[4] R. Viswanathan and K. S. Narendra, "Application of stochastic automata models to learning systems with multimodal performance criteria," Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-40, 1972.
[5] E. M. Vaysbord and D. B. Yudin, "Multiextremal stochastic approximation," *Eng. Cybern.*, vol. 6, No. 5, pp. 1–11, 1968.
[6] L. S. Gurin, "Random search in the presence of noise," *Engrg. Cybern.*, vol. 4, No. 3, pp. 252–260, 1966.
[7] L. S. Gurin and L. A. Rastrigin, "Convergence of the random search method in the presence of noise," *Automat. Remote Contr.*, vol. 26, No. 9, pp. 1505–1511, 1965.
[8] G. J. McMurtry and K. S. Fu, "A variable structure automaton used as a multimodal searching technique," *IEEE Trans. Automat. Contr.*, vol. AC-11, No. 3, pp. 379–387, July 1966.
[9] E. M. Braverman and L. I. Rozonoer, "Convergence of random processes in learning machines theory," *Automat. Remote Contr.*, vol. 30, No. 1, pp. 57–77, 1969.
[10] M. Loeve, *Probability Theory*. Princeton, N.J.: Van Nostrand, 1963.
[11] L. A. Rastrigin, "The convergence of the random search method in the extremal control of a many-parameter system," *Automat. Remote Contr.*, vol. 24, No. 11, pp. 1337–1342, 1963.
[12] A. S. Poznyak, "Use of learning automata for the control of random search," *Automat. Remote Contr.*, vol. 33, No. 12, pp. 1192–2000, 1972.
[13] G. R. Gucker, "Stochastic gradient algorithms for searching multidimensional multimodal surfaces," Stanford Univ., Stanford, Calif., Tech. Rep. TR-6778–7, 1969.
[14] E. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, No. 3, pp. 462–466, 1952.
[15] R. A. Jarvis, "Adaptive global search in a time-variant environment using a probabilistic automaton with pattern recognition supervision," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-6, pp. 209–217, July 1970.
[16] L. D. Cockrell and K. S. Fu, "On search techniques in adaptive systems," Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE-70-1, 1970.
[17] R. Viswanathan and K. S. Narendra, "Stochastic automata models with applications to learning systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 107–111, Jan. 1973.
[18] G. N. Saridis and H. D. Gilbert, "On the stochastic fuel-regulation problem," Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE-68-9, 1968.
[19] J. Matyas, "Random optimization," *Automat. Remote Contr.*, vol. 26, no. 2, pp. 244–251, 1965.
[20] R. Viswanathan and K. S. Narendra, "A note on the linear reinforcement scheme for variable structure stochastic automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 292–294, Apr. 1972.
[21] V. Ya Katkovnik and O. Yu. Kul'chitskii, "Convergence of a class of random search algorithms," *Automat. Remote Contr.*, vol. 33, no. 8, pp. 1321–1326, 1972.
[22] E. M. Braverman and L. I. Rozonoer, "Convergence of random processes in learning machines theory; part 2," *Automat. Remote Contr.*, vol. 30, no. 3, pp. 386–402, 1969.
[23] J. D. Hill, "A search technique for multimodal surfaces," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, pp. 1–11, Jan. 1969.
[24] L. P. Devroye, "A mixed stochastic optimization algorithm and its applications in pattern recognition," in *Proc. IEEE Conf. Decision Control*, San Diego, Calif., pp. 356–360, 1973.
[25] H. J. Kushner, "Stochastic approximation algorithms for the local optimization of functions with nonunique stationary points," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 646–654, Oct. 1972.
[26] B. T. Polyak and Ya. Z. Tsypkin, "Pseudogradient adaptation and training algorithms," *Automat. Remote Contr.*, vol. 34, no. 3, pp. 377–397, 1973.
[27] S. H. Brooks, "A discussion of random methods for seeking maxima," *Oper. Res.*, vol. 6, no. 2, pp. 244–251, 1958.
[28] A. N. Mucciardi, "A new class of search algorithms for adaptive computation," in *Proc. Conf. Decision Control*, San Diego, Calif., pp. 94–100, 1973.
[29] G. J. McMurtry, "Adaptive optimization procedures," in *Adaptive, Learning and Pattern Recognition Systems*, J. M. Mendel and K. S. Fu, Eds. New York: Academic, 1970.
[30] G. A. Bekey and M. T. Ung, "A comparative evaluation of two global search algorithms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, pp. 112–116, Jan. 1974.
[31] A. S. Poznyak, "Learning automata in stochastic programming problems," *Automat. Remote Contr.*, vol. 34, no. 9, part 2, pp. 1608–1619, 1973.
[32] G. V. Karumidze, "A method of random search for the solution of global extremum problems," *Engrg. Cybern.*, vol. 7, no. 6, pp. 27–31, 1969.
[33] H. B. Mann and A. Wald, "On stochastic limit-order relationships," *Ann. Math. Stat.*, vol. 14, 1943.

Correspondence

On the Complexity of Causal Models

B. R. GAINES, MEMBER, IEEE

***Abstract*—It is argued that a principle of casuality is fundamental to human thinking, and it has been observed experimentally that this assumption leads to complex hypothesis formation by human subjects attempting to solve comparatively simple problems involving acausal randomly generated events. This correspondence provides an automata-theoretic explanation of this phenomenon by analyzing the performance of an optimal modeler observing the behavior of a system and forming a minimal-state model of it.**

I. Introduction

The postulation of a principle of causality, "to every effect there is a cause," has been a continuing central problem for philosophy [1]. Its role as a source of contention in modern science [2] is epitomized by Einstein's remark that, "I can't believe that God plays dice." Many of the arguments about the application of the principle are very relevant to systems science and to problems of system identification and machine learning, on the one hand, and to epistemology and behavioral psychology, on the other. In current system science the theory of causal deterministic systems is most well developed and generally applied, while the theory of modeling with alternative structures, e.g., stochastic automata, indeterminate automata, products of asynchronous automata, etc., has not been developed to the same degree.

There is a significant parallel here in human psychology—some of Michotte's [3] experiments on perception demonstrate that causality can be a very low level percept, similar in status to shape and colour. The deep-felt assumption of causality, typified by Einstein's remark, is apparent in cognitive tasks also. In our research on machine learning [4]–[6] we used as one test environment a version of nim in which the learning machine, played against a partially random player of variable optimality. To demonstrate that the task was nontrivial for man we also

Manuscript received January 27, 1975; revised June 4, 1975.
The author is with the Man–Machine Systems Laboratory, Department of Electrical Engineering Science, University of Essex, Colchester, England.

built an automaton with four lamps (representing the number of matches left) and three keys (representing taking one to three matches) and a "reward" light indicating that the game had been won. In practice, human beings found the game virtually impossible to learn in these circumstances, whereas in the normal representation of nim it is trivial.

Much of the problem in learning seemed to stem from the partially random opponent in the automaton. Although this random element makes the game easier to win (the opponent makes fatal errors) it also makes the behavior of the automaton partially indeterminate. A hypothesis of randomness or indeterminism was *never* introduced by human players against the automaton, and the longer they played the more confused they became.

To investigate this further we built a simple automaton with two lamps and two keys such that one lamp came on for a short period when a key was depressed. The problem was to press the key under whichever lamp was going to come on next. In fact when a key was depressed either lamp came on at random with equal probability. Again in tests with a wide range of subjects this random element was never recognized, and individuals were prepared to pit their wits against the automaton for many hours and most seemed to delight in doing so.

A number of records were made of verbal introspection as subjects played against the automaton. One mathematician, after some 30 minutes of play, came to the conclusion that the automaton had two modes of behavior. In one the same light tended to come on as the key depressed while in the other the opposite light tended to come on—this is tautologous but was regarded as an empirical hypothesis. Most people tended to generate more elaborate models but felt that they were "not quite correct."

One of the most interesting records is that of a behavioral scientist who played with the box for some 30 minutes and kept a tally of the occasions when he was correct less the occasions when he was wrong. After that period he was over 20 ahead on his count and announced that he had a good strategy but felt that he should drop it for a while and experiment with alternatives to seek a better one. His accumulated count then rapidly declined and when it went negative he announced that he was going back to his "good strategy." The count continued to decline, however, and he finally announced that the box contained a system that modeled his strategy and then structured itself to act in the opposite manner and thus outwit him!

Such a form of "frustration automaton" is possible and a construction of a universal form for deterministic players has been given [7]. However, the hypothesis of such a complex structure on the basis of interaction with a simple two-state stochastic automaton is a fascinating phenomenon in its own right. The remainder of this correspondence will demonstrate that such a elaborate models are necessarily generated if causality is postulated in modeling acausal phenomena.

II. Expected Size of Causal Model

The following abstract situation captures all the relevant features of a modeler forming a causal model of some other system based on his observations. The observed behavior is some sequence drawn from the union of an input alphabet X, and an output alphabet Y, in which no two input symbols are adjacent (this constraint is applied to simplify the definition of the sequence "length"—it may easily be satisfied by introducing a null-output symbol). Only finite sequences of observations will be considered, and the *length* of such a sequence will be defined as the number of members of the output alphabet in it. For example, if $X = \{I,J\}$, $Y = \{A,B\}$, then a sequence of behavior might be

$$AIBBAJBIAAJ$$

where the "length" would be seven. Note that no assumptions are made about the structure of the system generating the observed behavior.

Now consider a *modeler* whose task is to explain the observed sequence of behavior in terms of some model from a class of possible models. We are interested in the case in which the modeler assumes *a priori* that the observed behavior is generated by a causal system and forms a deterministic model of it as a state-determined machine. For any finite sequence of observed behavior there will be an unbounded number of finite-state machines whose input–output behavior, starting from a specified state, is identical to that observed. Out of these possible models there will be some such that no other machine has a smaller number of states. An *optimal causal modeler* is defined as one who always chooses such a minimal-state machine as a model.

What we now wish to derive is some relationship between the length of the sequence of observations and the number of states in the model formed by an optimal causal modeler. The following simple properties may be stated:

1) the number of states is a monotonically nondecreasing function of the length of the observed sequence of behavior;
2) the number of states in the model of a sequence of behavior generated by an M-state deterministic automaton cannot exceed M;
3) the number of states cannot exceed the length of the sequence of observations;
4) there are observed behaviors of length N such that the model must have N states.[1]

The first two properties seem to express the intuitive expectations of a human causal modeler—after a sufficiently long sequence of observations his model should stabilize. The last two properties show that it is possible, however, for the size of model to grow at the highest possible rate, precisely as the number of observations. That is, in terms of the discussion of human behavior, an "unlucky" observer may obtain data requiring a very complex model. However, if the system generating the behavior is relatively simple, for example, finite state, even if acausal, we might still expect our optimal causal modeler to perform in a similar way to when modeling a causal system. Clearly a stochastic source will generate sequences requiring as many states in the model as the length of the sequence, but it is plausible to suppose that such "pathological" sequences might be generated only infrequently by a finite-state automaton.

Hence a more interesting characteristic of the observer's behavior would be, not the maximum complexity model he might generate, but instead the *expected* number of states in the model generated (the average complexity). This might have one of three possible behaviors, as a function of the length of the observation sequence N. The expected number of states in the model formed by an optimal causal modeler of the behavior of a finite-state stochastic automaton might be the following:

a) asymptotic to a finite number, i.e., closely similar to the corresponding situation when the behavior modeled is generated by a finite-state deterministic automaton;

[1] For example, consider the sequence $A^{N-1}B$ or the infinite sequence $ABA^2B^2A^3B^3$.

b) grow without limit but slower than the number of observations N itself, for example, as $\log N$; one might hypothesize that at least the ratio

$$R_N = \frac{\text{expected number of states}}{\text{number of observations}} \tag{1}$$

would tend to zero as the number of observations increased;

c) grow without limit at a rate similar to the maximum possible N; this would imply that nearly all sequences generated were "pathological" requiring maximum-size models growing as fast as the number of observations.

The recent development of theories of probability based on computational complexity [8]–[11] shows that case c) in fact occurs, and it is probably a measure of the extent to which the result is counterintuitive that this basis for probability has been so late in developing. We have so long accepted that a random sequence may have any structure, including possibly one that appears highly deterministic, that it comes as a shock when it is suggested that we can apply a test for randomness to an individual sequence rather than a distribution. The paradox is resolved, of course, because almost all randomly generated sequences may be shown to have high computational complexity [11].

The following additional property shows the significance of these results in the present context—case c) occurs and the ratio of (1) can tend to unity:

5) The expected number of states in the model formed by an optimal causal modeler observing a sequence of behavior of length N may be at least $N - \log_2 N - 2$.

Derivation: Consider a Bernoulli sequence of outputs of the form $(A + B)^*$ generated by a two-state stochastic automaton. There are 2^N possible sequences of N observations and if A and B are equiprobable so are the sequences. Now enumerate the number of different S-state deterministic automata that are available as models of these sequences. We note that a given automaton, starting in a given state, can act as a model of only one of the sequences, and we are concerned to find an upper bound on the variety of models available for a given S. This can be derived by noting that the general form of a model will be a transient chain followed by a cycle. If there are S states to be filled with two symbols and the cycle can commence in any state, there are at most $S2^S$ automata (some of which may generate the same sequence making this an upper bound on the number of different models available).

Since automata of this form with less than S states are also included amongst those having S states, $S2^S$ is an upper bound on the number of different automata with S states or less. Thus to have enough models available we must have

$$S2^S > 2^N. \tag{2}$$

Now consider the mean number of states in the minimal forms of this set of models. There are at most $R2^R$ models with R states, and hence a lower bound of the mean number of states in the set whose maximum number of states is S is given by

$$\mu_S > \frac{1}{(S-1)2^S} \sum_{R=1}^{S-1} R(R2^R - (R-1)2^{R-1}) \tag{3}$$

$$= S - 2 + \frac{2}{S-1} - \frac{2}{(S-1)2^{S-1}} \tag{4}$$

$$> S - 2. \tag{5}$$

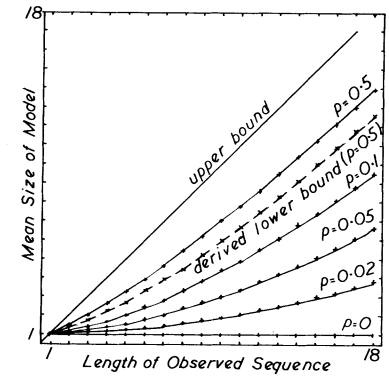

Fig. 1. Expected number of states in optimal causal model against number of observations for binary Bernoulli sequences of different generating probabilities.

Also, by taking logarithms to base 2 of inequality (2) we have

$$S + \log_2 S > N \tag{6}$$

so that

$$S > N - \log_2 S \tag{7}$$

but we have $S < N$ so that

$$S > N - \log_2 N \tag{8}$$

hence from (5)

$$\mu_S > S - 2 > N - \log_2 N - 2 \tag{9}$$

that is, the mean number of states in the ensemble of models necessary to account for all possible behaviors is greater than $N - \log_2 N - 2$, where N is the number of observations.

Note that the ratio of (1), of states to observations, is

$$R_N = 1 - \frac{(\log_2 N + 2)}{N} \tag{10}$$

which is asymptotic to unity as N increases.

III. Computational Results

The theoretical bound obtained could be sharpened, but it clearly indicates the closeness between the expected number of states and the maximum number. Sequences requiring complex models are the common ones in the example used in the derivation of Property 5. It is of interest to see how closely the derived lower bound follows the actual mean number of states in an optimal causal model. Minimal-state models were computed for each of the set of 2^N binary sequences for N from 1 through 18. The case analyzed theoretically, of a Bernoulli sequence with generating probability $p = 0.5$, is intuitively a worst case since it has no "structure." To examine the effect of a bias towards one symbol on the expected size of model the effect of varying the generating probability was also investigated.

Fig. 1 is a set of graphs of the expected number of states in an optimal causal model against the number of observations for binary Bernoulli sequences with generating probabilities ranging from 0.0 to 0.5. The maximal upper bound at which the number of states equals the length of the observed sequence is shown as a diagonal line. The dashed line is derived from (4) and (6) of Theorem 5 and represents the best bound derivable from the theoretical argument.

When the generating probability p is zero the sequence is determinate and always represented by a one-state model. However, the introduction of the second symbol with a probability

of one in fifty ($p = 0.02$) gives rise to a steady growth in the size of a model with the number of observations. It is clear that the rate of growth of states for nonzero nonunity p must ultimately become at least 1/2 since the only sequences of R A's and $N - R$ B's giving an average number of states less than $N/2$ are those with $R = 0$ and $R = N$. The probabilities of the occurrence of these sequences are p^N and $(1 - p)^N$, respectively, both of which decline to zero with increasing N unless p is zero or unity. Thus the slightest introduction of probabilistic acausality gives rise to models whose complexity is proportional to the length of observation.

IV. Conclusions

The results of Sections II and III show that the assumption of causality when modeling acausal systems can lead to indefinitely complex models of comparatively simple systems. It is not just that perfect prediction has become impossible, but rather, that not accepting this leads to essentially meaningless "models" that cannot reflect any stochastic structure present. The example used in the derivation of Property 5 is precisely the stochastic automaton used in some of the experiments with human subjects described in Section I. In Einstein's terms, if one assumes that God does not play dice, then when he does, one will obtain an over-complex view of the universe.

Acknowledgment

I am grateful to Ian Witten of this Department for his critical comments on the first draft of this correspondence, and to Dr. John Andreae of the University of Christchurch, New Zealand, for stimulating this line of argument.

References

[1] K. R. Popper, *Objective Knowledge*. Oxford: Clarendon, 1972.
[2] J. M. Jauch, "Determinism in classical and quantal physics," *Dialectica*, vol. 27, pp. 13–26, 1973.
[3] A. Michotte, *The Perception of Causality*. London: Methuen, 1963.
[4] B. R. Gaines and J. H. Andreae, "A learning machine in the context of the general control problem," in *Automatic Remote Control*, V. Broida, Ed. Proc. 2nd IFAC Congr. London: Butterworths, 1966, pp. 503–600.
[5] J. H. Andreae and P. M. Cashin, "A learning machine with monologue," *Int. J. Man–Machine Stud.*, vol. 1, pp. 1–20, 1969.
[6] J. H. Andreae, "Intelligent identification of signals from an unknown source," in *Proc. Nat. Electron. Conf.*, 1973.
[7] B. R. Gaines, "Memory minimisation in control with stochastic automata," *Electron. Lett.*, vol. 7, pp. 710–711, Nov. 1971.
[8] P. Martin-Lof, "The definition of random sequences," *Inform. Contr.*, vol. 9, pp. 602–619, 1966.
[9] A. N. Kolmogorov, "Logical basis for information theory and probability," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 662–664, Sept. 1968.
[10] G. Chaitin, "On the length of programs for computing finite binary sequences," *J. Assoc. Comput. Mach.*, vol. 16, pp. 145–159, 1969.
[11] D. Willis, "Computational complexity and probability constructions," *J. Assoc. Comput. Mach.*, vol. 17, pp. 241–259, 1970.

A Counter-Example in Linear Feature Selection Theory

D. R. BROWN AND M. J. O'MALLEY

***Abstract*—A family of examples is constructed to show that there may exist distinct $k \times n$ matrices, each of which maximizes transformed divergence which are not row-equivalent.**

Manuscript received March 12, 1975; revised July 2, 1975. This work was supported in part by NASA under Contract JSC-NAS-9-12777.
The authors are with the Department of Mathematics, University of Houston, Houston, TX 77004.

Introduction

The linear feature selection problem in multiclass pattern recognition can be regarded as that of linearly transforming statistical information from n-dimensional (real Euclidean) space into k-dimensional space, while requiring that average interclass divergence in the transformed space decrease as little as possible.

Divergence, as used in this paper, will be the expected interclass divergence derived from Hajek two-class divergence as defined, for example, in [4]. It is known [3] that there always exists a $k \times n$ matrix B such that the transformation determined by B maximizes the divergence in k-dimensional space. It is also known [3] that if Q is any $k \times k$ invertible matrix and B is as defined above, then QB again maximizes the divergence in k-space. The purpose of this correspondence is to show that the converse of this result is false; specifically, we shall show the existence of two matrices B_1 and B_2 each of which maximizes transformed divergence, which are not related in the fashion $B_2 = QB_1$ for any $k \times k$ matrix Q.

The negative resolution of this rather long standing conjecture is unfortunate from the computational standpoint, since derivation of matrices B that maximize transformed divergence is relatively inefficient. Several researchers have addressed the problem of obtaining such B ([1], [2], and [6]), but the latest and most efficient treatment known to us is [3]. A common error in examining special cases of the problem [2] is the incorrect assumption of equality between matrices of the forms

$$\sum_{i=1}^{m} (B\Omega_i B^T)^{-1}$$

and

$$B\left(\sum_{i=1}^{m} \Omega_i^{-1}\right) B^T.$$

Simple examples (see [5], for instance) show this to be false, even if all Ω_i are diagonal matrices.

In the sequel, we avoid this pitfall while computing "the best B" and assure the maximality of transformed divergence by selecting covariance matrices and the means for which it can be shown that divergence in the transformed space equals divergence in the original space. Since divergence is a monotone function of dimension [4], this is sufficient to establish maximality. While the choices of values are made with an eye toward computational simplicity, and are therefore subject to the charge of impracticality, it should be noted that the existence of inequivalent solutions in this restricted case casts doubt that there will ever arise a situation, however practical, in which only a single equivalent class of solutions may be assumed *a priori*.

I. Necessary Divergence Formulas

Let $\Omega_1, \cdots, \Omega_m$ and $\mu_1, \cdots, \mu_m$ be the covariance matrices and means for m classes, where for each $i = 1, \cdots, m$, Ω_i is an $n \times n$ positive definite matrix and μ_i is a column n-vector. Let

$$S_i = \sum_{j=1}^{m} (\Omega_j + \delta_{ij}\delta_{ij}^T)$$

where $\delta_{ij} = \mu_i - \mu_j$. Then, assuming equal *a priori* probabilities, the average interclass divergence for these m classes is given by

$$D = \tfrac{1}{2} \operatorname{tr}\left(\sum_{i=1}^{m} \Omega_i^{-1} S_i\right) - \tfrac{1}{2} m(m-1)n \qquad (1)$$