# COMPLEXITY AND APPROXIMATION: THE FOUNDATIONS OF PATTERN RECOGNITION

Brian R. Gaines

Department of Industrial Engineering
University of Toronto, Ontario, Canada M5S 1A4

This paper gives a general formulation of the problem of system modelling that encompasses many aspects of: pattern recognition; clustering, grammatical inference, system identification, and so on. The formulation allows various problems, and approaches in these diverse, but related, areas to be placed within a single canonical framework so that the similarities, and differences, in approaches become apparent and well-defined.

## INTRODUCTION

The **syntactic** approach to pattern recognition has attracted much interest in recent years[4,15]. Some pattern-recognition problems, such as waveform analysis[17], which have a natural sequential component leading to a linguistic approach based on grammatical inference (GI)[5]. Many problems that are not in themselves sequential can best be treated as such, particularly when the process of information acquisition itself involves costs.

Deterministic inference is inadequate for practical problems where observation uncertainties make the problem non-deterministic one. The key difficulty in any form of non-deterministic GI is that the notion of a **solution** to the problem is not well-defined. Whereas inferring a grammar for a complete set of productions from a member of a class of deterministic grammars has a well-defined solution[1,2,13], the problem of inferring the stochastic grammar that has generated the sample presented has no unique solution. At one extreme the generator might have been a zero-memory Bernoulli process, and at the other it could be a deterministic grammar generating just those productions observed.

In 1975 Gaines[6] resolved this problem by accepting the lack of unicity and defining the solution to be an **admissible set** of grammars such that a **simpler** grammar than one in the set was a **worse fit** to the observations, while a **better fit** grammar than one in the set was more **complex**. He showed this led to algorithms for stochastic GI that were effective in speed, compactness and in obtaining **correct** solutions to a variety of problems[7]. He developed computer programs[8,9] and showed their application to inductive inference and pattern recognition. The methodology was extended to **fuzzy** behavior generated by **possibilitic** systems[10], recursive behavior generated by push-down automata[22], and placed within a category-theoretic framework that applies to both stochastic and fuzzy systems[19].

This paper develops the general formulation, shows how the problem of system modelling common to a wide variety of decision problems splits into general sub-problems of: choice of descriptive vocabulary; choice of class of models; choice of complexity ordering over models; choice of approximation measures between observations and models; algorithms for determining the admissible set of models based on these choices and the data gathered; and algorithms for action, including that of gathering further data. The aim is to give a unified framework for diverse, but related, problem areas, allowing, and encouraging, the transfer of concepts and methodologies from one to another.

## THEORETICAL FOUNDATIONS

Let **E** be a set of possible exemplars of system behavior, e.g. the set of all possible patterns on a grid, or the set of all possible strings of input-output behavior of an automaton. Let **i: I→E** be a sub-object monomorphism from an **index-set I** to **E** that represents the **observed**, or **classified**, sub-set of **E**. Let **b: I→V** be a mapping from **I** to a **truth-set V** which represent possible observational assignments of values to elements of **E**. Goguen[14] calls such a mapping a **V-set** with **I** as carrier; **V** will normally be some ordered algebraic structure such as a lattice, the interval [0,1] with max/min, add/multiply, or logical operations[12]. For example, **V** might be a set of possible pattern classes, or the interval [0,1] with the mapping **b** corresponding to the observed relative frequencies of the exemplars in **E**. One might term **b** the **behavior** of a system, and take the problem to be one of inferring a **model** that accounts for **b**.

Let **M** be a set of mappings from **E** to **V** such that any **m∈M** represents a possible **model** of behaviors such as **b**. Fig. 1 shows the mappings under consideration. The problem is to find an **m** that is an extension of **b** from the sub-object of **E**, **I**, to the whole of **E**, so that **im=b**.

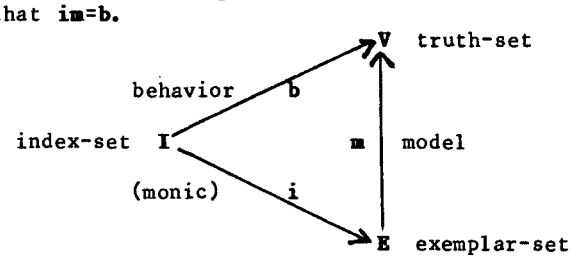


Fig. 1 Modelling as Extension of a Mapping

How do we select amongst the many possible extensions of **b**? In particular situations, there usually will be some criterion that, other things being equal, allows us to select one model as being preferable to another, e.g. that it has the least **parameters**, or the least **states**, or some other Ockham's razor type of criterion. In general, all such criteria can be subsumed under the requirement that there is some pre-order on **M**, $\leq$, such that the lesser models in the pre-order are intrinsically preferred to those greater, i.e. if $m,n \in M$ are two models that are both valid extensions of some behavior **b**, and $m \leq n$, then we will not prefer **n** to **m** as a model of **b**. This preference order is not necessarily of Ockham's razor type, for example it could be one based on elegance or the party-line, but it is so often one of complexity that it is convenient to use this terminology and say that if $m \leq n$ then **n** is not less **complex** than **m**.

At this stage our formulation of the problem of system identification might seem complete - "out of all models that are valid extensions of the observed behavior chose the least complex." However, while this is reasonable for the deterministic case where the behavioral mapping **b** is well-defined and can be precisely extended, it is not adequate for the non-deterministic case. For example, if **b** represents an observed behavior distribution for a stochastic system then determining a model that precisely generates the observed relative frequencies is a number-theoretic problem whose solution has little relevance to actual requirements. What we want is a model that approximates the observed frequencies to the extent that residual differences between behavior and model are statistically insignificant. Equally, if **b** represents a given **degree of membership** assigned to the possible behaviors of a fuzzy system, then what we require is a model that places the same fuzzy restriction on behavior within a reasonable tolerance. In general, we do not require **m** to be an exact extension of **b** but rather a reasonable approximation to such an extension. The diagram of Fig. 1 need not commute, only **almost commute**.

In a particular situation one will have highly specific concepts of what one means by **almost commutes**, i.e. to what extent **im** approximates **b**. However, in general the key notion is again one of order, that it should be possible to say that one model is a better, or worse, approximation to **b** than is another. To do this one needs to define there to be a mapping from the set of behaviors, to the set of pre-orders on models, $\mathbf{Ord_M}$, $\mathbf{f: B \rightarrow Ord_M}$, such that if $\leq_b$ is the image of **b** under **f**, then, for $m,n \in M$, $m \leq_b n$ means that **n** is not a better approximation to **b** than **m** is.

Having defined the pre-orders of complexity and approximation, we are in a position to define a solution to the identification problem in terms of their product, $\leq_b^*$, defined as:

$$\forall\, m,n \in M,\quad m \leq_b^* n \leftrightarrow m \leq n \text{ and } m \leq_b n$$

i.e. $m \leq_b n$ if and only if **n** is both neither less complex nor a better approximation to **b** than **m**. The minimal elements in this order are all **admissible** solutions to the identification problem because they cannot be bettered in complexity without worsening the approximation, and they cannot be bettered in approximation without worsening the complexity. They form the admissible subspace of models determined by **b**, $M^* \in M$, such that:

$$M^* \equiv \{m: \forall\, n \in M,\ n \leq_b^* n \rightarrow m \leq_b^* n\}$$

i.e. if any model is better than one in $M^*$ then it is equivalent to it.

The next section shows how this formulation allows the general problem to be broken down into sub-problems, and discusses the significance of each and the techniques available for their solution.

## ASPECTS OF IMPLEMENTATION

First, consider the set **M** of possible models. It is clear that the selection of **M** alone is a small step in setting up the modelling problem. The complexity ordering over the models must also be specified and variation of this affects the solution to the problem as much as does the choice of the models themselves. For example, finite automata may be ordered by number of states, number of links between states, or by some more complex function that takes into account hierarchical structure. Zeigler[23] gives several such order relations on automata and we have found that changing between them has significant effects in practical modelling situations, both on the speed of computation and the solutions found.

Second, consider the set **E** the set of possible exemplars. The **language** in which these are described is arbitrary and can be varied without affecting the essential modelling problem. However, it is clear that changes in **E** will be reflected by changes in the admissible sub-set of models, not only in the trivial sense that the models themselves must change to reflect the variations in **E**, but also in that the complexity ordering on models must also change if the situation is to remain invariant. This is unlikely to be so since this ordering will generally be derived from intensional, extra-linguistic, considerations.

Third, the implementation of a modelling scheme once the various ordered sets required have been defined is basically simple. One needs only an algorithm that generates possible models in order of increasing complexity. As each **m** is generated it is checked against **b** for its position in the approximation order. If this is is worse than that of simpler models previously found the model is rejected, but otherwise it is added to the admissible set. The key trick here is the derivation of a production algorithm for models for complete enumeration of models of increasing complexity, and Wharton[21] has given some techniques for doing this for a wide classes of models.

Fourth, the measures of approximation used are very significant and it is interesting that these seem to have a high degree of problem independence, largely because the set **V** is basically problem independent, a classification, or a distribution, or a fuzzy restriction. For example, consider **V** to be the interval [0,1] with the mappings **b** and **m** giving normalized relative frequency distributions. A suitable measure of approximation between **b** and **m** is

then one that has a minimum when $\forall$ e$\in$E, me=be (assuming be is suitably defined when e$\notin i^{-1}$E), and varies above this minimum to the extent that they differ. Examples of such measures are:

$$SE = \sum_{e \in E} (me-be)^2 \quad (1)$$

$$CE = \sum_{e \in E} (me-be)^2/me \quad (2)$$

$$LE = \sum_{e \in E} -be \times \ln(me) \quad (3)$$

Measure (1) is a standard least-mean-squares test used extensively in the literature on modelling and pattern recognition, and by de Finetti[3] as the foundation for his theory of probability. Measure (2) is the chi-square statistic used as a statistical test of similarity of distributions in GI[16,11]. Measure (3) is a Gibb's function generating a Shannon entropy measure widely used in GI and in Savage's model of subjective probability[20]. Pearl[18] has integrated these and other such functions into a single economic model.

Fifth, the determination of the admissible set is often not in itself the final solution to a modelling problem. Pattern recognition is carried out for a purpose and we have to decide which model to actually use out of the admissible set. Often this will involve action to obtain further data on further exemplars to provide for better discrimination between admissible models.

## CONCLUSIONS

The methodology of modelling proposed here is intended to clarify what most successful schemes actually do in practice, and to provide a rigorous basis for the comparison of a wide variety of schemes both within individual disciplines and across them. What we have found helpful in our own studies based on the conceptual framework outlined is the clear separation it effects between different sub-problems and the emphasis it places on the role of each. In particular it encourages one to vary the class of models and the complexity order on existent classes in order to obtain schema best-suited to the range of problems being studied. There is less scope for useful variation in the models of approximation, and the interesting feature of these is their general applicability over a wide range of problem areas.

## ACKNOWLEDGEMENTS

I am grateful to John Andreae, Michael Arbib, Richard Dawkins, Joe Goguen, Susan Haack, George Klir, Ladislav Kohout, Judea Pearl, Adrian Walker, Ian Witten, and Lotfi Zadeh, for discussions that have influenced the work discussed in this paper. Financial assistance for this work has been made available by the National Sciences and Engineering Research Council of Canada.

## REFERENCES

[1] Arbib, M.A. and Manes, E.G., Foundations of system theory: decomposable systems, **Automatica, 10** (1974) 285-302.

[2] Ehrig, H., **Universal Theory of Automata** (B.G.Teubner, Stuttgart, Germany, 1974).

[3] de Finetti, B., **Probability, Induction and Statistics** (John Wiley, London, UK, 1972).

[4] Fu, K.S., **Syntactic Pattern Recognition and Applications** (Prentice-Hall, New Jersey, 1982).

[5] Fu, K.S. and Booth, T.L., Grammatical inference: introduction and survey, **IEEE Transactions on Systems, Man & Cybernetics, SMC-5** (1975) 95-111 (Pt.I), 409-423 (Pt.II).

[6] Gaines, B.R., Approximate identification of automata, **Electronics Letters, 11** (1975) 444-445.

[7] Gaines, B.R., Behaviour/structure transformations under uncertainty, **International J. Man-Machine Studies, 8**(3) (1976) 337-365.

[8] Gaines, B.R., System identification, approximation and complexity, **International J. General Systems, 3** (1977) 145-174.

[9] Gaines, B.R., General system identification - fundamentals and results, in Klir, G.J. (ed.), **Applied General Systems Research: Recent Developments and Trends** (Plenum Press, New York, 1978) 91-104.

[10] Gaines, B.R., Sequential fuzzy system identification, **Fuzzy Sets and Systems, 2**(1) (1979) 15-24.

[11] Gaines, B.R., Maryanski's grammatical inferencer, **IEEE Trans. Computers, C-28**(1) (1979) 62-64.

[12] Gaines, B.R. and Kohout, L.J., The logic of automata, **International J. General Systems, 2**(4) (1976) 191-208.

[13] Goguen, J.A., Realization is universal, **Mathematical Systems Theory, 6** (1973) 359-374.

[14] Goguen, J.A., Concept representation in natural and artificial languages: axioms, extensions and applications for fuzzy sets, **International J. Man-Machine Studies, 6**(5) (1974) 513-561.

[15] Gonzalez, R.C. and Thomason, M.G., **Syntactic Pattern Recognition: An Introduction** (Addison-Wesley, Massachusetts, 1978).

[16] Maryanski, F.J. and Booth, T.L., Inference of finite-state probabilistic grammars, **IEEE Transactions on Computers, C-26** (1977) 521-536.

[17] Pavlidis, T., Linguistic analysis of waveforms, in Tou, J.T. (ed.), **Software Engineering** (Academic Press, New York, 1971) 203-225.

[18] Pearl, J., An economic basis for certain methods of evaluating probabilistic forecasts, **International Journal of Man-Machine Studies, 10**(2) (1978) 175-183.

[19] Ralescu, D., A system theoretic view of social identification, in Ericson, R.F. (ed.), **Improving the Human Condition: Quality and Stability in Social Systems** (Society for General Systems Research, Kentucky, 1979) 531-538.

[20] Savage, L.J., Elicitation of personal probabilities and their assessment, **Journal American Statistical Association, 66** (1970) 783-801.

[21] Wharton, R.M., Grammar enumeration and inference, **Information & Control, 33** (1977) 253-272.

[22] Witten, I.H., Some recent results on non-deterministic modelling of behaviour sequences, in Reckmeyer, W.J. (ed.), **General Systems Research and Design: Precursors and Futures** (Society for General Systems Research, Kentucky, 1981) 265-274.

[23] Zeigler, B.P., Simulation based structural complexity, **International J. General Systems, 2** (1976) 217-223.