INFERENCE FROM SEQUENTIAL DATA

MARCH 1974

B.R.Gaines,
Man-Machine Systems Laboratory,
Department of Electrical Engineering,
University of Essex, Colchester, Essex, U.K.

0.Preamble

These notes represent some preliminary reactions to the studies of learning automata based on string-predictors, PUSS & FELIX, by Andreae & Cleary (University of Canterbury, Christchurch, New Zealand - Electrical Engineering Reports, 14, 17, 20, 24 & 26), particularly in relation to complementary work (ATOM) in this Laboratory. I stress their informality since many of the questions stated herein are not well-posed and the analyses ill-conceived. However, this is (one half of) a conversation - or at least an experiment in co-operative, interactive research between antipodal workers.

These notes were started late in 1973, taken up again in late 1974, and I hope will now be completed under the stimulus of recent correspondence with John Andreae. In the meantime Luss and ATOM have both changed, not beyond recognition, and not necessarily for the better (an incentive to finish these notes).

To mark its transitory nature the definitive form of this document is a computer file (one day part of an interactive dialogue package !). Any printout (such as this) reflects only what used to be in it, not what is in it now.

0.1 Amble

The search for a universal learning machine, like that for the holy grail, is one where the (stimulating) means must justify the (mythical) end. It is no coincidence (except in Jung's (1972) or Koestler's (1972) terms) that a feline terminology permeates the Canterbury tales. Bast, the cat-goddess, is pre-christian and was worshipped in ancient Egypt as a symbol of life, maternity, and happiness. Interpreting this last through Virgil's felicitous phrase, 'Felix qui potuit rerum cognoscere causus' (happy is he who has been able to comprehend the causes of things), we can see that cats symbolize our concern in automata studies for human capabilities, self reproduction, and knowledge.

Thus the subconscious racial memories generating PUSS & FELIX go deep. ATOM is more Greek than Egyptian but through double entendre links the cultures - A (predatory) TOM cat in

one sense, but a unitary (and hence pure and perfect) constituent of learning molecules in another.

There are several strands to our work on learning machines and automata of which ATOM is one product. One goes back to my PhD thesis (Gaines 1972) and beyond, where a semigroup model of a purely behaviourist approach to psychology is developed, proved possible, but not applied to real data. ATOM actually implements that model and may be seen (and is being used) as a computer-based behavioural psychologist making very few assumptions and having no pre-conceptions. Another strand is that followed by Ian Witten who developed a STELLA-like learning system and compared it in perceptual-motor skills acquisition with human operators and other learning controllers (Witten & Corbin 1972). Ian became particularly interested in the 'dual control' aspects of acquiring knowledge whilst at the same time using it, and has since published several papers on this, particularly in its purest form as the 'two-armed bandit' problem (Witten zzzz).

The early PUSS studies (Andreae zzzz) stimulated us into developing optimal modelling/predicting algorithms to compare in behaviour and performance with PUSS on the published data. IW developed a deterministic automaton modeller (refined by Peter Facey who has been responsible for most of our computational tools, including the interactive string-handling language BASYS which has made it possible to program up new systems very quickly). The deterministic automaton modeller coped readily with the early examples of structured string prediction from Canterbury, but BPG was interested in modelling the same data without distinguishing inputs from outputs. This meant it became stochastic data which created new problems. One result of studying these was a paper demonstrating that the mean number of states in a deterministic modeller of a properly stochastic sequence was virtually the same as the string of observations (gaines 1975).

This ruled out deterministic modellers once and for all, but also suggested that it would be interesting to discover how human beings, with their predilections towards a causal world-view, would cope with modelling stochastic sequences. Chris Rogers, a graduate student, has been studying human behaviour in PUSS-like tasks. This has somewhat tangled links with the other work because we have pitted human beings against a variety of sequence generators including a PUSS predictor modelling the human's behaviour ! Also we have used the automaton modellers in ATOM to analyse the joint behaviour of human and sequence generator.

ATOM itself is a suite of programs written in BASYS for the PDP11/45 designed for general experimentation with string analysis and prediction in the PUSS problem environment. It runs both interactively for experiments with on-line string input, games and prediction, and in batch for long-term modelling experiments (ATOM currently absorbs all the night time running of the PDP11 doing odd jobs like in-depth analysis of the distribution of times necessary to solve Bill & Ben (Andreae 1974), and trying to predict the FT ordinary

share index !). The base level of ATOM is a data acquisition program which accepts commands and data, either from keyboard or from files, and can call upon any of ATOM's predictors to give estimates of the next data items, or its structure. This can also be slaved by other jobs to allow use of ATOM in other programs such as learning machines.

The main predictors implemented in ATOM are a FELIX-like algorithm which bases predictions on the immediately preceeding sub-string - it will either search for discriminating sub-strings and give the complete set, or produce a confusion matrix for a given restriction on the length of sub-string. This is used as a basic predictor and to examine the power of PUSS-like algorithms in particular cases (most of the data in the early Canterbury tales have been re-run through this system). We are developing 'optimal' PUSS-lie strategies based on this taking into account the balance between the frequency of sub-string occurrence and the limited memory resolution of short sub-strings. We have had some interesting experiences with the interactive predictor and certainly agree that such systems are full of surprises and hence highly interesting to work with interactively.

The pride of the ATOM suite is, naturally, a non-deterministic automaton modeller which is theoretically optimal according to some very well defined criteria (of this much more later). It will produce either Moore or Mealy models as required and gives the best fitting model with a given number of states. Thus the output is a graph of goodness-of-fit against complexity of model, a kind of 'spectral' analysis in which the structure of the input sequence generator (if it has one) is very apparent. Associated theoretical developments are: (a) very well-based criteria of goodness-of-fit linked closely to the literature on subjective probability and computational complexity (these should be applicable to all related work); (b) that the best model is always observable; (c) that we get the right model if the sequence generator is actually a stochastic automaton. Practical developments are that we do not have to tell the modeller which symbols are inputs and which outputs - it can tell us (we can tell ATOM about inputs, outputs and delimiters - it is much faster but less interesting), and that we have coped well with a lot of data, the Canterbury data, human generated data, and artificial examples. The interesting question is, of course, the relationship of these modellers to PUSS/FELIX.

0.3 Post Amble

0.3.1 The Philosophical Literature

One reason these notes have been so long in writing is that there is no much to say and it is not sequential in form. There are many byeways of varying significance and credibility. The most potent source of these is the

philosophical literature since in attempting to solve problems of machine learning we are formulating and realizing concepts whose philosophical foundations are insecure and extremely well-probed. The danger is that the relevant literature only provides arguments and evidence for the inadequacy of any particular approach. The more one investigates it the less likely one is actually to do anything.

I do not believe that this literature can be reasonably neglected by workers in the field of machine learning - it is an obstacle course that must be run sooner or later. The trick is to use one obstacle to surmount another, e.g. Popper's 'Objective Knowledge' (1972) followed by Feyerabend's (1974) critique of it coupled with Takatos's (1970) alternative approach and Gellner's (1974) dismissal of the lot of them !

When it comes to fundamentals it is clear that there are no rational foundations for even the weakest assumptions necessary to make machine learning (or human learning) possible. If the dryish arguments in the above references appear too verbal and legalistic, then the more impassioned leadership of Monod (1970) reinforces the point in a different way. His way out typifies the arguments one seems forced to use in avoiding sheer arbitrariness of postulates, but again it has its impact only through the counter-voices of its critics (Lewis (1974)). The third leg of the philosophical tripod might well be the Jung & Koestler works above which provide cogent arguments for links between events that do not have the cause/effect structure so axiomatic to our concepts of science (and, perhaps, so basic to the human psyche).

0.3.2 The Two Basic Postulates

Despite all this debate, in machine learning we still seem forced back to the position that Hume showed was untenable, that what has happened already is likely to happen again, coupled with some postulate of Ockham-like simplicity, that the best explanation is the least complex. For the present studies some formalization of these two postulates seems adequate.

0.3.3 Constrain The Problem Not The Solution

When it becomes necessary to add further postulates (hopefully not too soon before the surprising richness of these first two are fully explored and exploited) I feel strongly that these should relate to the problem rather than to its solution. At this stage of development (when we do not have theoretically sound or pragmatically useful algorithms) it is appropriate to ask , 'how will our system behave in this restricted environment', or 'how can we use these restrictions to improve its performance'.

It is totally inappropriate to raise questions as to the speed, complexity, cost, memory capacity, etc. of proposed

solutions. These 'engineering' considerations may one day become important, but for the present they only serve to confound phenomena whose source is of prime importance. For example, a proper algorithm may only require a limited memory capacity for its models - this will not be noticed if the capacity has already been restricted artificially. For example, many observed behavioural phenomena in natural species have been described in pseudo-causal terms (e.g. 'forgetting' is an observed behaviour not a causal activity) and there is a danger that these may be built into learning automata as causes when they should rather be looked for as effects. Any proper algorithm is liable to show 'illusions' & 'forgetting' - this is confounded if either is deliberately 'built-in'.

0.3.4 There Are No Limits To Memory Capacity

I object particularly to any attempt to limit the memory capacity of learning automata on 'engineering' or 'psychological' grounds. This has no justification - our memory technology is improving at a far faster rate than our learning automata requirements and we not justified in assuming anything except potentially infinite memories at virtually zero costs - equally there is no psychological evidence to demonstrate memory bounds in the human. The most reasonable assumption to make based on our current imperfect experimental techniques is that the human powers of retention are absolute and that all data perceived is completely retained throughout the lifetime. The phenomena commonly associated with 'forgetting' are most realistically ascribed to the mechanisms of retrieval than to the 'memory' (e.g. to an imperfect associative datum now) and may be expected to be shown by any associative retrieval system.

0.3.5 Subjective Evaluation Is Dangerous

One further danger is that of our own limitations. We are not good arbiters on the intelligence of others. The behaviour of even simple systems with random elements is beyond our unaided comprehension. It is essential to have objective normative standards by which the behaviour of a learning system may be judged. In particular human standards are important but inadequate since the task of inference from sequential data is one at which humans are poor.

0.3.6 But Interaction Is Vital

This is not an argument against the importance of interaction with our learning automata - far from it since the very weakness of our own powers of comprehension demands that we interact as closely as possible with the learning automaton we are attempting. to understand. Rather it is a warning against subjective evaluation, judgements based on our

interaction - we should only judge the power and success of our automata against external, normative standards.

1 Position

1.1 Fundamental Problem

To make inferences from data.

Formally I shall take a set of data to be an ordered multiset, that is a set consisting of elements, possibly repeated, such that an order relation is defined over the whole set (see Goguen (1974 p.541) for some remarks on multisets).

Formally I shall take the problem of inference to be that of adding elements to the ordered multiset.

Formally I shall base the notion of 'goodness of inference' on two further order relations:-

(a) One over specific inferencing methods (models) relating to their 'complexity';

(b) The other over the 'goodness of fit' of the inferencing models to the ordered multiset.

Notes

(1) This definition of data is general enough to include unordered and ordered sets, or multisets, and the ordering may clearly range from none, through partial orders, lattices, and chain orders.

(2) Since an ordered set is itself a category this problem defininition might be further generalized to that of adding elements to an arbitrary category. The theory of extensions of a variety of mathematical structures is quite rich. Until we have more results for the problem as stated, however, the value of such generalization cannot be evaluated.

(3) There is also a case for weakening the order over the multiset, cetainly to a pre-order, and perhaps to an arbitrary relation. Again the merits cannot yet be evaluated.

(4) Adding elements to the multiset within the order may be called 'interpolation' and outside it 'extrapolation'.

(5) The order need not necessarily be one of time since it is hoped in this formulation to capture problems of pattern recognition with parallel data as well as prediction from serial data.

(6) If the order represents a time sequence it need not

be a total order since this formulation is also intended to capture the problem of inference from multiple observers who do not necessarily also observe one another.

(7) If the order represents a time sequence then extrapolation to greater elements may be called 'prediction' and to lesser elements 'retrodiction'.

(8) In the basic formulation prediction and retrodiction are clearly symmetric. Any asymmetry introduced (Watanabe (1969)) is a function of the order relation on our models, not on the data.

(9) Prediction & retrodiction may be viewed as extrapolating on the basis of minimum iterpolation.

(10) The 'negative instances' (of non-allowed additions) so important in grammatical inference (Feldman 1972) may be incorporated in this framework. For the present however I feel that they have a better sematics as restrictions on allowable models rather than restrictions on allowed additions. The apparent symmetry between 'positive' and 'negative' instances hides an underlying epistemological difference in status which is far more important.

1.2 Structure Of Document

There are many interesting considerations as soon as one goes beyond pure inference and, for example, takes into account its purpose. I regard the ultimate application of PUSS-like predictors to STELLA-type (Gaines & Andreae 1966, Andreae & Cashin 1969) as of major interest. There is also an incredible variety of psychological phenomena (arousal, set, curiosity, forgetting, etc.) which seem to have natural explanations in terms of the consequences of algorithms solving the inference problem stated. There are also the applications of the inferencer to real data, and the theoretical consequences of assuming the existence of optimal inferences and analysing the 'science' that may be based on them.

There are two major chunks of material which naturally precede any of these developments - firstly, a detailed analysis of the simplest form of the problem stated - and secondly, some comparative results with the ATOM programs using the data previously fed to PUSS at Canterbury. This constitutes the following two sections of this report.

2 The Simplest Form Of Problem

This section is concerned to explore thoroughly the simplest significant form of the problem stated in 1.1 above, in which the order on the multiset is total and prediction is required. This was the original problem posed to PUSS and FELIX (string prediction) and is the sole function of each of

the algorithms in ATOM. It has a substantial literature, nowadays largely under the heading of the problem of 'grammatical inference' (zzzz) and interacts strongly with the problem of re-axiomatizing probability theory on a basis of 'computational complexity' (zzzz).

8 References

J.Feldman, Some decidability results on grammatical inference and complexity, Inf. & Contr. vol.20, pp.244-262, 1972.

P.Feyerabend, Popper's Objective Knowledge, Inquiry vol.17, pp.475-507, 1974.

E.Gellner, Legitimation of Belief, Cambridge: University Press, 1974.

J.A.Goguen, Concept representation in natural and artificial languages, Int.J.Man-Machine Studies, vol.6, pp.513-561, 1974.

C.G.Jung, Synchronity - An Acausal Connecting Principle, London: Routledge & Kegan Paul, 1972.

A.Koestler, The Roots Of Coincidence, London: Hutchinson, 1972.

J.Lewis (ed.), Beyond Chance And Necessity, London: Garnstone Press, 1974.

J.Monod, Chance And Necessity, London: Collins, 1972.

K.R.Popper, Objective Knowledge - An Evolutionary Approach, Oxford: University Press, 1972.

I.Takatos, Falsification and the methodology of scientific research programmes, in I.Takatos & A.Musgrave (eds.) Criticism And The Growth Of Knowledge, Cambridge: University Press, 1970.

S.Watanabe, Knowing And Guessing, New York: Wiley, 1969.