bar{\Delta}\sigma_{ij}^s > 0$ for all $j \in \Gamma_\alpha$ where u_α maximizes ξ payoff at the next trial. But we get convergence to precisely one $j \in \Gamma_\alpha$ using the n-action theorem 1.6.4. with multiple optima, noting $q_i^{\beta,s} = q_j^{\beta,s}$ $\forall i,j \in \Gamma_\alpha$, $\forall s, \beta$.

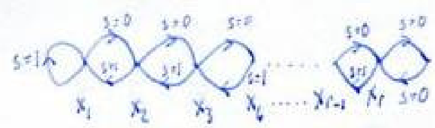
It is now easy to see that ${}_nL_1$ is stable and ${}_{m<n}L_1$ is not stable by lemma 3.4.2.

//

<u>Corollary</u> 3.4.5.

In $\mathcal{M}_n$, the automaton ${}_nL_r$ is stable iff

$\rho^1_{r-1}/\rho^2_{r-1} > q/p$ where :-



$x_i \in \Gamma_i$ $1 \leq i \leq r$.

<u>Proof</u>.

By the symmetry of the inequality, if it holds for arm Γ_1, then it is true for all arms Γ_α. Also $\rho^2_{r-1} > \rho^{k>2}_{r-1}$ since $\varphi_1^r \downarrow$ in r for ${}_nL_1$ by 3.4.2. a).

Then apply 3.4.4. for the result.

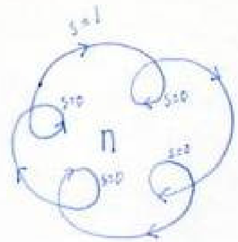
//

The extension of 3.3.7. b) requires the calculation of the equilibrium distribution for ${}_nL_2$ which has not been done. However, it still remains clear that $\bar{R}({}_nL_1) = \bar{R}({}_nL_2)$ when $\rho^1_{r-1}/\rho^2_{r-1} = q/p$, which we proved for n=2 in 3.3.2. d).

In general, the optimal v_{ij}^s automaton will not have the SOSA property although the $\mathcal{A}_0$ family may achieve close to optimal payoff. It is the discreteness of the formulation which causes the difficulties both here and in the general existence conjecture to be given later.

<u>Corollary</u> 3.4.6.

If $n\delta > 1$ then ${}_nL_1^c$ is a stable limiting structure.

$_nL_1^c$ ~ 

with convention, reward links on exterior, and penalty links on interior.

Proof.

Just note $K = (1 - n\delta)(q - p)$ changes sign, and we also check SOSA property is satisfied by $_nL_1^c$ by modifying analysis of 3.4.4.

For $q < p$ we actually get no change in stable structures, since sign changes all eventually cancel out. Thus 3.4.4 could be stated for all $q \in [0,1]$, with just the δ constraint. //

The $_nL_1^c$ is probably the unique limit, by symmetry requirements. However, in such enumeration problems of graph theory, there often seems to be the possibility of a subtle counter-example.

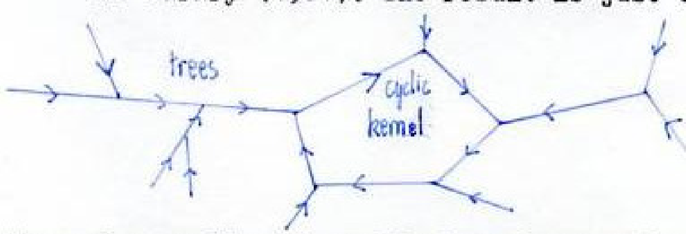
For fixed s, V_{ij}^s is a set of functional digraphs, and the following lemma gives the possible basic forms.

Lemma 3.4.7.

For each s, the graph V_{ij}^s consists of a set of disconnected subgraphs. Each has precisely one cyclic kernel with trees leading to it.

Proof.

See Harary (1964). The result is just observation. //



For s=0 we call each cyclic kernel an action switch.

For s=1 " " " " " the payoff.

We represent an arbitrary deterministic V_{ij}^s as:-

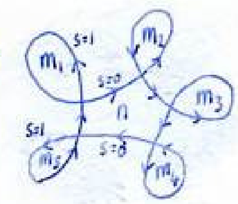
switch memory payoff

Such functional signed digraphs have not been investigated in the literature, but just from 3.4.7. , we see that we have a reasonably rich family of forms to adapt to a general environment $\mathfrak{M}$.

With a sufficiently large initial state space, we would expect to achieve near optimal memory depth, by combinatorial considerations of "cycle length" in each action class. However, such intuitive observations have not yet been rigorously proved. An analysis for a fixed finite state space of the asymtotic distribution over SOSA structures would appear extremely difficult to obtain, except for the proving of more general observations as above.

<u>Definition</u> 3.4.8.

i) An automaton is denoted <u>linear</u> if ${}_nL_{\underline{m}}$ $\sim$

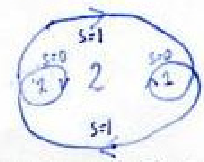


m_i = memory depth of i^{th} action.

n = number of actions in switch.

In particular ${}_2L_{\underline{m}}$ $\sim$

and ${}_2L^c_{\underline{1}}$ $\sim$



Previously we used ${}_nL_{\underline{r}} = {}_nL_{r\underline{1}}$ so all memory depths were equal, in $\mathfrak{M}_n$.

For arbitrary $\mathfrak{M}$ we cannot assert the existence of a limiting SOSA automaton but we can always achieve convergence using the rules $R_{\epsilon B}$. We then search for a SOSA automaton, yet we eventually converge to an unstable boundary if none exists. However, even if a SOSA automaton exists, under $R_{\epsilon B}$ rules we only have ϵ-optimality, so we might attain an unstable boundary. See 1.8. , rule 5).

ii) We use the notation $A_0(\mathfrak{M}) \triangleq$ the set of SOSA automata existing in $\mathfrak{M}$.

<u>Theorem</u> 3.4.9.

If $A_0(\mathfrak{M}) \neq \emptyset$ then $\sigma^s_{ij} \to v^s_{ij}$ for some SOSA automaton, under R_1 .

Proof.

We have seen that $\mathcal{A}_0(\mathfrak{M}) \neq \phi$ for all symmetric media and we shall indicate later why it is plausible that this is true for an arbitrary $\mathfrak{M}$.

Now $\Delta\sigma_{ij}^s(t) = \sigma_{ij}^s(t) \sum_\alpha w_i^\alpha(t)\, q_i^{\alpha,s} \sum_{k,\beta} \sigma_{ik}^s(t)\, \Delta_{\alpha\beta}\, \Theta_{jk}(\{\sigma_{ij}^s(t)\}) \left(q_j^{\beta,1} - q_k^{\beta,1}\right).$ 12.

We take the equilibrium drift by fixing σ_{ij}^s in some small neighbourhood N_{ij}^s of the boundary V_{ij}^s , and then apply 1.12.9. to each $\mathcal{B}_{(i,s)}$ associated with each σ_{ij}^s as discussed after 3.3.7. Thus we eliminate the transient effects, and consider the σ_{ij}^s - increment process w.r.t. the environmental equilibrium distribution in each state x_i which is +ve recurrent.

Thus $\overline{\Delta}\sigma_{ij}^s = \sigma_{ij}^s \sum_{\alpha,k,\beta} \Upsilon_i^\alpha\, q_i^{\alpha,s}\, \sigma_{ik}^s\, \Delta_{\alpha\beta}\, \Theta_{jk} \left(q_j^{\beta,1} - q_k^{\beta,1}\right).$ 13.

now $\Upsilon_i \sum_\alpha \xi_i^\alpha q_i^{\alpha,s} \Delta_{\alpha\beta} = c_i \Upsilon_i w_\beta^s$ where w_β^s =Pr(in E_β | s). $c_i = \sum_\alpha \xi_i^\alpha q_i^{\alpha,s}$.

Thus $\overline{\Delta}\sigma_{ij}^s = \sigma_{ij}^s c_i \Upsilon_i \sum_{\beta,k} w_\beta^s\, \sigma_{ik}^s\, \Theta_{jk}\left(q_j^{\beta,1} - q_k^{\beta,1}\right) = \sigma_{ij}^s c_i \Upsilon_i \sum_{\beta,k} \sigma_{ik}^s\, w_\beta^s\, R_{jk}^\beta.$ 14.

now SOSA $\Rightarrow$ $\sigma_{ij}^s \to 1$ only if u_j maximizes $\sum_\beta q_j^{\beta,1} w_\beta^s$.

Then $\sum_\beta R_{jk}^\beta w_\beta^s > 0,\ \forall k \notin \Gamma_j \Rightarrow \overline{\Delta}\sigma_{ij}^s > 0$ so we have stability and we can apply boundary learning theory as in 3.3.6. and 3.4.4., using 1.12.9.

Clearly, if u_j does not maximize expected payoff at the next trial, yet $\sigma_{ij}^s \to 1$ then we get a contradiction, since reinforcement $\mathcal{R}_0$ is optimal in a markov environment by 1.12.9.

//

Remarks 3.4.10.

i) We can replace 13) by $\overline{\Delta}\sigma_{ij}^s = \sigma_{ij}^s c_i \sum_{\alpha,k} \Psi_{k,i}^{\alpha,s} R_{jk}^\alpha.$

where $\Psi_{k,i}^{\alpha,s}$ = Pr(α, k at t+1 | Υ_i^β(t) and stimulus s),

and the stability of link σ_{ij}^s and hence $\mathcal{B}_{(i,s)}$ depends only on:-

$$\text{sign}\,(\bar{\Delta}\sigma^{s}_{ij}) = \text{sign}(\sum_{\alpha k} R^{\alpha}_{jk}\, \Sigma^{\alpha s}_{ki})$$

And boundary learning gives $\sigma^{s}_{ij} \to 1$ only if $\bar{\Delta}\sigma^{s}_{ij} \geq 0$ in arbitrarily small nbds of the boundary, under R_0. Note that we need the whole of the apparatus of chapter 1 to assert this. We need now only test for deterministic stability to obtain stability of the stochastic process, if we use R_0.

ii) We could by appropriate reinforcement achieve maximization over the next $r > 1$ trials, but it is our aim to keep the basic assumptions as natural as possible.

<u>Conjecture</u> 3.4.11.

a) For any $\mathcal{M}(\Delta_{\alpha\beta}, q^{\alpha s}_{u_i})$ $\exists$ a SOSA automaton.

b) The linear family $L^{(k_{ij})}_{\underline{m}}$ covers $\mathcal{M}$.

where m_i = memory depth of i^{th} action.

$k_{ij} = (k_{i1}, k_{i2})$ with k_{i1} = action we switch to at kernel under $s=0$ transition from u_i.

k_{i2} = action we switch to at kernel under $s=1$ transition from u_i.

so $k_{i2} = i$ under $m_i > 1$. (In many cases (k_{ij}) will be redundant.)

c) If $A_0(\mathcal{M}) \neq \phi$ then $\bar{R}(\nu^{s}_{ij}) \geq \max_i \underset{\sim}{e}.\underset{\sim}{q}_i$.

<u>Remarks</u>.

a) and b) go together since it is not possible to depart from $L^{(k_{ij})}_{\underline{m}}$ and still obtain SOSA automata. If a) were false, then it would indicate that the graphs ν^{s}_{ij} are acting as a "strait-jacket" and that we should reformulate to give a larger family.

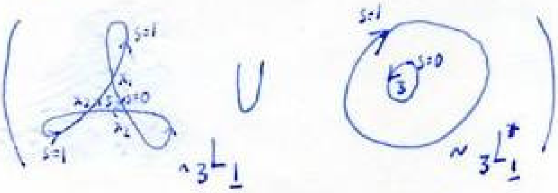
When $\Delta_{\alpha\alpha} \sim 1 \;\forall \alpha$ then we expect to be able to adjust $r_\alpha, \forall \alpha$. according to the equilibrium distribution e_α. Whilst if $\Delta_{\alpha\alpha} < 1/n$ with n= # media, then we adjust $k_{\alpha j}$ in an attempt to give a SOSA kernel. $(k_{ij}) \triangleq$ kernel switching matrix, $(m_i) \triangleq$ memory vector.

c) states that any limiting structure in a medium $\mathcal{M}$ will perform better than a singleton Π-cell. //

Examples 3.4.12.

a) For a cyclic 3-medium $\mathcal{M}_3^{cyc}(\Delta_{\alpha\beta}, q_{u_i}^{\alpha,s})$ with:-

$$\Delta_{\alpha\beta} = \begin{pmatrix} 1-\delta, & \delta, & 0 \\ 0, & 1-\delta, & \delta \\ \delta, & 0, & 1-\delta \end{pmatrix} \qquad \begin{matrix} q_i^{i1} = q \ , \ q_{j\neq i}^{i1} = p \\ q_i^{i0} = p \ , \ q_{j\neq i}^{i0} = q \end{matrix}$$

We have () as a covering for all δ, and $p<\frac{1}{2}$.

and for $p=0$, ${}_3L_{\underline{1}}$ is stable for $\delta < \frac{1}{2}$

$p=\frac{1}{2}$, ${}_3L_{\underline{1}}$ is stable for $\delta < (3-\sqrt{5})/2 \sim .766/2 = 0.382$.

It is interesting that the threashold is not at $\delta=\frac{1}{3}$.

${}_3L_{\underline{1}}$ is stable if $(1-\delta)(1-2q)(\delta^2(1-3q)+3\delta q-q) < 0$.

${}_3L_{\underline{1}}^*$ is stable if $(q-p)((3q-1)\delta^3-3q\delta^2+(3+pq)\delta-1) > 0$.

For $p > \frac{1}{2}$ we find that $({}_3L_{\underline{1}}^* \cup {}_3L_{\underline{1}}^A)$ is a covering for all δ.

For ${}_3L_{\underline{1}}^*$ we have $(*) \sim (k_{ij}) = \begin{pmatrix} 3,2 \\ 1,3 \\ 2,1 \end{pmatrix}$ the kernel switching matrix.

For ${}_3L_{\underline{1}}^A$ we have $(A) = \begin{pmatrix} 3,1 \\ 1,2 \\ 2,3 \end{pmatrix}$ so we cycle in the reverse direction to $\mathcal{M}_3^{cyc}$.

To obtain coverings we need to i) Calculate $\rho_i^\alpha = \gamma_i^\alpha / \gamma_i$

$$\gamma_i^\alpha = \sum_{s,j,\beta} \sigma_{ji}^s \, q_j^{\beta,s} \, \Delta_{\beta\alpha} \, \gamma_j^\beta$$

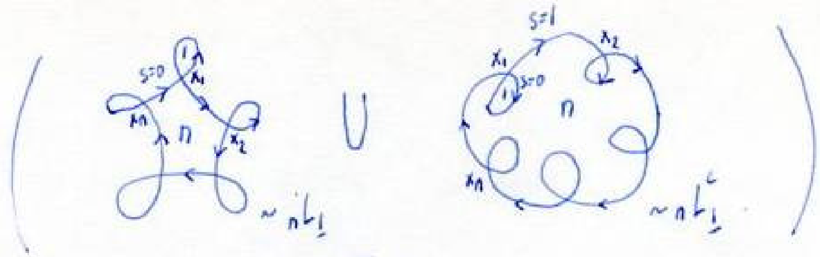
find ii) $\omega_\alpha^s = \sum_\beta \rho_i^\beta \Delta_{\beta\alpha} q_i^{\beta,s} \Big/ \left(\sum_\beta \rho_i^\beta q_i^{\beta,s}\right)$

and then ensure $\sigma_{ij}^s = 1$ only if u_j maximizes $\sum_\beta \omega_\beta^s q_{u_j}^{\beta,1}$.

We are just using 3.3.5. with ρ_i^α as starting point.

The calculations are lengthy, with difficulties arising in the cyclic case since ω_β^s terms all interact, whilst in the symmetric $\mathcal{M}_n$ we have ratios ω_i^s/ω_j^s determining structures.

b) For symmetric n-media $\mathcal{M}_n$, we have ${}_nL_{\underline{1}} \cup {}_nL_{\underline{1}}^c$ as a covering.



with $_nL_1$ stable for $n\delta \le 1$.

and $_nL_1^c$ stable for $n\delta \ge 1$.

//

Remarks. 3.4.13.

i) If we attempt to give a covering for $\mathcal{M}_n^{cyc}$, we get nth order polynomial inequalities in δ which are most difficult to handle. Thus it has not been possible to show whether

$({}_nL_1^A \cup {}_nL_1 \cup {}_nL_1^*)$ is a covering for $\mathcal{M}_n^{cyc}$.

ii) In section 3.9. we shall see how hierarchies of structures can use $\mathcal{M}$ coverings.

iii) In the case of $\mathcal{M}_3^{cyc}$ we actually obtain ${}_3L_1^A \cup {}_3L_1^*$ as a covering for all δ and p . If the calculations are correct, then it is of interest that no ${}_3L_1$ is required. However, for $\mathcal{M}_{n>3}^{cyc}$ we would still expect to require the full covering, as stated in i).

iv) The strategy of ${}_nL_1^*$ is of interest, since it has not appeared in the literature.

${}_nL_1^*$ ~

a) With $q > \frac{1}{2}$, then if we are in phase with $\mathcal{M}_n^{cyc}$ we shall receive s=1 with highest probability, so we switch through the actions x_i in the same order as the environment.

b) With $q > \frac{1}{2}$, then if we are out of phase we receive s=0 with highest probability so we switch through the actions in reverse order, since this is the quickest way to find the optimal action.

c) With $q < \frac{1}{2}$, we just have to avoid the one "bad" action so we attempt to be out of phase with $\mathcal{M}_n^{cyc}$, yet the ${}_nL_1^*$ is still SOSA.

v) Note that ${}_2L_1^A = {}_2L_1$, and that ${}_2L_1^{\bullet}$ just cycles $x_1 \to x_2 \to x_1 \ldots$ for all time, independent of stimulus. Thus only for $n > 2$ media do such automata exhibit useful adaptive behaviour.

3.5. Relationship between $\mathcal{A}_n$ and Likelihood Axis for a 2-Medium.

We shall consider in more detail the algorithm 3.3.5. and its relationship to our discrete structured automaton.

Lemma 3.5.1.

For 2-actions in $\mathcal{M}_2$ with arbitrary δ and q,

$\exists\ \omega_{max}$ and ω_{min} s.t. $\omega_{min} < \omega_1(t) < \omega_{max}$ for $t >> 0$.

where $\omega_1(t)$ = Pr(in E_1 at time t $|\ \omega_1$ (t-1), s(t-1), u(t-1)).

Proof.

We call the set $\mathcal{Z}_2 = \{\omega_1 : \omega_{min} < \omega_1 < \omega_{max}\}$. the operating zone for 2-actions.

Suppose $p, \delta < \frac{1}{2}$, then using the notation of 3.3.5.

$$\omega_1^{s=1}(t+1) = (\omega_1(t)(1-\delta)q + p\delta\omega_2(t)) / [\omega_1(t)q + \omega_2(t)q] = \omega_1(t) \quad \text{if :-}$$

$$(\omega_1)^2(2q-1) + \omega_1(1-2q+\delta) - p\delta = 0 \qquad 15)$$

and $\omega_{max} = ((2q-1+\delta) + (2q-1)(1-2\delta + (\delta/(1-2q))^2)^{\frac{1}{2}}) / 2(2q-1)$.

and for $\delta << \frac{1}{2}$, $\omega_{max} \sim 1 - \delta q/(2q-1)$.

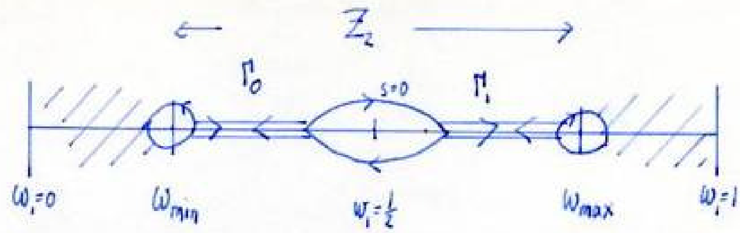
and similarly for $\omega_{min} = 1 - \omega_{max}$ by symmetry.

For $\delta = \frac{1}{2}$ we have $\omega_{min} = \omega_{max} = \frac{1}{2}$ so that $\mathcal{Z}_2$ is just a singleton point.

Whilst for $\delta > \frac{1}{2}$, $p < \frac{1}{2}$ we use $\omega_1^{s=0} = \omega_1$ to achieve bounds, so that if $\delta = 1$ then $\omega_{max} = 1/(1 + (p/q)^{\frac{1}{2}})$ and in general for $\delta \in [\frac{1}{2}, 1]$ just interchange q,p in 15).

Finally if $p > \frac{1}{2}$ we interchange q and p in all the bounds so

$$\omega_{max} \sim 1 - \delta p/(2p-1) \quad \text{if} \quad \delta << \frac{1}{2}. \qquad //$$



The Automaton ${}_2L_{\underset{\sim}{r}}$ embedded in Z_2.

Definitions 3.5.2.

i) The set of $\{\underset{\sim}{\omega}\}$ s.t. $\sum_\alpha \omega_\alpha (q_i^\alpha - q_j^\alpha) = 0$ for some j≠i and with $\sum_\alpha \omega_\alpha (q_i^\alpha - q_k^\alpha) > 0$ k≠j,i, is called the threashold set.

ii) The point $\omega = \lambda$ s.t. $\sum_\alpha \lambda_\alpha q_i^\alpha = \sum_\alpha \lambda_\alpha q_j^\alpha \ \forall i,j$ is called the action equilibrium point.

iii) For an n-medium the operating zone Z_n is defined to be the $\underset{\sim}{\omega}$ absorbing set, for the optimal bayesian algorithm 3.3.5.

In the case of $\mathcal{M}_2$, $\lambda = \varrho = \frac{1}{2}.\underline{1}$ and for $\mathcal{M}_n$ and $\mathcal{M}_n^{cyc}$ we have $\lambda = \varrho = \frac{1}{n}.\underline{1}$. It is for this reason that the symmetric and cyclic environments were treated in 3.4.12. , since it simplifies the analysis. Thus for $\mathcal{M}_2$ it is optimal to take u_1 if and only if $\omega_1 > \frac{1}{2}$.

Lemma 3.5.3.

For 2-actions Z_2 spans $\underset{\sim}{\varrho}$, where $\exists\, \omega'_\alpha < \varrho_\alpha < \omega''_\alpha$ with $\omega'_\alpha, \omega''_\alpha \in Z_2$.

Proof.

Suppose that $\underset{\sim}{\varrho}$ is not spanned by Z_2. Now by bayes' rule we must have $\overline{\omega_\alpha} = \varrho_\alpha$. So if $\omega_\alpha < \varrho_\alpha \ \forall \omega_\alpha \in Z_2$ we have a contradiction, so that Z_2 spans $\underset{\sim}{\varrho}$ as required, with $\underset{\sim}{e} = \underset{\sim}{e}\Delta$. //

It is unclear how this could be extended to give Z_n spans $\underset{\sim}{\varrho}$ or whether we can obtain the stronger result $\underset{\sim}{\varrho} \in Z_n$.

Lemma 3.5.4.

For 2-actions:-

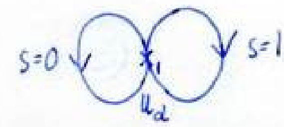
a) A structured automaton may evolve to a singleton action class Γ_α iff $\varrho_\alpha^s > \lambda_\alpha$ for s=0,1.

b) If $\underset{\sim}{\lambda}$ is not spanned by Z_2 then the maximum payoff is achieved by using a single action for <u>all</u> time.

<u>Proof</u>.

a) If $e^s_\alpha > \lambda_\alpha$ then define structure:-

$V^s_{11} = 1, \quad s = 0, 1.$



and this is clearly SOSA.

Suppose that $e^s_\alpha < \lambda_\alpha$ for s = 0 say, but that $\exists$ SOSA structure with equilibrium probabilities ξ^α_i = Pr(in E_α | in state i).

$\sum_i \xi^\alpha_i = e_\alpha$ so $\exists$ $\xi^\alpha_{i_0} \leq e_\alpha \leq \xi^\alpha_{i_1}$.

However, by continuity we have either $\{\xi^\alpha_{i_0}\}^{s=0} < \lambda_\alpha$ or $\{\xi^\alpha_{i_1}\}^{s=0} < \lambda_\alpha$, 16)

and hence we have a contradiction as this gives

$u_{\beta \neq \alpha}$ is optimal at the next trial, so the SOSA structure has multiple action classes.

Note that we get 16) since if $\omega'_1 > \omega''_1$ then $(\omega'_1)^s > (\omega''_1)^s$ iff $\Delta_{11} > \Delta_{21}$.

b) This is immediate from the definition of $\underset{\sim}{\lambda}$.

//

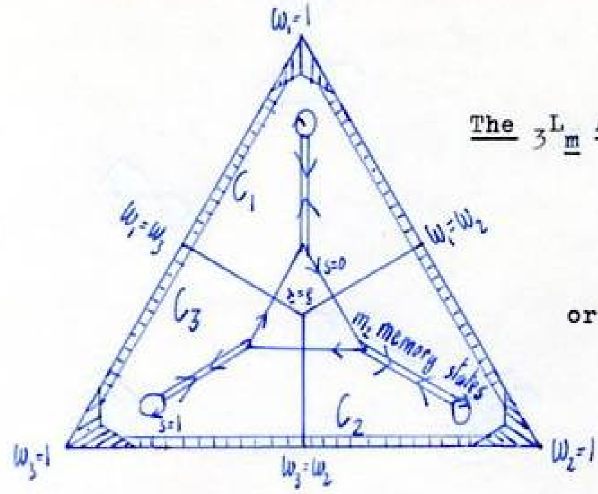
For a singleton action class Γ_α , we have $\bar{R}(V^s_{ij}) = \underset{\sim}{e}.\underset{\sim}{q}_\alpha$, as for a Π-cell. Then a) gives us conditions under which we have a stable limiting automaton in certain dynamic $\mathcal{M}$. The above lemmas have analogous forms for the n-action likelihood simplex but as yet, it is not possible to give such a precise characterization of their partitioning for automata learning.

3.6. <u>The n-Medium Likelihood Simplex</u>.

<u>Definition</u> 3.6.1.

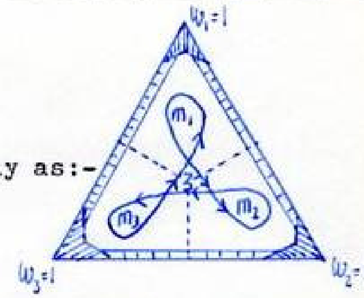
Let C_α be that part of the $\underset{\sim}{\omega}$ simplex in which u_α is the optimal action.

We first consider the symmetric medium $\mathcal{M}_n$ and the relation between the SOSA property and the $\underset{\sim}{\omega}$ likelihood simplex.



<u>The $_3L_m$ Automaton embedded in its ω - Simplex.</u>

or symbolically as:-



The boundary barriers in these diagrams are <u>not</u> intended to be exact representations.

<u>Observations</u> 3.6.2.

a) The action switch forms around $\underset{\sim}{\lambda} = \underset{\sim}{\varepsilon}$ for $_n\underset{\sim}{L_m}$, in M_n.

b) The Γ_α payoff is in C_α.

c) The memory states of u_α are in C_α and they extend from the action switch towards $\omega_\alpha=1$.

d) The $\underset{\sim}{\omega}$ -operating zone Z_n is around $\underset{\sim}{\lambda}=\underset{\sim}{\varepsilon}$ and it extends towards barriers around each $\omega_\alpha=1$ and $\omega_\alpha=0$.

If $\omega_\alpha=1,\ \omega^s_\beta = q_i^{\alpha,s}\Delta_{\alpha\beta}/q_i^{\alpha,s} = \Delta_{\alpha\beta}$ under u_i.

Thus $\omega^s_\alpha = \Delta_{\alpha\alpha}$ under any u_i, and this prevents Z_n from actually reaching $\omega_\alpha=1$. Similarly for $\omega_\alpha=0$.

e) If $n\delta=1,\ Z_n=\{\underset{\sim}{\omega}=(\frac{1}{n}\underline{1})\}$ whilst if $n\delta>1$ we have $_nL^c_{\underset{\sim}{1}}$ stable around $\underset{\sim}{\lambda}$ within $Z_n(M_n)$. d) and e) are the generalization of 3.5.1. from Z_2 to Z_n.

<u>Remarks</u>. 3.6.3.

a) The $\underset{\sim}{\lambda}$ point corresponds to the equilibrium strategy for "nature" if the process is considered as a game, since $\sum_\alpha \lambda_\alpha(q_i^\alpha - q_j^\alpha)=0\ \forall j$.
Then with n-actions, each of which is optimal in one of n-media, we have a "completely mixed" strategy. For $\bar{u}$-cell games we had $\sum_\alpha \lambda'_\alpha(q_{\alpha i}-q_{\alpha j})=0,\ \forall j$ giving equilibrium values in 2.3.1.

If $\underset{\sim}{\omega}=\underset{\sim}{\lambda}$ then all strategies u_i have the same expected payoff.

If $(\bar{u}_i)=(\lambda'_i)$ then all strategies u_j for player 2 have the same expected payoff.

b) If $(q_i^\alpha - q_j^\beta) < \epsilon \quad \forall i,\alpha,\beta$ then $\mathcal{Z}$ becomes an arbitrarily small region spanning $\underset{\sim}{e}$. For $\mathcal{M}_2$ we actually proved in 3.5.1 that $(w_{max} - w_{min}) \to 0$ as $(q-p) \to 0$, and the spanning of $\underset{\sim}{e}$ just follows from $\overline{\underset{\sim}{w}} = \underset{\sim}{e}$.

When $q_i^\alpha = q_j^\beta \quad \forall i,j,\alpha,\beta$ then $\underset{\sim}{w} = \underset{\sim}{e}$ and the statement above concerning the smallness of $\mathcal{Z}$ follows by the continuity implicit in bayes' rule.

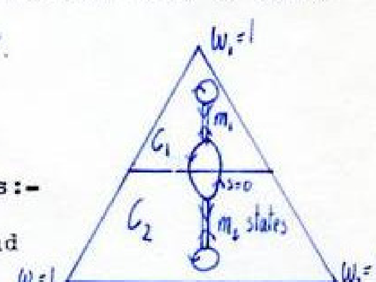
If $\underset{\sim}{e} \neq \underset{\sim}{\lambda}$ then using 3.5.4. we can achieve singleton action class Γ_α iff $e_\alpha^s > \lambda_\alpha$ s=0,1 so that $\underset{\sim}{e} \in C_\alpha$. In this sense, the structure emerges from $\underset{\sim}{w} = \underset{\sim}{e}$ and as q_i^α separate we reach $\underset{\sim}{\lambda}$ and an action switch forms, with memory extending out towards the vertices of the $\underset{\sim}{w}$ -simplex, always keeping within $\mathcal{Z}$.

c) Just as the $\underset{\sim}{\lambda}$ point gives nature an equilibrium strategy against the automaton, so with $q_i^\alpha = q_j^\beta \quad \forall i,j,\alpha,\beta$ we find that nature has no influence on the automaton. Here we are representing the environmental actions by the states of $\mathcal{M}$.

d) We have assumed that there is a unique optimal action for each environment.

i) If there are n+r actions in an n-medium then take u_i s.t. $q_{u_i}^\alpha \geq q_{u_j}^\alpha \quad \forall j$ for each α.

ii) If there are n-r actions in an n-medium, then we again just take u_i s.t. $q_{u_i}^\alpha \geq q_{u_j}^\alpha \quad \forall j$ for each α.



So with 2-actions in a 3-medium, we embed $_2L_{\underset{\sim}{m}}$ in the $\underset{\sim}{w}$-simplex as:-

The threashold is the line $\sum_\alpha w_\alpha q_1^\alpha = \sum_\alpha w_\alpha q_2^\alpha$, and if $q_{u_i}^2 = q_{u_i}^3$ for each u_i, then we effectively just have a 2-medium likelihood axis.

In general we obtain a hypersurface of dimension n-2 in an n-medium likelihood simplex, as the threashold set.

Summary 3.6.4.

We characterize the $\underset{\sim}{\omega}$ likelihood simplex by:-

i) The threashold set $\{\underset{\sim}{\omega} : \underset{\sim}{\omega}.\underset{\sim}{q}_i = \underset{\sim}{\omega}.\underset{\sim}{q}_{j\neq i} \geq \underset{\sim}{\omega}.\underset{\sim}{q}_k \ \forall h \neq j, i$ for each $u_i, u_j\}$,
and the $\underset{\sim}{\lambda}$ point, where equality holds in the above in the case of a completely mixed equilibrium strategy.

ii) The environment equilibrium point $\underset{\sim}{e}$. $e_\alpha = \sum_\beta e_\beta \Delta_{\beta\alpha}$.

iii) The operating zone Z which is the $\underset{\sim}{\omega}$ absorbing set for the optimal bayesian algorithm 3.3.5.

We have now informally analysed the $\mathcal{A}_0$ family, and so we turn now to networks of Π-cells $\mathcal{G}'(\otimes)$.

3.7. Networks of Π-cells.

Instead of a unique action at each state we now have a Π-cell $\otimes_k$ for each set of states x_i with $i \in \Gamma_k$. There is now an alternative way of defining an underlying graph.
Let $\beta_{i,j}^{s_z,k}(t) = \mathrm{Prob}(x_i(t) \to x_j(t+1) \mid s_z(t)$ and $u_k(t))$. Now since we asymtotically obtain the same family of structures, although with a different distribution, the greatly increased number of graphs $(2 \to 2n)$ does not seem justified. Indeed we shall see that we obtain all the desired results using $\sigma_{ij}^{s_z}$ as defined in 3.1.1.

The aim of this section is to show that even though we initially have all states possessing the same distribution over all actions, the actions automatically partition themselves through the tool of uniform learning. The actions interact through the equilibrium distribution γ_i^α of the structure. We first charaterize the $\mathcal{B}_0$ family if structures in 3.7.1. and then relate this to $\mathcal{A}_0(\mathcal{M}_n)$ in 3.7.2. and 3.7.3. , whilst 3.7.4. considers the general $\mathcal{A}_0(\mathcal{M})$.

As motivation for this abstract treatment we may consider the singleton Π-cells as entities which cluster together for their mutual benefit and then differentiate in their actions, in order to give structural adaptation in the markovian medium $\mathcal{M}$.

Theorem 3.7.1.

Let i be a +ve recurrent state in the limit $t \to \infty$ as defined in 3.2.1. , then under R_0 and in environment $\mathcal{M}$:-

a) $\sigma_{ij}^{s} \to 1$

with $i \in \Gamma_k$

and b) $\Pi_m^{k} \to 1$

only if $\max_1 \xi R$ w.r.t. the equilibrium distribution $\xi_i^{\alpha}(v_{ij}^{s})$, for $\sigma_{ij}^{s} \to 1$

and $\max_0 \xi R$ w.r.t. " " " $\xi_{\Theta_k}^{\alpha}$, " $\Pi_m^{k} \to 1$

We denote $\{v_{ij}^{s}, y_m^{k}\}$ family as $B_0(\mathcal{M})$, where $\sigma_{ij}^{s} \to v_{ij}^{s}$ and $\Pi_m^{k} \to y_m^{k}$.

And where above we have used:-

i) $\xi_{\Theta_k}^{\alpha} = \sum_{i \in \Gamma_k} \gamma_i^{\alpha} / \sum_{i \in \Gamma_k} \gamma_i$.

ii) $\max_r \xi R$ holds if we maximize the expected reward over the next r trials w.r.t the given environment equilibrium distribution.

Proof.

This is the basic convergence theorem for $B_0(\mathcal{M})$, and it is possible that a stable limit does not exist, as for $A_0(\mathcal{M})$. We shall assume here that $B_0(\mathcal{M}) \neq \phi$.

Note that $\max_1 \xi R$ is precisely the SOSA property of 3.3.8. and that $\max_0 \xi R$ corresponds to the result 1.12.9. in which a Π-cell asymtotically takes that action which maximizes its average payoff.

We essentially mimic 3.4.9. and replace $q_i^{\alpha,s}$ by $\sum_j \Pi_j^{k} q_j^{\alpha,s}$. throughout. We could write out $\Delta\sigma_{ij}^{s}, \Delta\Pi_j^{i}$ for the actual process as in 3.3. but the equations are very similar, with $\underset{\sim}{\Pi}^{k} \underset{\sim}{q}^{\alpha,s}$ replacing $q_i^{\alpha,s}$.

Consider the process in an arbitrarily small neighbourhood of the boundary absorbing set, in the usual way. Let the equilibrium distribution for the fixed process be $\gamma_i^{\alpha}(\sigma_{ij}^{s}, \Pi_j^{i})$.

Now $$\bar{\Delta}\Pi_m^{k} = \Pi_m^{k} \sum_{\alpha, n \in \Gamma_k} \gamma_n^{\alpha} \sum_i \Pi_i^{k} \theta_{im}(\underset{\sim}{\Pi}^{k}) (q_m^{\alpha,1} - q_i^{\alpha,1}). \qquad n \in \Gamma_k$$

$$= \Pi_m^{k} \Big(\sum_{n \in \Gamma_k} \gamma_n\Big) \sum_{\alpha, i} \xi_{\Theta_k}^{\alpha} \Pi_i^{k} \theta_{im}(\underset{\sim}{\Pi}^{k}) (q_m^{\alpha,1} - q_i^{\alpha,1}). \qquad 17).$$

and $\overline{\Delta}\sigma^s_{ij} = \sigma^s_{ij} \sum_{\alpha,k,\beta} \gamma^\alpha_i \left(\sum_m \bar{\Pi}^k_m q^{\alpha s}_m \right) \sigma^s_{ik} \Delta_{\alpha\beta} \theta_{jk}(\{\sigma^s_{ij}\}) \left(\sum_m \bar{\Pi}^a_m q^{\beta 1}_m - \sum_m \bar{\Pi}^b_m q^{\beta 1}_m \right)$ 18).

where $i \in \Gamma_k$, $j \in \Gamma_a$, $k \in \Gamma_b$.

From 17) we get $\bar{\Pi}^k_m \to 1$ only if $\sum_\alpha \rho^\alpha_{\theta_k} R^\alpha_{mi} \geq 0 \quad \forall i$

and hence $\max_0 \rho R$ w.r.t. $\rho^\alpha_{\theta_k}$ by 1.12.9.

For σ^s_{ij} we consider $\{\bar{\Pi}^k_m\}$ arbitrarily close to the boundary, in some nbd N^k_m of y^k_m say, so that $|\bar{\Pi}^k_m - y^k_m| < \epsilon^k_m$.

Then $\overline{\Delta}\sigma^s_{ij} = C_i \sigma^s_{ij} \sum_{m,k,\beta} \omega^s_\beta \sigma^s_{ik} \theta_{jk} \left(\bar{\Pi}^a_m - \bar{\Pi}^b_m \right) q^{\beta 1}_m$ 19)

where $C_i \omega^s_\beta = \sum_\alpha \rho^\alpha_i \sum_m \left(\bar{\Pi}^k_m q^{\alpha s}_m \right) \Delta_{\alpha\beta}$.

and $C_i = \sum_\alpha \rho^\alpha_i \left(\sum_m \bar{\Pi}^k_m q^{\alpha s}_m \right)$.

and further sign($\overline{\Delta}\sigma^s_{ij}$) > 0 iff sign($\sum_{\beta,k} \Psi^\beta_k R^\beta_{ab}$) > 0

with $k \in \Gamma_b$, $j \in \Gamma_a$.

and $R^\beta_{ab} = \sum_{k,m} \theta_{jk} q^{\beta 1}_m \left(\bar{\Pi}^a_m - \bar{\Pi}^b_m \right)$

and from 19) we get $\max_1 \rho R$ w.r.t. ρ^α_i as we require

$$\sum_\beta R^\beta_{ab} \omega^s_\beta \geq 0 \quad \forall\ \theta_b.$$

Convergence follows as in 3.4.9. , using a β-staircase construction and boundary learning theory in a markov medium, to give a form of 1.12.9. with the required equilibrium distribution for $\rho'(\theta)$.

//

Here v^s_{ij} has the SOSA property, but we have the additional constraint given on y^k_m , so it is unclear whether $A_0(M) = B_0(M)$. I shall indicate why I believe $B_0 \subset A_0$, but it is conceivable that $A_0 \not\subset B_0$ due to difficulties with discreteness.

We may find that u_j maximizes the expected reward at the next trial w.r.t. ρ^α_i , yet the Π-cell θ_k which executes this action at the next trial is unstable with respect to $\rho^{\alpha}_{\theta_k}$. However, no case of this has yet been found so the problem remains unresolved, and it seems clear that most structures will be well behaved.

The partitioning of actions is a stability result of a higher order than the boundary learning results we have so far considered. Consequently I shall present the analysis as a series of conjectures with an outlined "proof", since a truly rigorous proof would seem to be a deep excursion into probability theory.

Conjectures 3.7.2. and 3.7.3. are the results for $\rho'_j(\theta)$ which correspond to those for $\rho'_j(\theta)$ in 3.4.4. The corresponding extension of 3.4.9. is then 3.7.4. The π-cells themselves effectively become "mixed actions" consisting of a distribution over the "pure" actions.

Conjecture 3.7.2.

In M_n under R_0 :-

$$\pi^k_j \to 1 \Rightarrow \pi^m_j \not\to 1 \quad \forall m \neq k$$

Sketch.

This is a form of exclusion principle, so that only one π-cell is allowed in each memory.

Suppose that u_j is the action which maximizes the expected reward at the next trial so that $\sum_\alpha w^s_\alpha (q^\alpha_{u_j} - q^\alpha_{u_k}) \geq 0 \ \forall k$.
Then it is optimal to choose θ_k s.t. $\pi^k_j \geq \pi^m_j \ \forall m$. Hence the process is unstable if both θ_k and θ_m have u_j as the limiting action. Using R.W. theory it should be possible to prove that # axis crossings of $\pi^k_j = \pi^m_j$ is a.s. finite. Then after the last crossing θ_k say will always be the optimal π-cell to pick if u_j is required.

Now use boundary learning to give $\overline{\Delta\sigma^s_j} > 0$ for all $j \in \Gamma_k$, if θ_k gives $\max_{1\ell} \rho R$, when at the present trial we are in x_i.

For a rigorous formulation we need to consider $\psi^\alpha_i(\sigma^s_{ij}, \pi^k_m)$, so that for the instability in $\otimes_k$ and $\otimes_{m\neq k}$, when u_j is optimal we would expect $\gamma^{\alpha}_{\otimes_k}\uparrow$ and $\gamma^{\alpha}_{\otimes_m}\downarrow$ if $\pi^k_j \geq \pi^m_j \ \forall m$, $q^{\alpha,i}_{u_j} \geq q^{\alpha,i}_{u_k} \ \forall u_k$. //

Conjecture 3.7.3.

In $\mathcal{M}_n$ under R_0 :-

a) ${}_nL_{\underline{1}}$ is in $\mathcal{B}_0(\mathcal{M}_n)$.

b) ${}_{m<n}L_{\underline{1}} \notin \mathcal{B}_0(\mathcal{M}_n)$.

Sketch.

Using 3.7.2. we have one π-cell on each memory, so that if ${}_{m<n}L_{\underline{1}}$ is stable we have $\otimes_r$ say omitted. However, we know ${}_{m<n}L_{\underline{1}}$ is not SOSA by 3.4.2. , and hence given s=0, it will be optimal to make a transition to some $\otimes_r$ not "included in the structure". Only ${}_nL_{\underline{1}}$ is in $\mathcal{B}_0(\mathcal{M}_n)$, as it is SOSA and $\otimes_r$ is stable by inspection of $\rho^{\alpha}_{\otimes_r}$. //

This conjecture 3.7.3. is certainly true in itself, but it cannot be stated as a lemma since it relies on 3.7.2. for access to all $\otimes_r$ which are required.

Now let ${}_nK^s_r$ be the complete digraph on rn states with n π_n-cells and each link $\sigma^s_{ij}(0)$ having probability 1/rn . Then we have 3.7.2. and 3.7.3. giving ${}_nK^s_r \xrightarrow{\mathcal{M}_n} {}_nL_{\underline{m}}$ as a mathematical formalization of:- chaos $\xrightarrow{\text{environment}}$ structure adapted to its environment. The next conjecture considers the problem of π-cell differentiation for general $\mathcal{M}$.

Conjecture 3.7.4.

In $\mathcal{M}$ under R_0 :-

a) $\pi^k_j \to 1 \Rightarrow \pi^m_j \not\to 1 \ \forall m \neq k$

b) $\mathcal{B}_0(\mathcal{M}) \subset \mathcal{A}_0(\mathcal{M})$.

Sketch.

In 3.7.2. we only had to consider the optimal actions, but for general $\mathcal{M}$, we must consider all u_i. It is possible that $\bar{\pi}_i^k > \bar{\pi}_i^m$ with u_i optimal at the next trial, yet ϑ_k need not be the optimal $\bar{\pi}$-cell to pick. ϑ_k may allocate probability $(1-\bar{\pi}_i^k)$ to the action that $\min_1 \mathcal{E} R$, whilst ϑ_m allocates $(1-\bar{\pi}_i^m)$ to the action which $\max_1 \mathcal{E} R$ if u_i is excluded.

However, we can still consider axis crossings of $\mathcal{E}$ payoff (ϑ_k) against $\mathcal{E}$ payoff (ϑ_m) at the next trial, when we wish to use u_i and obtain the instability as in 3.7.2. which prevents $\bar{\pi}_i^k, \bar{\pi}_i^m \to 1$.

Thus using boundary learning we can consider the equilibrium $\gamma_i^{\nu}(\sigma_{ij}^{s}, \bar{\pi}_m^k)$ and still obtain the exclusion principle, after proving that the $\#$axis crossings $\overset{a.s.}{\to} \infty$ for $\mathcal{E}$ payoff(ϑ_k) - $\mathcal{E}$ payoff(ϑ_m) for each pair of $\bar{\pi}$-cells and each action u_i. (Construct a s/mg each side of the unstable axis.)

If $\mathcal{M}$ is static then for 3.2.2. we have that there exists a unique ϑ_k at $t = \infty$.

b) It is assumed that we have a structured automaton with every action u_i replaced by a $\bar{\pi}$-cell which incorporates all actions. If $\mathcal{B}_0 \neq \mathcal{A}_0$ let $\nu_{ij}^{s} \in \mathcal{B}_0 \setminus \mathcal{A}_0$ now ν_{ij}^{s} is SOSA, so that the only possibility is that $\exists\ u_v$ s.t. this is not included in ν_{ij}^{s} , yet would be in $\mathcal{A}_0$. However, just as in 3.7.3. , the remaining ϑ_k have $\bar{\pi}_v^k > 0$, and these are used if u_v is optimal at the next trial, just as if we had a structured automaton with fixed actions evolving to $\mathcal{A}_0$.

Hence we have a contradiction unless $\nu_{ij}^{s} \in \mathcal{A}_0$. Again it is a) which is difficult to prove, whilst b) is fairly immediate once we have a $\bar{\pi}$-cell available for each action in case it is required.

//

Networks of $\bar{\pi}$-cells would seem to act as a reasonable model for cellular differentiation. Waddington (1967) gives a useful basis for the biological framework.

It is perhaps useful to imagine the $\bar{\pi}$-cells of $f_j'(\vartheta)$ as initially representing white light, with a spectrum of colours (actions).

Then as $\mathcal{G}'(\mathcal{B})$ evolves, it adopts the colour most expedient in each environment, just as some animals, like the chameleon, change their colour according to their surroundings. Thus the white light within the e-m spectrum will partition itself into its component colours (frequencies) according to their uses within the environment.

The interation between $\mathcal{G}$ and $\mathcal{M}$ enables the environment to "pull out" actions with certain properties from a "pool of actions", as they are required.

We can prove an analogue of $U\gamma = \gamma$ for structured automata, although the result has not been of use so far in generating bounds on absorption probabilities.

<u>Theorem</u> 3.7.5.

a) $\lim\limits_{n\to\infty} U^n \gamma_{v^s_{ij}}(\sigma^s_{ij}, \omega^\alpha_i) = \gamma_{v^s_{ij}}(\sigma^s_{ij}, \omega^\alpha_i)$

b) $\Delta\gamma_{v^s_{ij}} \equiv 0.$

where $\gamma_{v^s_{ij}}(\sigma^s_{ij}) = 1$ if $\sigma^s_{ij} = v^s_{ij}$ with γ^s_{ij} a deterministic SOSA automaton.
$\qquad\qquad = 0$ if $\sigma^s_{ij} = (v^s_{ij})^*$ " $(\gamma^s_{ij})^*$ " " "

and $U\gamma_{v^s_{ij}}(\sigma^s_{ij}(t), \omega^\alpha_i(t)) = \mathcal{E}\left(\gamma_{v^s_{ij}}(\sigma^s_{ij}(t+1), \omega^\alpha_i(t+1)) \mid \{\sigma^s_{ij}(t)\}, \{\omega^\alpha_i(t)\}\right).$

$\Delta\gamma_{v^s_{ij}} = U\gamma_{v^s_{ij}} - \gamma_{v^s_{ij}}$

$\gamma_{v^s_{ij}}(\sigma^s_{ij}(0), \omega^\alpha_i(0)) = Pr\left(\sigma^s_{ij}(t) \to v^s_{ij} \mid \sigma^s_{ij}(0), \omega^\alpha_i(t)\right).$

$\omega^\alpha_i(t) = Pr(\text{in } E_\alpha(t) \text{ and } x_i(t) \text{ at time } t).$

<u>Proof</u>.

a) $\lim\limits_{n\to\infty} U^n \gamma_{v^s_{ij}} = 1.\gamma_{v^s_{ij}} + 0(1 - \gamma_{v^s_{ij}}) = \mathcal{E}(\gamma_{v^s_{ij}}(\sigma^s_{ij}(\infty), \omega^\alpha_i(\infty)) \mid \mathcal{F}_0)$

where it is assumed that $A_0(\mathcal{M}) \neq \phi$ or else R_{EB} is used.

b) Clearly $\lim\limits_{n\to\infty} U^n\gamma_{v^s_{ij}} = \gamma_{v^s_{ij}}$ and if $\Delta\gamma_{v^s_{ij}} \neq 0$ we must get a contradiction, for $\lim\limits_{n\to\infty} U^n\gamma_{v^s_{ij}} = \lim\limits_{n\to\infty} UU^{n-1}\gamma_{v^s_{ij}} = U \lim\limits_{n\to\infty} U^{n-1}\gamma_{v^s_{ij}} = U\gamma_{v^s_{ij}} = \gamma_{v^s_{ij}}$.

We can interchange the U operator and the limit $n \to \infty$ since as in 1.12.1. $U^n \gamma$ converges uniformly to γ, even though γ itself will have discontinuities, under R_0 reinforcement. So we could verify that for each $\epsilon > 0$ $\exists N$ s.t. $n > N$ gives $|U^n\gamma - \gamma| < \epsilon$. //

Remarks 3.7.6.

i) If V^s_{ij} is the unique SOSA automaton then clearly $\gamma_{V^s_{ij}} \equiv 1$ under R_0, if $\sigma^s_{ij}(0) \in (0,1)$, but generally we may obtain the "same" structure in many different ways by relabelling the states in each Γ_{V_i}. If we wish $Pr(\sigma^s_{ij} \to {}_2L_1)$ we have to enumerate the graphs, with labelled states, that give ${}_2L_1$ from the initial state space $\mathcal{X}$.

ii) Thus $\gamma_{V^s_{ij}}$ is the fixed functional solution of the operator U. Intuitively, this follows since in the limit $t \to \infty$, γ is an indicator function, assuming we have boundary convergence, with a probabilistic mass of $\gamma_{V^s_{ij}}(\sigma^s_{ij}(0), \omega^\alpha_i(0))$ on the set which has $\gamma_{V^s_{ij}}(\sigma^s_{ij}(\infty), \omega^\alpha_i)$ as unity. This is the fundamental equation of reinforcement learning, and if γ is defined appropriately, we shall achieve $\Delta\gamma \equiv 0$, in all models using U.L. π-cells, when boundary convergence is assured. Thus there would be a generalisation of 3.7.5. to the hierarchies $\mathcal{G}^n(\theta)$ of 3.9. and the games $\prod_{i=1}^{n} \mathcal{G}^1_i(\theta)$ of 3.10. So given any evolving stochastic automaton $\mathcal{G}$ with absorbing structures defined in some environment $\mathfrak{M}$, we can immediately form the functional equation which generates the absorption probabilities over the structures.

iii) This is essentially a discrete time result of probabilistic potential theory, where instead of a diffusion and Laplace's equation $\nabla^2\phi = 0$, we have stochastic difference equations and $\Delta\gamma = 0$. Thus any evolving stochastic automaton can be viewed as possessing a certain potential at any time during its environmental adaptation. Such equations are treated in an abstract form by Meyer (1966). We can then consider a reward (penalty) stimulus as the reception of a +ve(-ve) charge, say, with sign depending on boundary conditions.

iv) It is of interest to briefly note the relation between the complementary pairs (reward, penalty), (survival, extinction), (+ve, -ve), (structure, chaos), (fixed, fluid) and the Ancient Chinese philosophy of Yin and Yang. Here <u>all</u> such fundamental duals are united through the Yin-Yang symbol,

where the spots indicate that each side of any dual pair contains a little of the other. Gardener (1964) relates asymmetry in molecules to the possible origin of life, and considers the above symbol in relation to contemporary science.

<u>Corollory</u> 3.7.7.

$$\Delta\gamma_{g(\circledast)}(\sigma^s_{ij}, \pi^i_j, \omega^\alpha_i) \equiv 0.$$

with $\gamma_g(g^*) = \delta_{gg^*}$.

<u>Proof</u>.

Again $\lim_{n\to\infty} U^n\psi_g \to \gamma_g$ with $\psi_g(g^*) = \delta_{gg^*}$. g, g^* are deterministic limiting structures

so $\lim_{n\to\infty} U^n\gamma_g \to \gamma_g = U\gamma_g$ by the same reasoning as in 3.7.5.

//

We have reduced the analysis of structured automata, $g'(\circledast)$, and networks of π-cells, $g'(\circledast)$, evolving under R_0 in M, to the investigation of:-

i) sign $(\overline{\Delta}\sigma^s_{ij})$. ii) sign $(\overline{\Delta}\pi^k_m)$. iii) $\Delta\gamma_g \equiv 0$.

in the usual notation.

3.8. <u>π-cell Controllers and "Blueprint" Learning</u>.

We may have several evolving automata and we wish to pick the one with the highest average payoff. We take a π_m -cell $\circledast_c$ operating under R_0 which samples the $s_2(t)$ stimulus of the r^{th} automaton A_r, as its r^{th} action u_r at time t, $1 \leq r \leq m$.

$\circledast_c$ is then called a <u>π-cell</u> <u>controller</u>.

<u>Theorem 3.8.1</u>.

For controller $\circledast_c$:-

a) $\exists$ r s.t. $\pi_r \to 1$.

b) $\Pi_k \to 1$ only if $\overline{R}(\lim_{t\to\infty} A_k(t)) \geq \overline{R}(\lim_{t\to\infty} A_s(t))$, $\forall s$

where we assume $\mathcal{R}_0(\mathcal{M}) \neq \phi$ so that $\lim_{t\to\infty} A_r(t)$ has the SOSA property.

Proof.

We let $A_k(t)$ have transitions $\sigma^s_{ij}(k,t)$ and $\lim_{t\to\infty} A_k(t)$ is some SOSA automaton with transitions $\nu^s_{ij}(k)$ and equilibrium distribution $\gamma^\alpha_i(k)$. We could always ensure that $\lim_{t\to\infty} A_k(t)$ exists by using $\mathcal{R}_{\epsilon B}$ instead of $\mathcal{R}_0$ as in 3.4.9.

We have $\mathcal{B}_0$ in a markovian environment generated by $\prod_{r=1}^{m} A_r$
Now apply 1.12.9. to give $\Pi_i \uparrow 1$ iff $\underset{\sim}{e}.\underset{\sim}{q}_i > \underset{\sim}{e}.\underset{\sim}{q}_{j\neq i}$ where the equilibrium distribution $\underset{\sim}{e}$ now becomes that for the A_r. Now combining this with 1.9.1. so that we need only consider the limiting probabilities as $t \to \infty$, we obtain:-

$$\Pi_r \uparrow 1 \text{ only if } \sum_{\alpha,i} \gamma^\alpha_i(r) q^\alpha_i \geq \sum_{\alpha,i} \gamma^\alpha_i(s) q^\alpha_i \quad \forall s$$

If $\overline{R}(A_k(\infty)) = \overline{R}(A_h(\infty))$ for A_k and A_h say, then we still obtain convergence since we can relate the process to that of multiple optima in a static medium (1.6.4.) and use a β-staircase at each boundary to give $\#\text{upcrossings} \overset{a.s.}{<} \infty$.

Clearly the $A_r(t)$ could be π-cell networks $\mathcal{G}(\mathcal{B})$ and the result still follows if $\mathcal{B}_0(\mathcal{M}) \neq \phi$, or else we use $\mathcal{R}_{\epsilon B}$. //

This result is not elegant but a typical example of forced centralised learning, whilst we should prefer interactive behaviour between automata as the mechanism which increases their mutual adaptation.

If we are given a "blueprint" for any finite automaton then it is easy to formulate a learning process which converges to this structure. I am putting this result here because Suppes(1969) obtains a similar result with a different stimulus-response model. Using $\mathcal{R}_0$ the theorem is easily proved. Kieras (1976) considers the theorem of Suppes in great detail and extends the limiting family of automata.

Theorem 3.8.2.

Given $\{\nu_{ij}^{s_\alpha}\}$ as a finite deterministic automaton where:-

$s_\alpha(t)$ = input stimulus which gives transition $i \rightarrow j$ if $x_i(t)$ and $\nu_{ij}^{s_\alpha} = 1$

Then if we put $q_{ij}^{s_\alpha} > q_{ik}^{s_\alpha}$ whenever $\nu_{ij}^{s_\alpha} = 1$ then $\sigma_{ij}^{s_\alpha} \rightarrow \nu_{ij}^{s_\alpha}$ under R_1.

Here $q_{ij}^{s_\alpha}$ = Pr($\sigma_{ij}^{s_\alpha}$ is positively reinforced (reward), if we receive $s_\alpha(t)$ when in $x_i(t)$ and $i \rightarrow j$).

Proof.

We apply the n-choice optimality theorem 1.6.4. for unstructured automata in a static medium, for each link $\sigma_{ij}^{s_\alpha}$, and the result is immediate. Note that the $q_{ij}^{s_\alpha}$ above differs from the usual $q_i^{\alpha,s}$ since here we must differentiate between the "blueprint" $\{\nu_{ij}^{s_\alpha}\}$ and the learning mechanism. //

Remarks 3.8.3.

i) Essentially, we ensure that there is a unique absorbing state so that the evolving automaton is "always at the same potential" in that $\mathcal{V}_{BP} \equiv 1$ for the given "blueprint" (B.P.) if $\sigma_{ij}^{s_\alpha}(0) \in (0,1)$.

ii) The essence of $\bar{\pi}$-cell learning is that $\mathfrak{M}(\Delta_{\alpha\beta}, q_{u_i}^{\alpha,s})$ is unknown so that learning is "blind" just as in our own evolution. If we allow the automaton to have access to $q_{u_i}^{\alpha,s}$ we get a similar rather artificial "forced" result to 3.8.2. for it easy to obtain $q_{u_i}^{\alpha,s} \rightarrow 1 \ \forall \alpha, u_i$ through reinforcing $q_{u_i}^{\alpha,s}$ whenever $u_i(t)$, $E_\alpha(t)$ hold at time t.

iii) We thus see that if we deviate from the $\bar{\pi}$-cell formulation through allowing the automaton to possess fuller knowledge of $\mathfrak{M}$ then the results degenerate almost to remarks rather than leading to a fuller insight. So in representing structural adaptation through an evolving automaton $\mathcal{G}$, we must ensure that we have no access to "global" information of $\mathfrak{M}$.

3.9 Hierarchical Automata.

We have considered static and markovian environments $\mathfrak{M}$. In general $\Delta_{\alpha\beta}$ may consist of sub-environments, so that it would be natural to define a hierarchical automaton to adapt to such an environment. Tsetlin (1965) and Narendra and Viswanathan (1972) both used forms of 2-level systems for periodic media. The 1st level determines the period whilst the 2nd level operates in the selected environment. The $\mathcal{G}^q(\theta)$ considered here is at present of conceptual rather than practical value and indicates how the theory of structural adaptation may develop in the future. This section contains work which is still rather speculative with many properties remaining to be investigated. We first motivate the theory through two simple examples.

Examples 3.9.1.

Here we take the structured automaton and substitute automata themselves as the actions executed in the states of the higher order automaton. Thus instead of a Π-cell being allocated to a set of states, we may replace it by a network of Π-cells. In these examples the automata are fixed and so we do not require a reinforcement stimulus; just a stimulus to determine state switching.

a) $\Delta_{\alpha\beta} = \begin{pmatrix} \Delta^1 & \Delta^\epsilon \\ \Delta^\epsilon & \Delta^2 \end{pmatrix}$ with $\Delta^i = \begin{pmatrix} 1-\delta_i-\epsilon, & \delta_i \\ \delta_i & , 1-\delta_i-\epsilon \end{pmatrix}$ $\Delta^\epsilon = \begin{pmatrix} \epsilon, & 0 \\ 0 & \epsilon \end{pmatrix}$

and where $\epsilon \ll \delta_i \quad i = 1, 2.$

and $\delta_1 \ll \frac{1}{2} \quad \delta_2 > \frac{1}{2}$

We denote the environment states as:- E_{11}, E_{12}, E_{21}, E_{22}.

Then for Δ^1 we use ${}_2L_{\underset{\sim}{r}}(1)$.

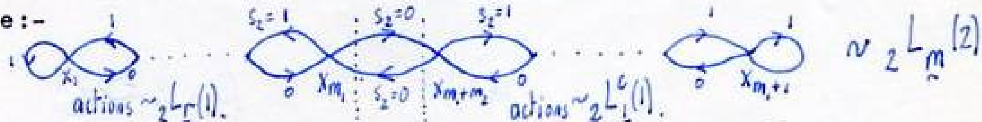
" " Δ^2 " " ${}_2L^c_{\underset{\sim}{1}}(1)$.

and for Δ use ${}_2L_{\underset{\sim}{m}}(2)$ with $\underset{\sim}{m}$ chosen according to ϵ.

We define stimulus $s(t) = (s_1(t), s_2(t))$, where $s_j(t)$ is applied to the j^{th}-level in the hierarchy.

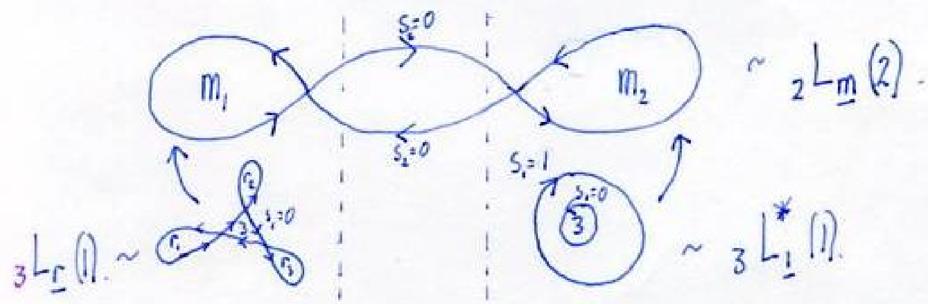
For level 1 automata in E_{ab} we put $q^{b,1}_{i=b} = q = q^{b,0}_{j\neq b}$ else $q_i^{\alpha s} = p = 1-q$. And similarly for level 2 we put $q^{a1}_{i=a} = q = q^{a0}_{j\neq a}$ else $q_i^{\alpha s} = p$.

Thus we have:-



The actions on ${}_2L_m(2)$ are replaced by automata ${}_2L_r(1)$ and ${}_2L_1^c(1)$ of the first level.

b) We can form a hierarchical automaton to operate in a cyclic 3-medium in a similar manner, where the δ parameter switches between say $\delta = \frac{1}{10}$ and $\delta = \frac{2}{3}$. Thus we have:-



//

The technique of 3.9.1. allows us to adapt to environments which split naturally into sub-environments. If we have $\mathfrak{M}(\Delta_{\alpha\beta}, q_i^{\alpha,s})$ and the parameters change, then we switch between the covering automata of $A_1(\mathfrak{M})$ as in 3.9.1., where the coverings were investigated in 3.4.12.

Definitions 3.9.2.

i) Let $\mathcal{G}^n(\mathbb{O})$ be an n-level hierarchical automaton of actions, so that the r^{th} level has $\mathcal{G}_i^{r-1}(\mathbb{O})$ as actions.

ii) Let $\sigma_{ij}^{s_2,r}(r,t)$ be the transition function for the r^{th} level at time t, with reinforcement under U.L. rules.

iii) Let $\Delta^{m}_{\alpha\beta}(r)$ =Pr($E_\alpha \to E_\beta$ in r^{th} environmental level when n^{th} level is in m_1, and r^{th} level is in m_{n-r+1}.)

iv) Let ${}^{\alpha}q_i^{m,s_{j,r}} = \Pr$(stimulus $s_{j,r}$ in level r on using action i in E with $m_r = \alpha$, and (r+s)th environment level in $m_{n-r+s+1}$.)

v) We have matrix stimulus $s_{\cdot j} = (s_{1j}, s_{2j})$ with $s_{ij} = 1$ representing reward, as usual.

Here s_{1j} is used for r^{th} level reinforcement and s_{2j} is used for r^{th} level transitions, each being independent with the same stimulus probability. Thus we have a vector simulus for each level.

We have defined a stimulus for each level rather than a single global stimulus since the latter would require large memory in many cases and would fail to operate if we required ${}_nL_1^c$, say, as actions.

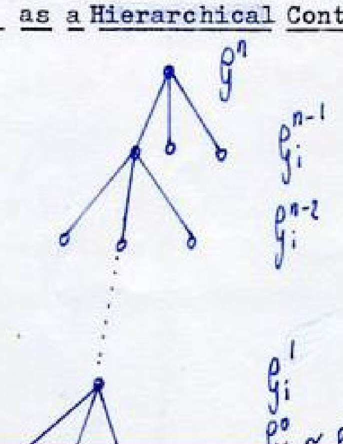
Using s_{ij} we get in phase with $\mathfrak{M}$ by successive levels, rather than requiring the whole automaton to be in phase before we benefit from the increased rewards, as in the case of global $s=(s_1, s_2)$.

vi) We define Γ_α^r as before so that $\mathcal{G}_\alpha^{r-1}$ operates at time t if and only if $x_i^r(t) \in \Gamma_\alpha^r$ with x_i^r an r-level state. Thus Γ_i^r is the set of states in the r^{th} level which use action i. If each level is held in Π-cells then Γ_k^r denotes the r-level $\Theta_k(r)$.

If we change $x_i^r(t) \longrightarrow x_j^r(t+1)$ with $i \in \Gamma_\alpha^r$ and $j \in \Gamma_{\beta\neq\alpha}^r$ then the automaton action $\mathcal{G}_\beta^{r-1}$ will initially require time to "acclimatize" after its period of inactivity.

We consider each level r, with its associated memory held in each Γ_α^r, reaching out to "tap" different environmental facets.

a) $\mathcal{G}^n$ as a Hierarchical Controller.



b) Fully Extended $\mathcal{G}^n$.

action switch n

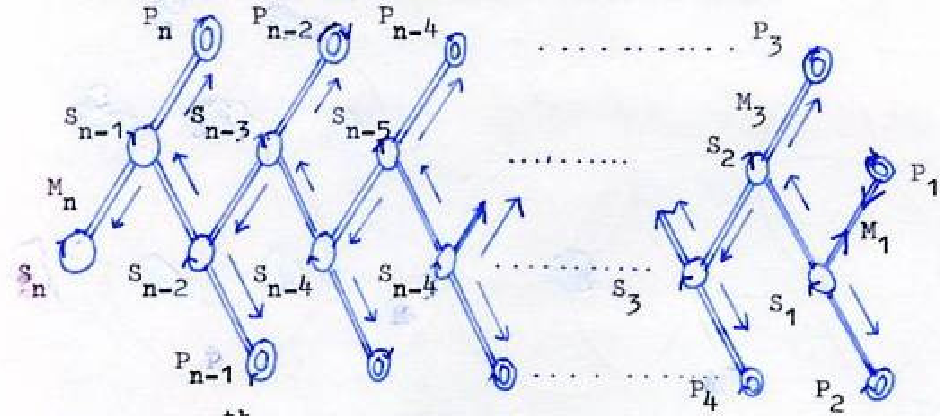
memory n

payoff n and action switch n-1.

levels r s.t. 1<r<n.

payoff level 1.

c) Representation of $\mathcal{G}^n$ Hierarchy in General State.



$$\left.\begin{array}{l} \text{where } S_i = i^{th} \text{ action switch.} \\ M_i = i^{th} \text{ memory.} \\ P_i = i^{th} \text{ payoff.} \end{array}\right\} \text{ and } \mathcal{G}^r = \bigcup_{i=1}^{i=r} (S_i, M_i, P_i) \text{ in terms of these "elements".}$$

We have an abstract "expanding tongs" exploring the environment, which can be likened to a robot under hierarchical control, as considered by Albus and Evans (1976), in a practical context.

On the memory of each $\mathcal{G}_j^r$ we have automaton actions $\mathcal{G}_i^{r-1}$.

Remarks 3.9.3.

i) The environment transition matrix $\Delta^m_{\alpha\beta}$ as defined in 3.9.2. iii) differs from that used in the motivating examples 3.9.1.

In our new notation we write:-

$$\Delta^1_{\alpha\beta} = \begin{pmatrix} 1-\delta_1, & \delta_1 \\ \delta_1, & 1-\delta_1 \end{pmatrix} \qquad \Delta^2_{\alpha\beta} = \begin{pmatrix} 1-\delta_2, & \delta_2 \\ \delta_2, & 1-\delta_2 \end{pmatrix}$$

$$\text{and } \Delta_{\alpha\beta} = \begin{pmatrix} 1-\epsilon, & \epsilon \\ \epsilon, & 1-\epsilon \end{pmatrix}$$

with ϵ small:- $0 < \epsilon \ll \delta_i$, $i = 1, 2$.

ii) When we hold the actions at each level in Π-cells we can write:-

$$\mathcal{G}^n(\circledast) = \underbrace{\mathcal{G}(\mathcal{G}(\mathcal{G} \;\ldots\ldots\; \mathcal{G}(\circledast))\ldots)}_{n\text{-times.}}$$

So for $\mathcal{G}^n$ we have Π-cells of $\mathcal{G}_i^{n-1}$. This is called Π-cellising the automaton actions at each level.

Then we obtain the results for hierarchies of Π-cells by the theory previously developed for $\mathcal{G}(\circledast)$.

We define $\Pi_i^k(r,t)$=Prob(use action $\mathcal{G}_i^{r-1}$ at r^{th} level on using Π-cell $\otimes_k(r)$, at time t.)

With the stimulus defined for each level we can immediately modify the action partitioning conjectures of 3.7. to the case of $\mathcal{G}^n(\otimes)$. Thus we can Π-cellise the actions at each level to obtain the full $\mathcal{G}^n(\otimes)$, or define a fixed action for each level state, to give $\mathcal{G}^n(\mathcal{O})$. These are then the natural generalisations of Π-cell networks and structured automata, respectively, which are 1-level hierarchies $\mathcal{G}^1$. Clearly we could also construct mixed hierarchies so that we Π-cellise some levels and leave the remaining levels as fixed actions.

iii) The actions at the r^{th} level are denoted $u_i(r)$ and are held in the $\Pi_i^k(r,t)$ if we Π-cellise. The $u_i(r)$ are $\mathcal{G}_i^{r-1}$ automata which receive stimuli s_{ir} with probability ${}^{\alpha}q_{u_i}^{m\ s_{ir}}$

We put $\bar{R} = \sum_{r,\alpha,i} \gamma_i^{\alpha}\, q_{u_i(r)}^{\alpha, s_{i,r}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{r=1}^{n} s_{i,r}(t)$. where γ_i^{α} = eqm dsn for $\mathcal{G}^n$ over $E_{\underset{\sim}{\alpha}}$ and states $\underset{\sim}{i}$.

So the average reward is obtained by summing over each level r.

iv) For hierarchical automata, we reinforce transition links $\sigma_{ij}^{s_{i,r}}$ only if we use the same action automaton $\mathcal{G}^r$ on ≥ 2 successive trials.

v) We note that biological systems themselves have a hierarchically stable organisation, so that each level is stable within the appropriate environmental level.

We can obtain a generalised SOSA limit theorem, with the stable limiting family defined as A_0^n for an n-level hierarchy $\mathcal{G}^n(\mathcal{O})$.

Theorem 3.9.4.

If $A_0^n(M) \neq \phi$ then under R_0, $\sigma_{ij}^{s_{i,r}} \to \nu_{ij}^{s_{i,r}}$ for some hierarchy that is SOSA at each level, for $\mathcal{G}^n(\mathcal{O})$.

Proof.

As in 3.7.1. we can write the SOSA property as $\max_1 \mathcal{E} R$ at each level with respect to the equilibrium distribution over the complete hierarchy.

We essentially need only apply 3.4.9. to <u>each</u> level r.

Thus $\sigma_{ij}^{s_2,r} \to 1$ only if $\max_1 \xi R$ w.r.t. $\gamma_i^\alpha(r)$ where:-

$_\beta\gamma_i^\alpha(r)$ = Pr (in E_β at r^{th} level and $m_{n-r+1+s.}$ at r+s level , and in automaton state i_{r+s} at level r+s, with $0 \le s \le n-r$.)

So the full equilibrium distribution is $\gamma_i^\alpha(0)$,yet for level r we are only dependent on environment and automaton levels $s \geq r$ above us.

Now, as in 3.4.9. , we take +ve recurrent state x_i^r at level r and consider $\sigma_{ij}^{s_2,r}$ in an arbitrarily small neighbourhood of the absorbing boundary. Then we fix $\{\sigma_{ij}^{s_2,r}\}$ and find $\overline{\Delta\sigma}_{ij}^{s_2,r}$ in equilibrium, on considering the hierarchical automaton as a markov process. Note that each level r must have at least one +ve recurrent state, since automaton has finite state space.

Now we have $\overline{\Delta\sigma}_{ij}^{s_2,r} > 0$ with $j \in \Gamma_\gamma^r$ if and only if $u_\gamma(r)$ gives $\max_1 \xi R$ w.r.t. the equilibrium distribution $\gamma_i^\alpha(r)$.

Then under $\mathcal{R}_0$, we use 1.12.9. , with $\underline{e}$ being replaced by $\gamma_i^\alpha(r)$, and hence the limiting structure at each level is SOSA as required. //

We can prove a similar theorem to 3.7.1. for $\mathcal{G}^n(\otimes)$ which will give us the limiting family $\mathcal{B}_0^n$ for an evolving n-level hierarchy.

<u>Theorem</u> 3.9.5.

Let i_r be a +ve recurrent state as $t \to \infty$ at level r, for each r. Then under $\mathcal{R}_0$ and in $\mathcal{M}$:-

i) $\sigma_{ij}^{s_2,r} \to 1$ with $i \in \Gamma_k^r$, for $\mathcal{G}^n(\otimes)$.

ii) $\pi_m^k(r) \to 1$

only if $\max_1 \xi R$ w.r.t the equilibrium distribution $\gamma_i^\alpha(r)$ for $\sigma_{ij}^{s_2,r} \to 1$

and $\max_0 \xi R$ " " " " $\xi_{\otimes_k(r)}^\alpha$ " $\pi_m^k(r) \to 1$

where $\xi_{\otimes_k(r)}^\alpha$ =Pr(environment level $s \geq r$ is in m_{n-s+1} | we use $\otimes_k(r)$,in eqm)

We denote $\{\nu_{ij}^{s_2,r}, y_m^k(r)\}$ family as $\mathcal{B}_0^n(\mathcal{M})$, where $\sigma_{ij}^{s_2,r} \to \nu_{ij}^{s_2,r}$ and $\pi_m^k(r) \to y_m^k(r)$.

Proof.

This is just a restatement of 3.7.1 for n-levels, and the same method of proof applies in each level. Thus we extend $\mathcal{B}_0$ to $\mathcal{B}_0^n$ just as 3.9.4. gives us $\mathcal{A}_0 \to \mathcal{A}_0^n$. //

Now at each level we conjecture that the Π-cellised actions partition themselves, so that 3.7.2. $\longrightarrow$ 3.7.4. have natural extensions. Thus on each r-level memory Γ_α^r we conjecture that only one $\mathcal{G}_\alpha^{r-1}$ is allowed, and then that $\mathcal{B}_0^n(\mathfrak{M}) \subset \mathcal{A}_0^n(\mathfrak{M})$, for each $\mathfrak{M}$.

An evolving $\mathcal{G}^n(\circledast)$ will first adapt to the higher level environments before proceeding to "explore" the more minor environmental facets of the lower levels with the lower $\mathcal{G}^r(\circledast)$ hierarchies. Thus, intuitively, we can imagine a "wave" of adaptation and action partitioning travelling down the "branches" of the hierarchical controller $\mathcal{G}^n$.

3.10 Games between Structured Automata.

In the Russian Literature there are many papers which have been published on games between fixed deterministic automata, which were first formulated by Tsetlin (1963). Subsequent work has been published by:- Butrimenko (1967), Gersht (1967), Ginzburg and Stefanyuk (1970), Gurvich (1975), Kalinin (1965, 1966), Krinskii(1963,1966), Takeuchi (1974), Tsertsvadze (1970), Tsetlin (1964, 1965, 1974), Vaisbord (1968), Varshavskii (1972) and Volkonskii (1965). Some related papers, including applications, are also to be found in the bibliography at the end of the thesis.

We define ${}_i g_{\beta}^{k\prime}$ as before, in 2.5.1., and then we obtain 3.10.1. quite easily. The game between Π-cell networks is denoted $\prod_{i=1}^{n} \mathcal{G}_i^{1}(\circledast)$.

Theorem 3.10.1.

For n , structured automata, $\mathcal{G}_k^1(0)$ $1 \le k \le n$, evolving under $\mathcal{R}_0$ with game matrix ${}_i g_{\underset{\sim}{\beta}}^{k\prime}$ then:-

a) A pure strategy $\underset{\sim}{\beta}^*$ is stable in the limit ~~iv~~ if and only if it is a Nash Point of ${}_i g_{\underset{\sim}{\beta}}^{k\prime}$.

b) $\sigma^{s}_{ij}(r) \to 1$ only if the limiting structure has the SOSA property w.r.t $\underset{\sim}{\gamma_r}$, where $\underset{\sim}{\gamma_r}$ is the equilibrium distribution for $\prod_{i=1}^{n} \mathcal{G}'_i(\Theta)$.

<u>Proof</u>.

a) This does <u>not</u> assert uniqueness; only that any Nash Point is stable. We follow 2.5.1. and 3.2.2. in considering $\sigma^{s}_{ij}(r,t)$ near an absorbing barrier of pure strategies $\underset{\sim}{\beta}^{*}$ for each automaton r at time t. Then $\Delta\sigma^{s}_{i,\beta^{*}_{r}}(r) > 0 \ \forall r, i$, in a sufficiently small nbd of the boundary $\{\sigma^{s}_{i,\beta^{*}_{r}}(r) = 1, \forall r, i\}$, iff $\underset{\sim}{\beta}^{*}$ is a Nash strategy. Now we use the boundary learning theory of 1.7.5. under Θ_0 to give convergence to $\underset{\sim}{\beta}^{*}$ only if it is a Nash Point, so that each $\mathcal{G}'_r$ just contains $\Gamma^{r}_{\beta^{*}_{r}}$, in the limiting structure.

It is unresolved whether other limiting stable structures exist if ${}_{i}g^{k}_{\underset{\sim}{\beta}}{}'$ has Nash Points. For a singleton Π-cell this is solved in 2.5.1.

b) This does not give the existence of limiting deterministic structures for it is possible that such limits do not always exist, as in the periodic trajectories of 2.3.

The result is obtained in the same way as the theorem 3.4.9. , by writing out $\overline{\Delta\sigma^{s}_{ij}(r)}$, where the average is taken w.r.t. the equilibrium distribution $\underset{\sim}{\gamma_r}$. Then apply 1.12.9. , with $\underset{\sim}{e}$ replaced by $\underset{\sim}{\gamma_r}$ to obtain the SOSA convergence constraint. //

<u>Remarks</u> 3.10.2.

i) Again, we could extend 3.10.1. to $\prod_{r=1}^{n} \mathcal{G}'_r(\Theta)$ on defining $\mathcal{G}^{\underset{\sim}{\beta}}_{\Theta_k(r)}$ for the product gaming environment, with strategy vector $\underset{\sim}{\beta}$.

ii) We would like the limiting structures to resemble the automata of Krinskii (1963), which have a deterministic "circular" behaviour, but it is unclear whether they have the SOSA property in a gaming environment. Thus although the linear automata are ideally suited to learning in markovian media, we need cyclic behaviour for automata games, with the digraph resembling the Φ_{*} =const trajectories of section 2.3.

iii) A structural analysis of $\prod_{i=1}^{n} \mathcal{G}'_i(\Theta)$ still remains to be done, but it seems most unlikely that they will achieve worse results than the gaming singleton π-cells, just as we proved 3.2.2. for games against "nature".

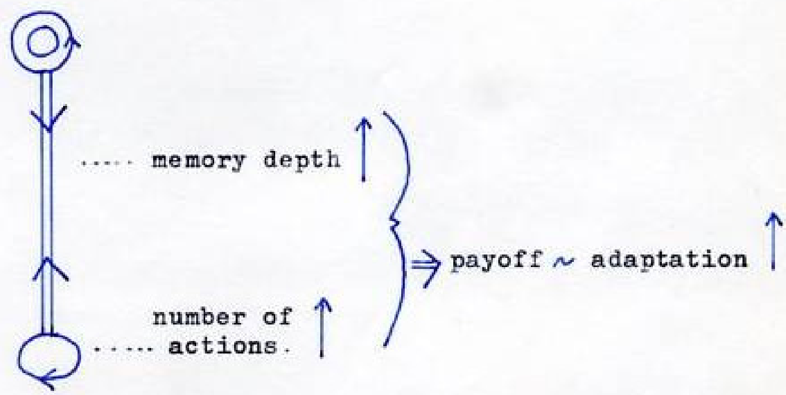
3.11. Concluding Remarks.

3.11.1. There are essentially two means by which automata can increase their average payoff in the formulation of this chapter.

a) Memory extension. ("Tree" ↑)

b) Action switch enlargement. ("Kernel" ↑)

We represent this diagrammatically as:-



We saw in 3.8. that efforts to improve on suboptimality using controllers, were rather artificial. It would be preferable to postulate community interaction so that automata learn by their mutual errors and successes.

For a) we need to "fluidise the structure" and increase the adaptation through observing other automata with larger memories. In addition, we could allow automata to possess individual curiosity so that some may act as pioneer automata in trying out new structures. A preliminary model for community behaviour is presented as 3.11.2.

For b) new actions may arise which are expedient in some environment and which automata can now incorporate in their structure. The cities of Paris and Rome have new "actions" in their "kernel of operation"; La Défense and the E.U.R. respectively. Bacon (1974) adopts such an evolutionary approach in his qualitative study of cities.

Model 3.11.2. ("The Sheep Effect".)

Here we are not concerned explicitly with the automaton structure but just in their relative expediency in an environment. In this very simple model we assume perfect mutual observation between automata, rather than the usual observation with uncertainty as in n-cell theory.

i) We define states $x_i = 1,2, \ldots m$ with state 1 "best" and then the utility monotonically decreases to give state m "worst".

ii) We define automata $a_1, \ldots\ldots a_n$.

iii) If an automata observes another automata doing "better" than itself, then it jumps with probability 1/m to some x_1.

iv) Let $A_{j_1 \ldots j_n}(s) = Pr($ all automata are eventually absorbed in state s where initially a_i is in state j_i.)

v) Clearly $A_{\underline{j}}(s)$ is only a function of $\min_i j_i$ and the number which attain this.

So define $A_{jr} = Pr($ all automata eventually absorbed in x=1, given r start in $x=j > 1$.)

Thus
$$A_{jr} = A_{jr}\left(\frac{m-j}{m}\right)^{n-r} + \sum_{s=1}^{n-r} A_{j,r+s}\binom{n-r}{s}\left(\frac{1}{m}\right)^{s}\left(\frac{m-j}{m}\right)^{n-r-s}$$
$$+ \sum_{k<j}\sum_{s=1}^{n-r} A_{ks}\binom{n-r}{s}\left(\frac{1}{m}\right)^{s}\left(\frac{m-k}{m}\right)^{n-r-s}.$$

or
$$A_{jr}\left(m^{n-r} - (m-j)^{n-r}\right) = \sum_{s=1}^{n-r} A_{j,r+s}\binom{n-r}{s}(m-j)^{n-r-s}.$$
$$+ \sum_{k<j}\sum_{s=1}^{n-r} A_{ks}\binom{n-r}{s}(m-k)^{n-r-s}.$$

And if $A_{1n} = 1$, $A_{jn} = 0$, $j \neq 1$, then:-

$A_{jr} \sim 1 - (1/j)^{n-r}$ for $j > 1$ and n large. So that the probability of non-optimal absorption becomes geometrically small as n increases.

vi) Thus we have a bootstrapping effect in that the "pioneer" automata "pull up" the trailing "sheep" automata, yet which themselves may change rôles and act as the new pioneers. Since the automata are only guided by each other, there is always a finite probability that they will <u>all</u> become absorbed together in the <u>same</u> sub-optimal state. //

3.11.3. As motivation for extending the π-cell theory to hierarchies of automata $\mathcal{G}^n(\circledast)$ in 3.9. , we provide the following brief intuitive outline.

Initially we suppose there are singleton π-cells which proceed to cluster for their mutual benefit and subsequently differentiate in their rôles as in 3.7. Now to form $\mathcal{G}^{n+1}$ we require a mechanism of collective behaviour between the $\mathcal{G}^r_i$ which eventually gives rise to a critical point and the gelation of structural links, followed by rôle differentiation. (Whittle 1971). In this way automata can attain adaptation in higher environmental levels of $\mathfrak{M}$, within a single entity $\mathcal{G}^n$. The automaton $\mathcal{G}^n$ acts as a hierarchical controller with autonomous behaviour within each lower level, and which we could now embed in the appropriate ω -likelihood simplex, as in 3.6., for $\mathcal{G}^1$.

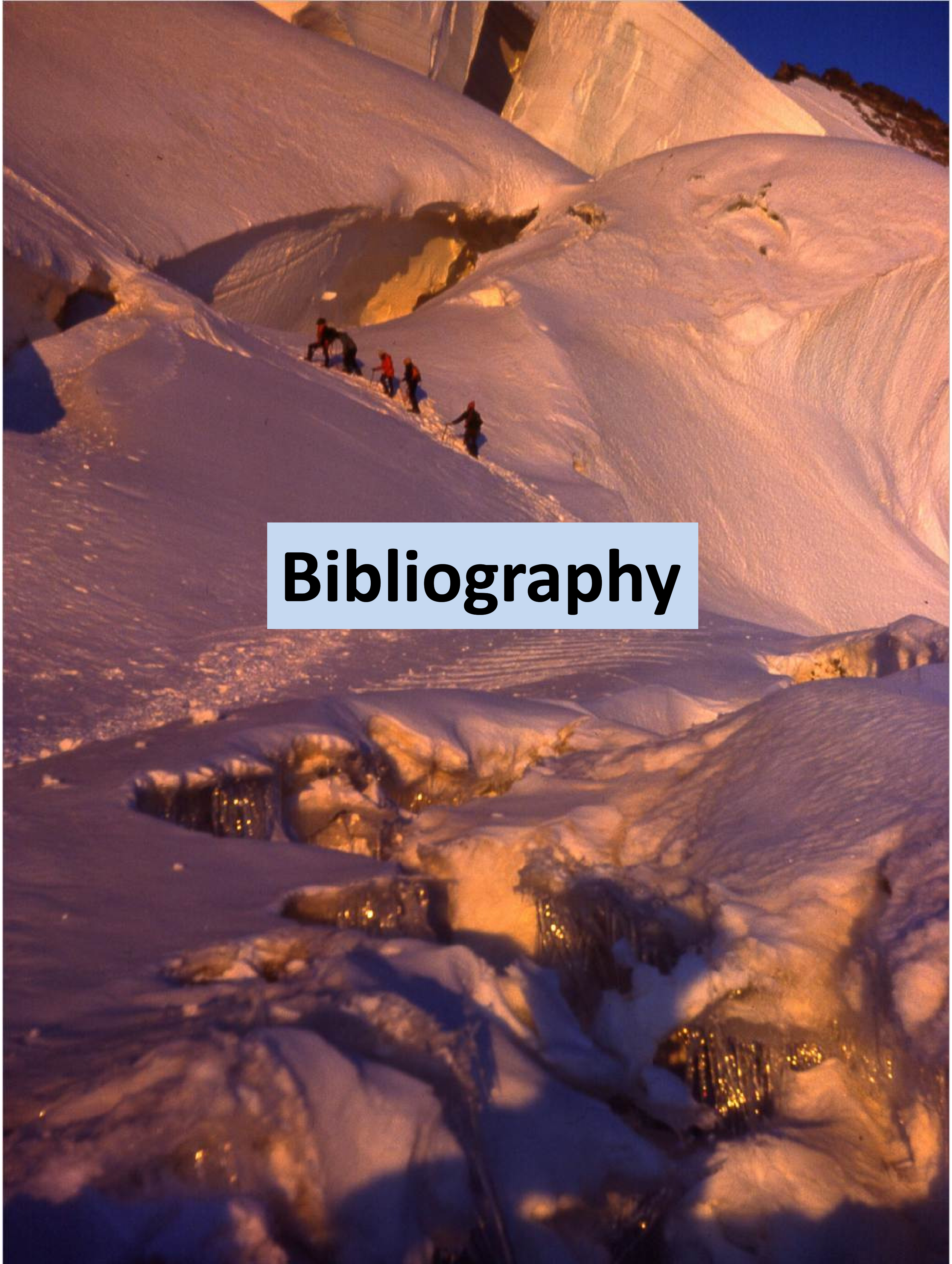
Such a process of clustering and subsequent rôle differentiation appears frequently in nature. In particular, we may proceed up the evolutionary tree from the single-celled amoeba, eventually obtaining the multicellular mammals and the cultural society of man. The actual mechanism of reinforcement here is natural selection over many generations with the genetic encoding of this information giving embryonic cell differentiation and subsequent learning over the lifetime of each individual.

3.11.4. At present, most models in psychological learning theory are based on the unstructured automaton, yet semantic ideas express the brain as a <u>structure</u> evolving through the recption of environmental stimuli, as in the simple qualitative models of De Bono (1971).

So as a simple conceptual model, we have investigated the π-cell as a basis for structural learning, with its family of learning functions.

3.11.5. Probability expresses our environmental uncertainty so that through learning, our ideas become rigid, and we act according to that structure.

Research requires us to "fluidise" this structure to enable us to "extend the trees" and "enlarge the kernels", in our "blind" attempts at environmental understanding and adaptation.



Bibliography

4 Bibliography.

.... even as the sand that is upon the sea shore

in multitude.

Joshua 11 v 4.

4 Bibliography.

A comprehensive bibliography is given here including some papers not actually cited in the text, although they all helped in the development of the underlying theme of the thesis.

Much of this literature was gathered before the survey paper of Narendra and Thathachar (1974), yet the field covered is remarkably similar. This includes the work of mathematical psychologists, American and Japanese control engineers and Russian automaton theorists.

Prior to 1973, the literature was virtually partitioned into the three categories above, but it is now clear that all approaches are merging to form a common framework of learning automata and structural adaptation. More abstractly, we could view the field within Thom's (1975) theory of morphogenesis.

Abbreviations:-

a) A.R.C. – Automation and Remote Control.
(Translation of Avtomatika i Telemekhanika.)

b) P.I.T. – Problems of Information Transmission.
(Translation of Problemy Peredachi Informatsii.)

c) S.S.C. – I.E.E.E. Transactions on Systems, Science and Cybernetics.
(1965-1970)

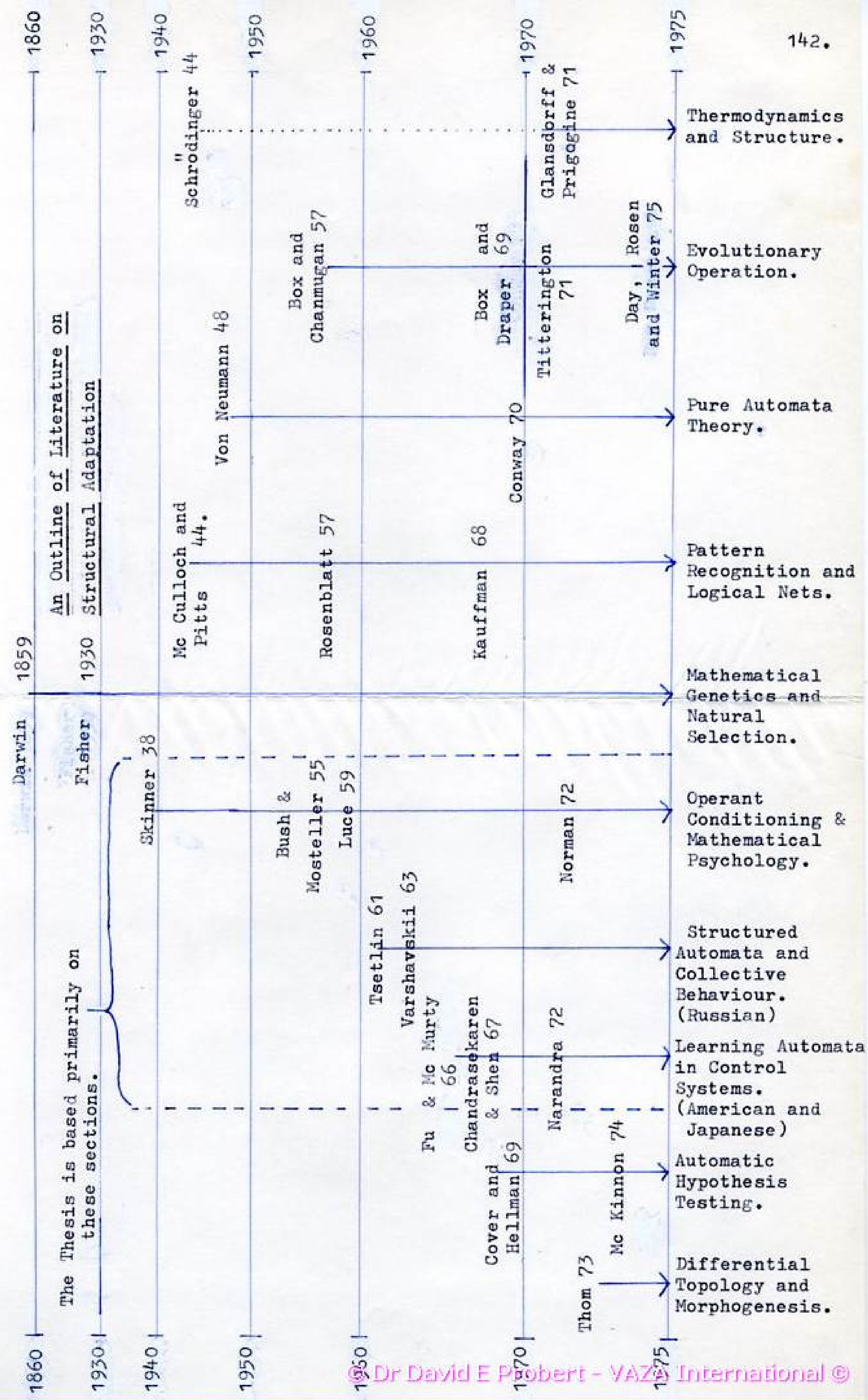
d) S.M.C. – I.E.E.E. Transactions on Systems, Man and Cybernetics.
(1971-)

e) A.C. – I.E.E.E. Transactions on Automatic Control.

f) I.T. – I.E.E.E. Transactions on Information Theory.

g) J. of Math Psych – Journal of Mathematical Psychology.

h) J. of Theor Biol – Journal of Theoretical Biology.


An Outline of Literature on Structural Adaptation
The Thesis is based primarily on these sections.
1860
1930
1940
1950
1960
1970
1975
Darwin 1859
Fisher 1930
Skinner 38
Bush & Mosteller 55
Luce 59
Norman 72
Mc Culloch and Pitts 44.
Rosenblatt 57
Kauffman 68
Von Neumann 48
Conway 70
Box and Chanmugan 57
Box and Draper 69
Titterington 71
Day, Rosen and Winter 75
Schrödinger 44
Glansdorff & Prigogine 71
Tsetlin 61
Varshavskii 63
Fu & Mc Murty 66
Chandrasekaren & Shen 67
Narandra 72
Cover and Hellman 69
Mc Kinnon 74
Thom 73
Thermodynamics and Structure.
Evolutionary Operation.
Pure Automata Theory.
Pattern Recognition and Logical Nets.
Mathematical Genetics and Natural Selection.
Operant Conditioning & Mathematical Psychology.
Structured Automata and Collective Behaviour. (Russian)
Learning Automata in Control Systems. (American and Japanese)
Automatic Hypothesis Testing.
Differential Topology and Morphogenesis.

1) Accardi L. (1974). Description of Probabilistic Evolution Dependent on the Past. A.R.C. 4 50-61.

2) Asai K and Kilajima S (1971). A Method of Optimising Control of Multimodal Systems using Fuzzy Automata. Information Sciences 3 343-353.

3) Albus J and Evans J.M. (1976). Robot Systems. Scientific American 234 2.

4) Aleksander I (1971) Electronics for Intelligent Machines. New Scientist 11th March.

5) Aiserman M.A et Al (1971). Logic, Automata and Algorithms. Academic Press.

6) Aiserman M.A. (1975). A Man and a Collective as Elements of a Control System. A.R.C. 36 No 5 Pt 1.

7) Agasandyan G.A. (1969) Some Aspects of the Theory of Automata with Variable Structure. P.I.T. 5 No 1 71-78.

8) Arbib M.A. (1969) Theories of Abstract Automata. Prentice-Hall Inc.

9) Atkinson R.C. (Ed 1964) Studies in Mathematical Psychology. Stanford University Press.

10) Atkinson R.C. et Al (1974). Contemporary Developments in Mathematical Psychology. Freeman San Francisco.

11) Bacon E.N. (1974). The Design of Cities. Thames and Hudson.

12) Barbour A. (1973). The Principle of the Diffusion of Arbitrary Constants. J. of Applied Probability. 9 519-541.

13) Bellman R (1956). A Problem in the Sequential Design of Experiments. Sankhya. 16 221-229.

14) Bellman R (1975). Communication and Mind. Biosciences. 27 p347-

15) De Bono E (1971). The Mechanism of Mind. Penguin Books.

16) Bernal J.D. (1967). The Origin of Life. W.Clowes and Son.

17) Borovikov V.A. and Bryzgalov V.I. Simplest Symmetrical Game of Many Automata. A.R.C. 26 No 4.

18) Breiman L (1968). Probability. Addison Wesley.

19) Bush R.R. and Mosteller F. (1955) Stochastic Models for Learning. Wiley.

20) Butrimenko A.V. (1967). Games of Automata Possessing Different Activities. P.I.T. 3 No 4 81-88.

21) Buyakas V.I. (1974) Problem of Controlling Collective Behaviour. A.R.C. 35

22) Box G. and Jenkins G (1962). Some Statistical Aspects of Adaptive Optimisation and Control. J.R.S.S. B. 24 297-343.

23) Box G and Draper N. (1969). Evolutionary Operation. Wiley.

24) Cairns-Smith (1971). The Life Puzzle. Oliver and Boyd.

25) Calvin (1969) Chemical Evolution. Oxford University Press.

26) Chaikovskii Yu.V. (1970) A Group of Gain Comparing Automata. P.I.T. 6 No 4 56-65.

27) Chandrasekaren B and Shen D.W.C. (1967) Adaptation of Stochastic Automata in Nonstationary Environments. Proc of National Electronics Conference.

28) Chandrasekaren B and Shen D.W.C. (1968) On Expediency and Convergence in Variable Structure Automata. S.S.C. 4 No 1.

29) Chandrasekaren B and Shen D.W.C. (1969) Stochastic Automata Games. S.S.C. 5 No 2.

30) Chandrasekaren B. (1970). Finite-Memory Testing. A Critique. I.T. 16.

31) Chandrasekaren B. et Al. (1974). Artificial Intelligence. S.M.C. 4 p88-103.

32) Chien Y.T. and Fu K.S. (1967a) Learning in Nonstationary Environments using Dynamic Stochastic Approximation. Proc 5th Annual Allerton Conference on Circuit and System Theory.

33) Chien Y.T. and Fu K.S. (1967b) On Bayesian Learning and Stochastic Approximation S.S.C. 3 28-38.

34) Chien Y.T. and Fu K.S. (1969) Stochastic Learning of Time-Varying Parameters in Random Environments. S.S.C. 5 No 3 237-246.

35) Codd E.F. (1968) Cellular Automata. Academic Press.

36) Coxeter et Al (1953) Uniform Polyhedra. Phil Trans A. 401-450.

37) Day R and Groves T. Eds (1975) Adaptive Economic Models. Academic Press.

38) Dodson M.M. (1976) Darwin's Law of Natural Selection and Thom's Theory of Catasrophes. Math Biosciences. 28

39) Doob (1953). Stochastic Processes. Wiley.

40) Dubins L.E. and Savage L.J. (1965). How to Gamble if you Must. Mc Graw Hill.

41) Estes W. (1970).Learning Theory and Mental Development. Academic Press.

42) Feichtinger G. (1970). Lernprozesse in Stochastischen Automaten. Springer-Verlag.

43) Fisher R.A. (1930). The Genetical Theory of Natural Selection. Oxford.

44) Fogel L. et Al (1966) Artificial Intelligence through Simulated Evolution. Wiley.

45) Freedman D. (1973). Another Note on the Borel-Cantelli Lemma and Strong Law. Annals of Probability 1 No 6 910-925.

46) Fu K.S. and Wee W.G. (1967). A Formulation of Fuzzy Automata and its Application to Learning Systems. Proc 5th Annual Allerton Conference on Circuit and Systems Theory.

47) Fu K.S. and Li T.J. (1969) Formulation of Learning Automata and Automata Games. Information Science 1 237-256.

48) Fu K.S. (1970). Learning Control Systems : Review and Outlook. A.C. 15 2.

49) Gardener M. (1970) The Ambedextrous Universe. Pelican Books.

50) Gardener M. (1971). Mathematical Games-"Life".Sci American 224 2 112-117.

51) Gersht A.M. (1967). Games of Continuous Automata. P.I.T. 3 No1 50-56.

52) Ginzburg S.L. and Stefanyuk V.L. (1970). n-Automata Games with Single Nash Point. A.R.C. 31 Pt 2 1264-1272.

53) Glorioso R.M. and Grueneich G.R. (1971) A Training Algorithm for Systems described by Stochastic Transition Matrices. S.M.C. 1 86-87.

54) Glushkov V.M. (1966). Introduction to Cybernetics. Academic Press.

55) Goel N.S. et Al (1971). Non-Linear Models of Interacting Populations.

56) Golovchenko V.B. (1974). Self-Organisation of a Collective of Probabilistic Automata with the Two Simplest Motives of Behaviour. A.R.C. 35 No 4 151-156.

57) Gurvich E.T. (1975). Method for Asymtotic Investigation of Games between Automata. A.R.C. 36 No 2 80-94.

58) Harary F. et Al (1965). Structural Models. Wiley.

59) Hellman M.E. and Cover T.M. (1970a).The Two-Armed Bandit Problem with Time Invariant Finite Memory. I.T. 16 No 2.

60) Glansdorff and Prigogine. (1971). Thermodynamic Theory of Structure and Fluctuations. Wiley Interscience.

61) Hellman M.E. and Cover T.M. (1970b) Learning with Finite Memory. Annals of Math Stat 41 No 3 765-782.

61) " " " (1970c) Comments on Automata in Random Media. P.I.T. 6 No 2 21-30.

62) " " " (1971) On Memory Saved by Randomisation. Annals of Math Stat 42 1075-1078.

63) Hellman M.E. (1972) The Effects of Randomisation on Finite Memory Decision Schemes. I.T. 18 No 4.

64) Hirschler D. and Cover T (1975) A Finite Memory Test of the Irrationality of the Parameter of a Coin. Annals of Statistics 13 No 4.

65) Holman E.W. (1969) Asymtotic Properties of Monotonic Learning Models. J. of Math Psych. 6 456-469.

66) Iosifescu M. and Theodorescu (1969) Random Processes. Springer-Verlag.

67) Jarvis R.A. (1975) Adaptive Global Search by the Process of Competitive Evolution. S.M.C. 5 297-311.

68) Jones P.W. (1975) Two-Armed Bandit. Miscellanea, Biometrika 62 No 2

69) Kalinin D.I. and Epshtein I.M. (1965) A Note on an Automaton Game with Partner using Correlated Mixed Strategy. A.R.C. 26 No 11.

70) Kalinin D.I. and Rotvain I.M. (1966) Some Asymtotic Estimates for Games of Automata in Distribution. 27 No 4 642-644.

71) Kanal L. (1962) A Functional Equation Analysis of Two Learning Models. Psychometrika. 27 89-109.

72) Kandelaki N.P. and Tsertsvadze G.N. (1966) Behaviour of Certain Classes of Stochastic Automata in Random Media. A.R.C. 27 No 6.

73) Karlin S. (1959) Matrix Games, Programming and Mathematical Economics. Vol 1. Addison Wesley.

74) Kauffman S.A. (1969). Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. J. of Theor Biol. 22 437-467.

75) Keilson J and Wishart D. (1964). A Central Limit Theorem for Processes defined on a Finite Markov Chain. Proc Cam Phil Soc 60 p547-567.

76) Keilson J and Wishart D. (1965). Boundary Problems for Additive Processes defined on a Finite Markov Chain. Proc Cam Phil Soc 61 173-190.

77) Keilson J and Wishart D. (1967). Addenda to Processes defined on a Finite Markov Chain. Proc Cam Phil Soc. 63 187-193.

78) Kieras D.E. (1976). Finite Automata and S-R Models. J. of Math Psych. 13 127-147.

79) Koganov A.V. (1973). Automata that Distinguish Random Media. P.I.T. 9 No 2 68-80.

80) Krinskii V.I. (1963). An Asymtotically Optimum Automaton. Biofizika 6 No 6 484-487. (In Russian)

81) Krinskii V.I. (1966). Zero-Sum Games for Two Asymtotically Optimal Sequences of Automata. P.I.T. 1 No 2 43-53.

82) Krylov V.Y. (1963) On One Automaton that is asymtotically Optimal in a Random Media. A.R.C. 24 1226-1228.

83) Kushner H.J. (1967) Stochastic Stability and Control. Academic Press.

84) Laksmivarahan S and Thathachar M.L. (1972) Optimal Non-Linear Reinforcement Schemes for Stochastic Automata. Information Sciences 4 121-128.

85) Laksmivarahan S. and Thathachar M.L. (1973). Absolutely Expedient Learning Algorithms for Stochastic Automata. S.M.C. 3

86) Linkin V. (1972) An Adaptive Algorithm for Determining Variations of the Characteristics of an Observed Random Process. P.I.T. 8 40-45.

87) Luce R.D. (1959) Individual Choice Behaviour.

88) Lyubihik L.M. and Poznyak A.S. (1974) Learning Automata in Stochastic Plant Control Problems. A.R.C. 35 95-109.

89) Mc Culloch W and Pitts W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. Bull Math Biophysics. 5 113-133.

90) Mc Kinnon K.I.M. (1974) Optimal Finite-State Discrimination Procedures. Phd Thesis Cambridge.

91) Mc Laren R,W. (1966) A Stochastic Automaton Model for the Synthesis of Learning Systems. S.S.C. 2 No 2.

92) Mc Murty C.J. and Fu K.S. (1966) A Variable Structure Automaton used as a Multimodal Searching Technique. A.C. 11 No 3.

93) Malishevski A.V. and Tenisberg P.D. (1969) One Class of Games connected with Models of Collective Behaviour. A.R.C. 30 1828-1837.

94) Marley A.A.J. (1967) Abstract One-Parameter Families of Commuting Learning Operators. J. of Math Psych. 4 414-429.

95) Mason I.G. (1973) An Optimal Learning Algorithm for S-Model Environments. A.C. 18 493-496.

96) Maynard-Smith J. (1974) The Theory of Games and the Evolution of Animal Conflicts. J. of Theor Biol. 47 p207-

97) Meleshina M.V. (1969) Automaton Model of Client Interaction Organisation in Queuing Systems with Waiting. A.R.C. 30.1 782-783.

98) Mendal J.M. and Fu K.S. (1970) Adaptive, Learning and Pattern Recognition Systems Theory and Applications. Academic Press.

99) Mendal J.M. (1973) Reinforcement Learning Models with their Applications to Control Problems. Joint Automatic Control Conference 1973 p 3-18.

100) Meyer P.A. (1966) Probability and Potentials. Ginn (Blaisdell).

101) Miller H.D. (1962b). Absorption Probabilities for Sums of Random Variables defined on a Finite Markov Chain. Proc Cam Phil Soc. 58 286-316.

102) Miller H.D. (1962a) A Matrix Factorization Problem in the Theory of Random Variables defined on a Finite Markov Chain. Proc Cam Phil Soc 58 268-285.

103) Milutin A.A. (1965) On the Automaton with Optimal Expedient Behaviour in a Random Environment. A.R.C. 26 No 1.

104) Monod J. (1971). Chance and Necessity. A.A. Knopf Inc.

105) Narendra K.S. and Viswanathan (1972) A Two-Level System of Stochastic Automaton for Periodic Random Environments. S.M.C. 2.

106) Narendra K.S. and Thathachar M (1974) Learning Automata- A Survey. S.M.C. 4 No 4

107) Neimark E and Estes W.K. (1967) Stimulus Sampling Theory. Holden Day.

108) Neumann Von J (1948) The General and Logical Theory of Automata. Collected Works Vol 5 Pergamon Press 1963.

109) Nilsson N.J. (1965) Learning Machines. Mc Graw Hill.

110) Norman M.F. and Yellott J.I. (1966) Probability Matching. Psychometrika 31 43-60.

111) Norman M.F. (1968a) Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators. J. of Math Psych. 5 61-101.

112) Norman M.F. (1968b) On the Linear Model with Two Absorbing Barriers. J. of Math Psych. 5 224-241.

113) " " (1970) Limit Theorems for Additive Learning Models. J. of Math Psych 7 1-11.

114) " " (1971) Slow Learning with Small Drift in Two Absorbing Barrier Model. J. of Math Psych. 8 1-21.

115) Norman M.F. (1972) Markov Processes and Learning Models. Academic Press.

116) Norman M.F. (1974) Markovian Learning Processes. SIAM Review 16 No 2.

117) " " (1975) An Ergodic Theorem for Evolution in a Random Environment. J. Applied Probability. 12 661-672.

118) Owen G. (1968) Game Theory. Saunders Company.

119) Page C.U. (1965) Equivalences between Probabilistic and Deterministic Machines. Information and Control. 9 469-520.

120) Paz A. (1971) Introduction to Probabilistic Automata. Academic Press.

121) Pittel B.G. (1965) The Asymtotic Properties of One Form of Gur Game. P.I.T. 1 No 3 76-89.

122) Ponomarev V.A. (1964) On a Finite Automaton Asymtotically Optimal in a Stationary Random Media. Biofizika 9 No 1 104-110.

123) Pyatetskii-Shapiro I.I. and Vasil'ev N.B. (1967). The Time for an Automaton to Adapt to the External Media. A.R.C. 28.

124) Pyatetskii-Shapiro I.I. (1970) Mathematical Problems associated with Morphogenesis. P.I.T. 4 94-107.

125) Rabin M.O. (1963) Probabilistic Automata. Information and Control. 6 No 3 230-245.

126) Richardson D (1976) Self-Replication by Template. Math Biosciences. 28 1-24.

127) Riordon J.S. (1969) An Adaptive Automaton Controller for Discrete Time Markov Processes. IFAC Journal 5 No 6 721-730.

128) Sawaragi Y. and Baba N. (1973) A Consideration on the Learning Behaviour of Variable Structure Stochastic Automata. S.M.C. 3 644-647.

129) Sawaragi Y. and Baba N. (1974) Two ϵ-Optimal Non-Linear Reinforcement Schemes for Stochastic Automata. S.M.C. 4 126-131.

130) Scarfe H. (1974) The Computation of Economic Equilibria. Yale Press.

131) Schmukler Y. (1970) Unidimensional and Bidimensional Gur Games. A.R.C. 31 Pt 2 1634-1642.

132) Schrödinger (1945) What is Life? C.U.P.

133) Shapiro I.J. and Narendra K.S. (1969) Use of Stochastic Automata for Parameter Self-Optimisation with Multimodal Performance Criteria. S.S.C. 5 No 4.

134) Skilling (1975) The Complete Enumeration of Uniform Polyhedra. Phil Trans A 111-137.

135) Skinner B.F. (1938) The Behaviour of Organisms. Appleton. New York.

136) Stefanyuk V.L. (1963). Example of a Problem in Joint Behaviour of Two Automata. A.R.C. 24 No 6 716-719.

137) Stefanyuk V.L. (1971) Description of Games of Two ϵ-Optimal Automata. A.R.C. 32 Pt 1 No 4.

138) Stewart D.J. et Al (1969) Automaton Theory and Learning Systems. Academic Press.

139) Stratonovich R.L. and Dobrovidov A.V. (1964) On the Synthesis of Optimal Automata to Operate in Random Media. A.R.C. 25 1289-1296.

140) Suppes P. and Lamperti J. (1960) Some Asymtotic Properties of Luce's β- Learning Model. Psychometrika 25 233-241.

141) Suppes P. (1969) Stimulus - Response Theory of Finite Automata. J. of Math Psych. 6 327-355.

142) Sushkov B.G. (1973). Symmetric Games for Large Aggregates of Probabilistic Automata. P.I.T. 9 95-99.

143) Takeuchi et Al (1974) Two Automata Games. Information Sciences. 7 81-93.

144) Tenisberg Yu. D. (1969) Some Models of Collective Behaviour in Dynamic Processes of Market Price Formation. A.R.C. 30 2 1140-1149.

145) Thom R. (1975) Structural Stability and Morphogenesis. Benjamin Inc.

146) Trakhtenbrot B.A. and Barzdin Ya. (1973) The Behaviour and Synthesis of Finite Automata. North Holland Press.

147) Tsertsvadze G.N. (1968) On the Asymtotic Proportion of Purposive Automata in a Stationary Environment. A.R.C. 29 Pt 2 No 8.

148) Tsertsvadze G.N. (1970) Rate of Establishment of the Final Distribution in a Game of Many Identical Automata. A.R.C. 31 1 583-586.

149) Tsypkin Yu.Z. (1966) Adaptation, Training and Self-Organisation in Automatic Systems. A.R.C. 27 No 1 23-61.

150) Tou J.T. et Al (1968) Applied Automata Theory. Academic Press.

151) Tsetlin M.L. (1961) On the Behaviour of Finite Automata in a Random Medium. A.R.C. 22 No 10.

152) Tsetlin M.L. and Krylov V.Yu. (1963) On Games of Automata. A.R.C. 24 No 7.

153) Tsetlin M.L. (1963) Finite Automata and Simple Models of Behaviour. Russian Mathematical Surveys. 18 No 4 1-27.

154) Tsetlin M.L. et Al (1964a). One Example of a Game for Many Identical Automata. A.R.C. 25 No 5 608-611.

155) Tsetlin M.L. et Al (1964b). Homogenous Games for Automata and their Computer Simulation. A.R.C. 25 No 11.

156) TsetlinM.L. and Ginzburg S.L. (1965) Some Examples of Simulation of the Collective Behaviour of Automata. P.I.T. 1 No 2 54-62.

157) Tsetlin M.L. et Al (1965). The Behaviour of Automata in Periodic Random Media and the Problem of Synchronisation in the Prescence of Noise. P.I.T. 1 No 1 65-71.

158) Tsetlin M.L. and Stefanyuk V.L. (1967) Power Regulation in a Group of Radio Stations. P.I.T. 3 No 4 49-57.

159) Tsetlin M.L. (1974). Automaton Theory and Modelling of Biological Systems. Academic Press.

160) Titterington (1971). Adaptive Optimisation of Yield. Phd Thesis Cambridge.

161) Tsuji H. et Al. (1973) An Automaton in the Non-Stationary Random Environment. Information Science. 6 123-142.

163) Vaisbord E.M. (1968a). Game of Two Automata with Differing Memory Depths. A.R.C. 29 Pt 1 440-451.

164) " " (1968b). Game of Many Automata with Differing Memory Depths. A.R.C. 29 Pt 2 1938-1943.

165) Varshavskii V.I. and Vorontsova I.P. (1963) On the Behaviour of Stochastic Automata with Variable Structure. A.R.C. 24 353-360.

166) Varshavskii V.I. and Gersht A.M. (1966) Behaviour of Continuous Automata in Stochastic Media. P.I.T. 2 No 3 68-75.

167) Varshavskii V.I. (1968a) Collective Bahaviour and Control Problems. Machine Intelligence 3 No 14.

168) Varshavskii V.I. et Al (1968) Priority Organisation in Queuing Systems using a Model of the Collective Behaviour of Automata. P.I.T. 4 No 1 73-76.

169) Varshavskii V.I. et Al (1968b). Some Variants on the Problem of Synchronisation of Automata. P.I.T. 4 No 3 73-83.

170) Varshavskii V.I. (1969a) Synchronisation of a Collection of Automata with Random Pairwise Interaction. A.R.C. 30 Pt 1 224-228.

171) Varshavskii V.I. (1969b) The Organisation of Interaction in Collectives of Automata. Machine Intelligence. 4 No 16 285-311

172) Varshavskii V.I. et Al (1969c) Use of a Model of Collective Behaviour in the Problem of Resource Allocation. A.R.C. 30 Pt 1 1107-1114.

173) Varshavskii V.I. (1972a) Some Effects in the Collective Behaviour of Automata. Machine Intelligence. 5 No 22.

174) Varshavskii V.I. (1972b) Automata Games and Control Problems. Proceedings of the 5th World IFAC Congress Pt 3 37.4.

175) Vasershtein (1969) Markov Processes over Denumerable Products of Spaces,Describing Large Systems of Automata. P.I.T. 5 No 3 64-72.

176) Vasil'ev N.B. (1969). The Limiting Behaviour of a Random Media. P.I.T. 5 No 4 68-74.

177) Vasil'ev N.B. and Koganov A.V. (1973). A Model of Optimal Behaviour in an Unknown Media. P.I.T. 9 No 4 58-65.

178) Viswanathan R. and Narendra K.S. (1972) A Note on the Linear Reinforcement Scheme for Variable Structure Automata. S.M.C. 2 292-294.

179) Viswanathan R. and Narendra K.S. (1973). Stochastic Automata Models with Applications to Learning Systems. S.M.C. 3

180) Viswanathan R. and Narendra K.S. (1974). Games of Stochastic Automata. S.M.C. 4 131-135.

181) Volkonskii V.A. (1965). Asymtotic Properties of the Behaviour of Elementary Automata in a Game. P.I.T. 1 No 2 36-53.

182) Vorontsova I.P. (1965). Algorithms for Changing Stochastic Automata Transition Probabilities. P.I.T. 1 No 3 122-126.

183) Waddington C.H. (1966). Principles of Development and Differentiation. Macmillan.

184) Watanabe S. (1975). Creative Learning and Propensity Automaton. S.M.C. 5 No 6.

185) Whittle P. (1971). Stochastic Automata and Co-operative Effects. Technical Report No 69. Stanford University.

186) Wiener N. (1948). Cybernetics. Wiley. New York.

187) Witten I.H. (1973). Finite-Time Performance of Some Two-Armed Bandit Controllers. S.M.C. 3 194-197.

188) Witten I.H. (1974). On the Asymtotic Performances of Finite-State Two-Armed Bandit Controllers. S.M.C. 4.

189) Yasii T. and Yajima S. (1970). Two-State, Two-Symbol Probabilistic Automata. Information and Control 16 203-224.

List of Frequently Occurring Symbols. (with brief statement of meaning)

$\circledast_k$	k^{th} π-cell, 11
π_i	action distribution, 11
u_i	action i, 11
E_i	environment i, 11
s=0	penalty stimulus, 11
s=1	reward stimulus, 11
$q_i^{\alpha,s}$	stimulus probability, 11
$\Delta_{\alpha\beta}$	environment transition matrix, 11
$\mathfrak{M}(\Delta_{\alpha\beta}, q_i^{\alpha,s})$	environment, 11
$R(\underline{\pi})$	average reward, 12
$\Delta X(t)$	expected increment in X(t), 12
U.L.	uniform learning, 12
$\theta_{ij}(\underline{\pi})$	n-action learning function, 13
s/mg	semi-martingale, 15
$\gamma(\underline{\pi})$	absorption function, 19
$U_{\theta_{ij}(\underline{\pi})}$	learning operator for $\theta_{ij}(\underline{\pi})$ rules, 18
$\mathcal{R}_0$	optimal U.L. family, 23
$i_1(\pi_i)$	probability take same action i for all time, 28
#alts	number of response alternations, 27
$\chi(\underline{\pi})$	expected #alts, 30
$\mathcal{R}_\epsilon$	ϵ-optimal U.L. family, 32
Pr(aBb|c)	skeleton absorption function, 56
g_{ij}	stochastic game matrix, 72
π_i^k	action distribution for $\circledast_k$, 72
e_i	equilibrium vector for $\Delta_{\alpha\beta}$, 62
$V^k(\underline{\pi}^i)$	average reward received by $\circledast_k$, 76
λ_j^i	Von Neumann saddle strategy, 76
Φ^*:const	approximate trajectories for mixed strategy saddles, 77
${}^i g_\ell^k$	game matrix for $\circledast_k$ in n-automata game, 85
$\underset{\sim}{s}(t)$	vector stimulus, 89
$\sigma_{ij}^s(t)$	structure transition matrix, 90
Γ_k	set of states which use $\circledast_k$, 90
x_i	state i, 90
$\mathcal{G}^1(\circledast)$	network of π-cells, 91
$\odot$	π_1-cell or pure action u, 92
ξ_i^α	environmental equilibrium probability when in x_i, 95
χ_i^α	joint probability of E_α and x_i, 95
R_{jk}^α	reward comparison matrix for E_α, 98
$\overline{\Delta\sigma_{ij}}$	equilibrium increment, 99
$\omega_\alpha(t)$	probability in E_α at time t, 98
SOSA	self-one-step-ahead, 101
$\mathfrak{M}_n$	symmetric n-medium, 102
L_{R-P}	linear reward-penalty rule, 104
$\mathcal{A}_0$	limiting family of $\mathcal{G}^1(\odot)$ structures under $\mathcal{R}_0$, 109
${}_nL_{\underline{m}}^{(k_{ij})}$	linear family of automata with kernel (k_{ij}), memory m_i and n-actions, 111
$\mathcal{Z}_n$	operating zone for n-medium, 115
$\mathcal{B}_0$	limiting family of $\mathcal{G}^1(\circledast)$ structures under $\mathcal{R}_0$, 120
$\xi_{\circledast_k}^\alpha$	environmental equilibrium probability if in $\circledast_k$, 120
$\mathcal{G}^n(\circledast)$	n-level hierarchy of π-cells, 131

The Evolution of Stochastic Automata. D.E.Probert.

The fundamental idea underlying this research is that an initially randomly structured stochastic automaton placed in an environment can change its structure to increase its adaptation. The automaton gains information from the environment by executing actions and consequently receives reward or penalty stimuli with some probabilistic distribution.

The first part of the thesis concerns optimal reinforcement rules for markovian learning and a comprehensive theory is developed which also gives new insight into the operation of the many existing non-optimal rules. The behaviour of optimal rules in both static and dynamic environments is considered. An automaton evolving through "uniformly learning" rules which embrace both the new non-linear optimal rules and existing linear rules, is defined and called a Π-cell.

Next, a model for games between Π-cells is formulated and it is proved that they converge to pure saddle strategies when they exist. Deterministic approximations are found in the case of games with equilibrium mixed strategies and the trajectories are shown to be closely related to those of the Volterra model for predator-prey behaviour.

The thesis then develops a theory for the operation of networks of Π-cells. It is proved that starting from a random structure, the automaton evolves to an expedient structure which represents a discrete approximation to bayesian updating. The structure characteristically consists of a central action switch, surrounded by "arms" of memory states which reach out to the vertices of the likelihood simplex in which it is embedded. It is shown how such evolving automata can serve as a simple model of cellular differentiation. Finally it is shown how the theorems have natural extensions to Π-cells arranged hierarchically, which give greater adaptation in hierarchical environments.