Int. J. General Systems
1978, Vol. 4, pp. 255–264

ON THE CONNECTION BETWEEN THE COMPLEXITY AND CREDIBILITY OF INFERRED MODELS

JUDEA PEARL
School of Engineering and Applied Science, University of California, Los Angeles, U.S.A.†

(Received June 15, 1977; in final form October 21, 1977)

The connection between the simplicity of scientific theories and the credence attributed to their predictions seems to permeate the practice of scientific discovery. When a scientist succeeds in explaining a set of n observations using a model M of complexity c then it is generally believed that the likelihood of finding another explanatory model with similar complexity but leading to opposite predictions decreases with increasing n and decreasing c. This paper derives formal relationships between n, c and the probability of ambiguous predictions by examining three modeling languages under binary classification tasks: perceptrons, Boolean formulae, and Boolean networks. Bounds are also derived for the probability of error associated with the policy of accepting only models of complexity not exceeding c. Human tendency to regard the simpler as the more trustworthy is given a qualified justification.

INDEX TERMS Inductive inference, complexity, credibility, error probability, discriminating capacity, ambiguous generalization, simplicity, modeling, theory formation, confirmation.

1 INTRODUCTION

The subject matter under discussion can hardly be introduced in a more concise fashion than quoting Quine:[1]

"It is not to be wondered that theory makers seek simplicity. When two theories are equally defensible on other counts, certainly the simpler of the two is to be preferred on the score of both beauty and convenience. But what is remarkable is that the simpler of two theories is generally regarded not only as the more desirable but also as the more probable. If two theories conform equally to past observations, the simpler of the two is seen as standing the better chance of confirmation in future observations. Such is the maxim of the simplicity of nature. It seems to be implicitly assumed in every extrapolation and interpolation, every drawing of a smooth curve through plotted points. And the maxim of the uniformity of nature is of a piece with it, uniformity being a species of simplicity."

Aside of the philosophical interest raised by the phenomena above, it has assumed an increasing practical importance. Much of today's data is being processed by electronic computers and an increasing part of the modelling activity is being delegated to mechanical procedures. In order for an automatic device to satisfactorily manage the generation and selection of competing hypotheses, the programmer-user can no longer hide his inductive procedures and preference criteria in the realm of intuition, but ought to explicate those in a formal, mechanizable way. The criteria for hypothesis selection carry even a greater significance in the area of Robotics. The operation of an industrial robot involves a continuous generation and selection of "explanations", or microtheories, for all sorts of non-anticipated inputs. The criteria for selecting among such competing explanations, their credibility and complexity, would significantly affect the robot performance in its industrial environment.

The philosopher who attempts to explain our natural compulsion to regard the simpler as the more truthful (e.g., as evidenced by the decisive role simplicity had in shaping the historical development of science[2]) inevitably finds himself facing a blind alley. Complexity is a concept variable with language while truth refers to something absolute outside the confines of languages. A theory which seems complex in one language would appear simple in another if only one redefines the atomic variables of one language in terms of the derivatives of another. "This being so, how can simplicity carry any peculiar presumption of objective truth?"[1]

The famous paradoxes of induction[3,4] are products of that same disparity. People tend to form theories in line with the particular language they happen to possess, while inductive logic attempts to capture the process of theory formation by language-independent rules. The two will forever

†This work was performed while the author was visiting the Department of Applied Mathematics at the Weizmann Institute of Science, Rehovot, Israel. The work was supported in part by the National Science Foundation, under Grant MCS75-18734 and MCS74-12208 A01.

remain incompatible and, similarly, one must resign to the idea that no logical argument can possibly connect simplicity with credibility.

But even assuming that the association between the simple and the truthful is purely psychological, one should still be justified in exploring the origin of such perceptual illusion. Apparently, by a long process of evolution our race has learned to associate the simple with the trustworthy. The two must, therefore, possess some common qualities which make them seem to occur conjunctively.

Certainly, part of the answer lies in what Quine[1] calls "subjective selectivity that makes us tend to see the simple and miss the complex." Another factor lies with the flexibility of our language; when a theory becomes workable we "force" it to become simple. When the need arises, we invent new concepts (e.g., ellipses, electrons, wave-functions) which get "entrenched" in our language as elementary entities, in terms of which our theories appear simpler. This phenomena, though, may account for only part of the answer since (as is demonstrated in Section 3) there is a definite limit to the process of simplification by intermediate variables. A point must eventually be reached where the added complexity associated with defining any new variable would overshadow the simplicity it may introduce.

The central property upon which this paper focuses is that of *uniqueness vs. ambiguity*. Simply stated, uniqueness may be exemplified by the fact that through any two points it is possible to pass many second degree polynomials but only one straight line. More generally, there usually are many complex theories which can explain a given set of observations but only a few simple theories (if any). Consequently, if one succeeds in finding a simple explanation to empirical data he is not likely to find another rival explanation, equally simple, which also explains that data. The simpler the theory at hand the lower the likelihood of refuting it with another theory of equivalent complexity. Likewise, we expect the likelihood of committing a prediction error on account of selecting the wrong theory to be lower the simpler the class from which theories are chosen.

In this paper we give these intuitive notions a quantitative formulation, and derive relations between the number of observations, the complexity of models and their credibility. The relations are derived for a binary discrimination task and for three different languages, as defined in Section 2.

Section 3 addresses the question of how many observations one ought to have before becoming fairly certain that any rival theory agreeing with the data must either give the same prediction on the next observation or be more complex than the one at hand. Section 4 analyzes the reliability of probabilistic assertions made by theories as a function of their complexity. We also bound from above the probability of error (on future observations) that any model of a given complexity might possibly make if it agrees with the data at hand.

2 FORMAL NOTATION FOR INDUCTION, MODELS AND LANGUAGES

In this section we give a simple formal description to the elements of inductive reasoning and the role of languages in the search for explanatory models. We imagine a scientist searching for a physical law ro explain a growing body of physical data. Each datum is represented by a pair (x, y), where x stands for the experimental condition and y denotes the experimental result. After collecting n observations the scientist possesses a total evidence $e_n \triangleq \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$, which he attempts to capture with a model f. Let $x \in X$, $y \in Y$ and let F^* stand for the set of all functions $X \to Y$. We assume that the data e_n is generated by an underlying model $f_0 (e_n \subseteq f_0 \in F^*)$ which the scientist wishes to discover. By a scientific method we mean an algorithm A which accepts the evidence e_n and computes a function $f = A(e_n) \in F^*$ which meets some criteria of fitness and complexity.

Fitness criteria measure the extent to which the model agrees with the evidence at hand. It is usually expressed in the form of a distance function,[5] $d[e_n, A(e_n)]$, which is zero whenever $e_n \subseteq A(e_n)$. In Section 3 we shall demand a perfect agreement between evidence and model, that is:

$$e_n \subseteq A(e_n)$$

and will relax it in Section 4.

The search for a model f, as well as the computation of predictions based on f, are usually performed within some linguistic structure which provides a symbolic representation of the space of potential models F^*. It is with respect to such a language that model complexity is usually defined. Let a language L be a pair (T, I) where

T is the set of sentences in the language, and I is its interpreter $I: T \to F^*$. Every sentence $t \in T$ of the language represents a model via its interpretation: $I(t) = f \in F^*$. On each $t \in T$ we define a *complexity* measure: $C: T \to R^+$ which may represent either the syntactic aspect of the sentence t, or the work required for the computation of $I(t)$. Given $C(t)$, we define *model complexity* by:

$$C(f) = \min_{t: I(t)=f} C(t).$$

Thus, the complexity of a model f with respect to a language L is defined as the complexity of the simplest sentence which represents that model.

We now exemplify these notions using three different languages which have frequently been used in Pattern Recognition.

L^1—*Perceptrons*[6]—the data is given as a conjunction of N-dimensional real vector $x = (x^1, x^2, \ldots, x^N)$ and a class label $y \in \{0, 1\}$. The models describable in this language are linear half spaces defined by a list of $N+1$ real coefficients. Thus, each sentence $t \in T$ is an ordered list of $N+1$ real numbers $t = (\omega_1, \omega_2, \ldots, \omega_{N+1})$, and

$$I(t) = \begin{cases} 1 & \text{if } \omega_1 x^1 + \omega_2 x^2 + \ldots \omega_N x^N > \omega_{N+1} \\ 0 & \text{elsewhere} \end{cases}$$

Generalization of L^1 is often more useful, whereby a set of d features from some feature-set $\phi_1(x), \phi_2(x), \ldots$ is first computed and then a linear discrimination is performed in ϕ-space[7]:

$I(t) =$

$$\begin{cases} 1 & \text{if } \omega_1 \phi_1(x) + \omega_2 \phi_2(x) + \ldots \omega_d \phi_d(x) > \omega_{d+1} \\ 0 & \text{elsewhere.} \end{cases}$$

We shall denote this language by L^1_ϕ. The complexity of a sentence in L^1_ϕ is usually equated with the number of features it invokes, that is $C(t) = d$.

L^2 – *Logical formula.* Each data point (x, y) is represented by a Boolean N-vector $x = (x^1, x^2, \ldots, x^N)$, $x^i \in \{0, 1\}$, accompanied by its truth value $y \in \{0, 1\}$. T is the set of Boolean formula on N variables containing negation, conjunction, and disjunction connectives. T can be identified recursively by:

$$T = \begin{cases} x^i \in T & i = 1, 2, \ldots, N \\ t_1, t_2 \in T \Rightarrow t_1, t_1 \cap t_2, t_1 \cup t_2 \in T \end{cases}$$

and $I(t)$ corresponds to the Boolean function represented by t.

Various complexity measures can be defined with respect to L^2. The most common ones are: (1) formula complexity—the number of connectives in t, 2) combinational complexity—the minimum number of gates necessary for a circuit realizing $I(t)$, and (3) time complexity—the minimum time delay in a circuit realizing $I(t)$. Combinational complexity is treated more directly using the next language, L^3. Time complexity is known[8] to be related to combinational complexity in a rather simple manner. We, therefore, take formula complexity to represent the complexity of L^2.

L^3—*Logical formula with intermediate variables.* This language is similar to L^2 with the exception that each sentence may contain several Boolean expressions; the main one defines the model-function while the rest define the variables appearing in the main formula. Each sentence t, therefore, constitutes an explicit blueprint for a logical circuit which computes $I(t)$. The complexity of L^3 will be taken to be the number of connectives in t and it also equals the number of gates in the corresponding circuit. L^2 is sometimes[9] regarded as a subset of L^3 with the restriction of unity fanout. t can also be regarded as a program for computing the Boolean function $I(t)$. The intermediate variables would then represent results of intermediate computations, and $C(t)$ would measure the program execution time.

The three languages, L^1, L^2 and L^3, will next be used as test vehicles to examine the connection between credibility and complexity. Note that while L^2 and L^3 have a full power of expression, as $I(T) = F^*$, this is not the case for L^1; only linear half spaces can be captured by $I(T)$. However, if the feature space is properly chosen (e.g., $\phi_i(x)$ are polynomials of increasing order) every $f \in F^*$ can be approximated at will by an $I(t)$, by taking a large enough d. Moreover, for any finite n we can find a $t \in T$ such that $\bar{e}_n \subseteq I(t)$.

3 CAPACITY AND AMBIGUOUS GENERALIZATION

In the following two sections we imagine a scientist who uses simplicity as a criterion for selecting models in each of the three languages defined above. We wish to find the general laws which govern his performance.

The first question we wish to answer addresses the likelihood of finding a simple model explain-

ing an evidence e_n. Assume that e_n is drawn at random (according to some probability distribution function), what can be said about the probability of finding a model with complexity not exceeding c which explains e_n?

DEFINITION 1 A *complexity bounded sublanguage* of L is a sublanguage $L_c=(T_c, I_c)$ such that $T_c \subseteq T$, $I_c \subseteq I$ and $C(f) \leqq c$ for all $f \in I_c$.

The partition of a language into its simple part L_c and its complex part, $L-L_c$, induces a similar partition on the data space. We shall say that a data e_n is c-simple (denoted by $e_n \in E_c$) if there exists a model $f \in I_c$ such that $e_n \subseteq f$.

DEFINITION 2 The capacity of a complexity-bounded language is given by the number of observations n_c satisfying

$$P(e_{n_c} \in E_c)=\tfrac{1}{2}. \tag{2}$$

This definition of capacity is a slight generalization over the concept originated by Cover[7] for L_ϕ^1. Loosely speaking, capacity measures the maximum number of samples that the scientist should collect if he wishes to guarantee at least 50% chance of finding a c-simple explanation for the data. Clearly, n_c is sensitive to the probability distribution which governs the data generation, and reasonable assumptions must be made before capacity computations can be carried out. For L_ϕ^1, Cover[7] has shown that if $\{x_1, x_2 \ldots x_n\}$ is in ϕ-general position and if the class labels $\{y_1, y_2, \ldots y_n\}$ are chosen at random with equal probability for the 2^n equiprobable possible labeling patterns, then

$$P(e_n \in E_c)=\left(\frac{1}{2}\right)^{n-1} \sum_{k=0}^{C-1} \binom{n-1}{k}. \tag{3}$$

Since, for L_ϕ^1, $P(e_n \in E_c)$ is independent of the exact location of the vectors $\{x_1, x_2 \ldots x_n\}$ (except for the loose requirement that $\{x_1, x_2 \ldots x_n\}$ be in ϕ-general position with probability 1), we can state that (3) holds for any distribution of e_n in which the y's are uniformly and independently distributed.

From (3), it is easy to show that the capacity of L_ϕ^1 is given by

$$n_c=2c-1 \tag{4}$$

and that the probability $P(e_n \in E_c)$ shows a pronounced threshold effect in the neighborhood of $n=2c$. For large c, the addition of each feature function results in capturing an average of two additional samples. Moreover, almost all data can be modeled by L_c if $n<2c$ and hardly any data can be modeled when $n>2c$.

For languages L^2 and L^3, $P(e_n \in E_c)$ is no longer independent on the input $\{x_1, x_2 \ldots x_n\}$, and one must assume a uniform distribution of e_n, in order to calculate the capacity. On the other hand, since both X and L_c are finite we can write

$$P(e_n \in E_c)=\frac{|\{e_n : e_n \in E_c\}|}{|\{e_n\}|}. \tag{5}$$

Denoting the total number of distinct evidences in E_c by $S(n, c)$ and its relative number by $s(n, c)$ we have

$$P(e_n \in E_c)=\frac{S(n, c)}{S(n, \infty)}=s(n, c). \tag{6}$$

The quantity $s(n, c)$ is not easy to compute for either L^2 or L^3, however, asymptotic expressions may be obtained from the literature on the complexity of Boolean functions.

Lower bound: Let n_0 be the highest integer n such that $\forall e_n$ $e_n \in E_c$, then

$$s(n, c) \geqq \begin{cases} 2^{n_0-n} & n \geqq n_0 \\ 1 & n \leqq n_0 \end{cases}. \tag{7}$$

The reason for (7) is that $e_n \in E_c$ implies that either $\{e_n, (x_{n+1}, 0)\}$ or $\{e_n, (x_{n+1}, 1)\}$ (or both) must also be in E_c as an extension of some model f in I_c. Therefore, $s(n, c)$ cannot decrease by a factor smaller than 1/2 for each additional observation.

Upper bound: Each model $f \in I_c$ agrees with exactly $\binom{2^N}{n}$ data sets (corresponding to the $\binom{2^N}{n}$ ways of choosing n out of 2^N possible input combinations, with the truth value determined by f). Therefore, the total number of data sets covered by I_c, $S(n, c)$, is at most (assuming no overlap) $|I_c|\binom{2^N}{n}$, and so

$$s(n, c)=\frac{S(n, c)}{2^n \binom{2^N}{n}} \leqq |I_c| 2^{-n}=2^{n_1-n} \tag{8}$$

where

$$n_1 \triangleq \log_2 |I_c|. \tag{9}$$

For L^2 and L^3 n_1 can be upper bounded by:[10]

$$n_1 \leqq (c+1)(4+\log_2 N) \quad \text{for} \quad L^2, \tag{10}$$

and

$$n_1 \leqq c[4-\log_2 c+2\log_2(N+c)] \quad \text{for} \quad L^3. \tag{11}$$

Since $s(n,c)$ is bounded between two exponential functions, the capacity is likewise bounded by

$$n_0+1 \leqq n_c \leqq n_1+1. \tag{12}$$

It is clearly the proximity between n_0 and n_1 that determines our ability to compute the capacity. Fortunately, the analyses of Sholomov[11] and Pippinger[10] show that n_0/n_1 approaches 1 asymptotically as $N\to\infty$.

Consider the set e_n of all partial Boolean functions of N variables specified on n points. Sholomov[10] has shown that every element of $\{e_n\}$ can be realized by a circuit of complexity not exceeding

$$c=\frac{n}{N}\left[1+0\left(\frac{\log N}{N}\right)\right] \tag{13}$$

if n has a larger order of growth than $N\cdot\log N\cdot\log\log N\cdot\log\log\log N$. Hence, we have

$$n_0=Nc\left[1+0\left(\frac{\log N}{N}\right)\right]. \tag{14}$$

At the same time (11) implies that, for $N<c<2^N$, n_1 is bounded by

$$n_1 \leqq Nc(1+6/N) \tag{15}$$

and so, using (12), the asymptotic capacity of L^3 becomes

$$n_c=Nc[1+0\left(\frac{\log N}{N}\right)]. \tag{16}$$

Thus, allowing the complexity of L_c^3 to increase by one unit (one binary gate) would increase the length of the observation sequences by N observations before models of higher complexity are likely to be needed. Likewise, models of complexity not exceeding n/N should be sufficient to capture about 50% of all observation sequences of length n.

In a similar way one can arrive at the capacity of L^2. Here, a recent result by Pippenger[10] would be necessary, stating that for L^2 all members of e_n would be captured by a formula of complexity not exceeding

$$c=\frac{n}{\log N}\left[1+0\left(\frac{\log\log N}{\log N}\right)\right] \tag{17}$$

and therefore

$$n_0=c\log N\left[1+0\left(\frac{\log\log N}{\log N}\right)\right]. \tag{18}$$

This, coupled with (10) and (16), yields the asymptotic capacity of L_c^2:

$$n_c=c\log N\left[1+0\left(\frac{\log\log N}{\log N}\right)\right]. \tag{19}$$

Several points should be noted in comparing L^2 with L^3. The complexity of a logical circuit with unrestricted fanout would, in most cases, be about $\log N/N$ times lower than an equivalent circuit with fanout one. Equivalently, programs for evaluating logical expressions would be about $\log N/N$ times shorter if the use of intermediate variables is allowed. From these statements one may get an idea of the degree of simplification expected as a result of enriching the language with new "entrenched" predicates.

The capacity of a language is closely related to another measure of performance introduced by Cover—*Probability of Ambiguous Generalization.* Imagine a scientist who succeeds in finding $f\in I_c$ to fit the data e_n. What is the probability that another model exists, $f_1\in I_c$, which also agrees with the past data but which contradicts f on the next sample to be observed? Intuitively, if $n\gg n_c$ then most data can be fitted by only one model in I_c and therefore the probability of ambiguity should be low. Likewise, for $n\ll n_c$ most data can be fitted by more than one model in I_c and so the probability of ambiguity ought to be high.

DEFINITION x_{n+1} *is said to be ambiguous* with respect to evidence e_n in I_c iff both $\{e_n, (x_{n+1}, 0)\}$ and $\{e_n, (x_{n+1}, 1)\}$ are in E_c.

DEFINITION Given a probability distribution on $\{e_n\}$ and $\{e_{n+1}\}$ we define the *probability of ambiguity* $P_a(n,c)$ as the probability that x_{n+1} is ambiguous with respect to a random evidence e_n in I_c.

The language L_ϕ^1 possesses a symmetry property which facilitates a ready calculation of $P_a(n,c)$. Here, each x_{n+1} is ambiguous with respect to a fixed number of ϕ-separable dichotomies of $\{x_1, x_2\ldots x_n\}$ regardless of the location of $\{x_1, x_2\ldots x_n, x_{n+1}\}$ (as long as it is in ϕ-general position). Based on this property, Cover[7] showed that if each ϕ-separable dichotomy of $\{x_1,$

$x_2 \ldots x_n\}$ has equal probability then $P_a(n,c)$ is given by

$$P_a(n,c) = \frac{\sum_{k=0}^{c-2} \binom{n-1}{k}}{\sum_{k=0}^{c-1} \binom{n-1}{k}} \tag{20}$$

and

$$\lim_{\substack{c \to \infty \\ c/n = \text{const.}}} P_a(n,c) = \begin{cases} 1 & 0 \leqq \frac{n}{c} \leqq 2 \\ \frac{1}{\frac{n}{c}-1} & \frac{n}{c} \geqq 2 \end{cases} \tag{21}$$

Thus, as long as the number of observations is below the capacity $2c$, the probability of ambiguity remains unity. For a higher number of observations, P_a decreases at a rate inversely proportional to n.

For L^2 and L^3 the number of dichotomies of $\{x_1, x_2 \ldots x_n\}$ with respect to which a given point x_{n+1} is ambiguous usually varies with $\{x_1, x_2 \ldots x_n\}$ and x_{n+1}. A separate analysis is therefore needed, to express $P_a(n,c)$ in terms of $s(n,c)$. Consider the set of all distinct ordered pairs (e_n, x_{n+1}) for which $e_n \in E_c$, and assume all such pairs to be equally probable. Let a total of c_1 such data-pairs be ambiguous and c_2 of them non-ambiguous. Clearly,

$$P_a(n,c) = \frac{\text{number of ambiguous pairs } (e_n, x_{n+1})}{\text{total number of pairs } (e_n, x_{n+1}) : e_n \in E_c} = \frac{c_1}{c_1 + c_2}. \tag{22}$$

Each ambiguous pair corresponds to two labelled pairs (y_{n+1} specified) which are in E_c, while each non-ambiguous pair corresponds to only one such labelled pair. Also, each data set e_{n+1} appears exactly $n+1$ times in the set of $2c_1 + c_2$ ordered pairs $(e_n, (x_{n+1}, y_{n+1}))$. Therefore,

$$2c_1 + c_2 = (n+1)\, S(n+1,c). \tag{23}$$

At the same time each of the $S(n,c)$ members of $\{e_n\}$ gives rise to $2^N - n$ ordered pairs (e_n, x_{n+1}), and we can write

$$c_1 + c_2 = (2^N - n)\, S(n,c). \tag{24}$$

Combining (22), (23), (24) and (6), we obtain

$$P_a(n,c) = \frac{(n+1)\, S(n+1,c)}{(2^N - n)\, S(n,c)} - 1 = 2\frac{s(n+1,c)}{s(n,c)} - 1. \tag{25}$$

For small sample size, $n < n_0$, $s(n,c)$ is equal to unity and

$$P_a(n,c) = 1 \quad \text{for} \quad n < n_0(c) - 1. \tag{26}$$

For sample sizes exceeding the language capacity a more detailed behavior of $s(n,c)$ is needed before the rate of decrease of $P_a(n,c)$ can be determined. An exponentially decaying $s(n,c)$, for example, would yield $P_a(n,c) = 0$. Had the exponential bounds of (7) and (8) been sufficiently tight one would expect to find a sharp drop in P_a for $n > n_1$. However, the asymptotic results of Sholomov and Pippenger only guarantee

$$\lim_{n \to \infty} \frac{n_1 - n_0}{n_0} = 0$$

not the vanishing of the absolute difference $n_1 - n_0$. Consequently, the exact behavior of $P_a(n,c)$ for $n > n_c$ remains an open question for L^2 and L^3.

Several features of $P_a(n,c)$, however, can be determined directly from the upper bound of (8). A simple analysis of (25), (7) and (8) reveals that $\log[1 + P_a(n,c)]$ must be bounded by:

$$\sum_{n=n_0}^{2^N - 1} \log_2[P_a(n,c) + 1] \leqq n_1 - n_0. \tag{27}$$

On the other hand (14) and (15) imply that $n_1 - n_0$ must be of order at most $c \log N$, and hence (using $\log_2(1+P) \geqq P$) $P_a(n,c)$ should satisfy:

$$\sum_{n=n_0}^{2^N - 1} P_a(n,c) \leqq c\, 0(\log N) \quad \text{for} \quad L^3 \tag{28}$$

and

$$\sum_{n=n_0}^{2^N - 1} P_a(n,c) \leqq c\, 0(\log \log N) \quad \text{for} \quad L^2. \tag{29}$$

The languages L^2 and L^3 exhibit faster decay rates for $P_a(n,c)$ than L^1. An inverse law relation such as the one found for L^1 in (21) would render the left hand sides of (28) and (29) of order N, thus violating the inequalities. A stronger rate of fall,

e.g. an inverse square law, is needed to satisfy (28) and (29).

It is not to be wondered that finite languages such as L^2 and L^3, exhibit a sharper cutoff for ambiguity than infinite languages employing real parameters such as L^1. Clearly, when one exhausts exploring all input combinations (e.g. $n=2^N$) the model is fully specified and no more ambiguity exists. What is significant, though, is that the point of diminishing ambiguity is reached much earlier, at the neighborhood of $n=n_c$, and the threshold in this neighborhood is more pronounced for L^2 and L^3 than L^1. The latter is a consequence of the tightness of the combinatorial bound (8) as expressed in (14) and (18). The significance of a sharper threshold for ambiguity is that for a given complexity bound c, a smaller number of observations is needed in order to achieve a certain level of credibility in the model at hand.

4 COMPLEXITY AND PROBABILITY OF ERROR

Whereas $P_a(n,c)$ may, in many cases, constitute an adequate measure of model credibility, it is a rather loose measure. To compute $P_a(n,c)$ we assumed that all $e_n \in E_c$ are equiprobable and excluded $e_n \notin E_c$. We now wish to extend the credibility measure in three directions. (1) We wish to include considerations of evidence-data not capturable by I_c, $e_n \notin E_c$, like those generated by either more complex models or by nondeterministic processes. (2) We wish to perform a "worst case" analysis assuming that Nature herself, in what might be regarded as a "hostile" manner, may select the observation sequence in accordance with some fixed distribution law. Indeed, it is rather unrealistic to assume equiprobable observation sequences for the mere fact that some experimental conditions are harder to satisfy than others. (3) We wish to define credibility not merely in terms of the number of competing models but rather directly in terms of the degree of agreement between the true underlying model and the one at hand.

Consider a scientist with a complexity bounded language L_c observing data e_n, and attempting to fit it with a theory $A(e_n) \in I_c$. Since e_n may be generated by a model $f_0 \notin I_c$ (or by a nondeterministic model) we must give up the requirement of perfect fit, and instead assume that the scientist only attempts to posit a theory which reasonably approximates the data (e.g. that which minimizes the number of mistakes: $(x_i, y_i) \notin f$), and report the degree of approximation. This scheme closely reflects Reichenbach's[12] concept of induction whereby the aim of science is viewed not as that of discovering true theories but of positing probabilistic assertions about nature with an ever increasing accuracy.

Suppose the scientist reports that a model $f \in I_c$ approximates an evidence e_n and that it disagrees with a fraction $\nu_f(e_n)$ of the n observed samples. Denoting by Π_f the true probability of disagreement (according to the underlying distribution which governs the data generation), we first wish to bound the probability of disparity $P(|\Pi_f - \nu_f| \geqq \varepsilon)$ as a function of ε, n, and the complexity bound c. It is intuitively believed that the simpler the model f the closer would ν_f be to Π_f, i.e., one can often find complex models for which $\nu_f = 0$, and which stand in no relation to Π_f.

If the samples (x_i, y_i) were drawn independently of each other, and if f were kept constant throughout the observation sequence one could then invoke Bernoulli's theorem[13] and write

$$P(|\Pi_f - \nu_f| \geqq \varepsilon) \leqq 2e^{-n\varepsilon^2/2} \tag{30}$$

This theorem is indeed the basis of Reichenbach's "vindication" of induction, demonstrating that as long as an underlying probability Π_f exists the probability that the reported frequency ν_f deviates from Π_f by any finite amount decreases exponentially with the number of observations. Unfortunately, the assumption of fixed f misses the most significant aspect of scientific activity. Scientists continuously modify their theories as experiments progress. In fact, the act of inventing a new theory to fit an existing data has, traditionally, been given much greater esteem than the painstaking effort of measuring ν_f for a fixed hypothesis. Fortunately, a recent work of Vapnik and Chervonenkis[14] permits the bounding of $P(|\Pi_f - \nu_f| \geqq \varepsilon)$ even under conditions of data fitting. Vapnik and Chervonenkis theorem, which can be termed "the Bernoulli theorem for the hindsighted scientist", will be briefly stated using their terminolology:

THEOREM *Let S be a collection of subsets of a space X on which a probability measure P_x is defined. Each sample $x_1, \ldots, x_l$ and event $A \in S$ determine a relative frequency for A equal to the*

quotient of the number n_A of those elements of the sample which belong to A and the total size l of the sample: $v_A^{(l)}(x_1,\ldots,x_l)=n_A/l$. If the samples are drawn independently then the probability that at least one event in S differs from its probability P_A by more than ε, for $l>2\varepsilon^2$, satisfies

$$P\left[\sup_{A\in S}|P_A-V_A|\geqq\varepsilon\right]\leqq 4m^s(2l)e^{-\varepsilon^2 l/8} \tag{31}$$

where $m^s(l)$ is the maximum over $(x_1, x_2,\ldots,x_l)$ of the number of distinct sets in $\{\{x_1, x_2,\ldots,x_l\}\cap A: A\in S\}$. In other words, $m^s(l)$ is the maximum number of ways that any sample of size l can be dichotomized by the elements of S.

In order to use (31) for bounding $P(|\Pi_f-\nu_f|>\varepsilon)$ we simply replace X with the space of sample pairs $X\times Y$ and identify S with I_c. $m^s(n)$ would then measure the maximum over e_n of the number of distinct dichotomies (agree vs. disagree) of e_n induced by f as it spans I_c. For a I_c we have

$$m^{I_c}(n)\leqq|I_c| \tag{32}$$

because every distinct dichotomy of e_n must be induced by a different $f\in I_c$. Moreover, for $n\leqq n_0$, all dichotomies can be matched by some $f\in I_c$, hence

$$m^{I_c}_{(n)}\leqq\begin{cases}2^n & n\leqq n_0\\ 2^{n_1} & n\geqq n_0\end{cases} \tag{33}$$

and

$$P(|\Pi_f-\nu_f|\geqq\varepsilon)\leqq\begin{cases}1 & n\leqq 8\,ln2[n_1(c)+2]/\varepsilon^2\\ 4e^{n_1(c)ln2-\varepsilon^2 n/8} & n\geqq 8\,ln2[n_1(c)+2]/\varepsilon^2\end{cases} \tag{34}$$

Equation (34) exhibits a sharp threshold effect; the bound on P remains unity up to about $8\,ln2/\epsilon^2$ times the language capacity, from which point on it decays exponentially with n. In an analogy paralleling the classical Cantellis theorem,[13] one may ask what sample size n would guarantee that $P(|\Pi-\nu|\geqq\varepsilon]$ would remain below some given level η for all succeeding observations. The answer is given by

$$n\geqq 2+\frac{8}{\varepsilon^2}\left\{\log 8/\varepsilon^2\eta+[n_1(c)+2]\,ln2\right\}. \tag{35}$$

Thus, for L^3 and large c, the addition of one gate to the model would necessitate roughly $8ln2/\varepsilon^2\log_2 c$ additional observations in order to maintain the same level of η, (see Eq. 11). For L^2 an addition of one connective to the model formula would require a uniform increase of $8ln2/\varepsilon^2\ (4+\log_2 N)$ observations.

For L^1_ϕ,

$$m^{I_c}(n)=2\sum_{k=0}^{c}\binom{n}{k}<2n^c \tag{36}$$

and so (31) becomes

$$P(|\Pi_f-\nu_f|\geqq\varepsilon)\leqq 8\,(2^c n^c)e^{-\varepsilon^2 n/8}. \tag{37}$$

Equation (37) is similar to the one used by Devroye and Wagner[15] to obtain performance bounds in error estimation for linear discrimination procedures. Note that the exponential drop is somewhat slowed down by the polynomial $(2n)^c$, and so, one should expect that more observations would be needed to maintain P at a certain level n. The exact expression determining n is:

$$n\geqq\frac{16}{\varepsilon^2}\left(c\log\frac{16c}{\varepsilon^2}-\log\eta/8\right). \tag{38}$$

The use of each additional feature would necessitate roughly

$$\frac{16}{\varepsilon^2}\log\frac{16c}{\varepsilon^2}$$

additional samples (for large c).

It is important to note that (34) and (37) hold for any f in I_c regardless of the method used by the scientists to discover f. The convergence of (34) and (37) for large n is a product of the limited expressional power of the languages considered. The lower $m^s(2n)$ the less flexible is the scientist to tailor his model around the data and the higher the reliability of the reported ν_f.

At this point one may consider the case of c varying with n. That reflects the natural phenomena that scientific terminology tends to become more and more complex as more data is collected. We may ask how fast can $c(n)$ be allowed to increase with n before the convergence of ν_f to Π_f is endangered. The answer can be obtained directly from (31) and (37), showing that the conditions:

$$\begin{cases}\lim_{n\to\infty}\dfrac{n_1[c(n)}{n}=0 & \text{for } L^2 \text{ and } L^3\\ c(n)=\sigma\left(\dfrac{n}{\log n}\right) & \text{for } L^1\end{cases} \tag{39}$$

would retain the convergence:

$$P(|\Pi_f - \nu_f| > \varepsilon) \underset{n\to\infty}{\to} 0. \qquad \varepsilon > 0 \quad (40)$$

The last concept we wish to explore is the effect of complexity on the probability of error. Assume that we know *a priori* that the underlying model f_0 is in I_c. In this case the scientist can perfectly match every e_n by at least one model in I_c. Assuming the scientist discovers such a perfect fit model f and subscribes to it, what is the probability of errors in future predictions? Since $e_n \subset f$, $|\Pi_f - \nu_f| = \Pi_f$ would represent the error frequency in future predictions. Π_f is a random variable since, in general, f is chosen by some algorithm on the basis of the evidence e_n, which is random. The overall probability of error P_e can be obtained by taking the expectation of Π_f,

$$P_e = E(\Pi_f) = \int_{\varepsilon=0}^{1} P(\Pi_f \geqq \varepsilon)\, d\varepsilon.$$

Since for $\nu = 0$ $P(\Pi_f \geqq \varepsilon)$ is bounded by (31) we obtain (for large n and large c):

$$P_e \leqq \begin{cases} \sqrt{8\ln 2}\ \sqrt{\dfrac{c\log_2 c}{n}} & \text{for } L^3 \\ \sqrt{8\ln 2}\ \sqrt{4+\log_2 N}\ \sqrt{\log_2 c/n} & \text{for } L^2 \\ \sqrt{8c}\ \sqrt{\dfrac{\log n}{n}} & \text{for } L^1 \end{cases} \quad (41)$$

Note the relatively large number of samples required to achieve low error probabilities for all languages, especially L^1. It is not surprising though that (41) exhibits slower drops of P_e than those obtained for $P_a(n, c)$ as (41), unlike (25), represents a worst case analysis for both f and P_x.

5 DISCUSSION

Sections 3 and 4 demonstrate that under rather simple and general descriptions of scientific inference several accepted norms of credibility are correlated with model's simplicity. The exact nature of this relationship though, depends on the language used by the modeller to construct theories with. From a practical viewpoint the analysis reported helps extend the classical notion of statistical confidence level to three commonly used languages with model complexity taking the role traditionally played by the "degrees of freedom" measure. The relations developed in Sections 3 and 4 should enable the modeller to determine the number of observations required for achieving a desired level of credibility for a model of given complexity, in much the same way that statisticians determine confidence intervals for linear regression models.

From a philosophical viewpoint it is essential to note that in all cases examined the role of *simplicity* was only incidental to the analysis. We would have gotten identical results if instead of L_c being a complexity bounded sublanguage we were to substitute an arbitrary sublanguage with equal number of functions. It is not the nature of the functions in I_c but their number $|I_c|$ (more precisely, the number of sample dichotomies induced by the members of I_c) which affects the various plausibility measures considered. As long as the scientist commits himself to a language of limited expressional power his data-fitting maneuverability would be curtailed, and consequently, any theory he may generate that can stand empirical test carries a high degree of credibility even when the language employs some very complex function.

Why, then, do people exhibit a higher trust in simpler theories? When a theory is reported we automatically assume that a certain *procedure* was followed by the scientist prior to its discovery. We assume that prior to discovery the scientist confines his attention to the class of theories with complexity not exceeding the one reported. If such a procedure is indeed adhered to, then the simplicity of the reported theory would reflect the limitation on the scientists maneuverability while trying to fit the data. In this case (and this case only) would a greater simplicity also mean a more falsifiable, more testable and so, more plausible theory?

The illusion that Nature seems to "talk our language" and behave as though She adopts the same complexity scale used by people seems to arise each time we face a phenomenon which depends on the *number of configurations* within a given set. The second law of Thermodynamics, for example, has been interpreted in many textbooks as though Nature exhibits an incurable tendency to disrupt order. Nature, of course, could not prefer one state of affairs to any other simply because we found an elegant description to the former, not more than the sequence HHHHHH is preferred to any other sequence in

a coin-flipping experiment. The second law implies only that a thermodynamic system tends to "escape" from any narrow region of phase space toward regions of larger volume. The illusion of an irreversible trend toward disorder originates with the fact that the volume occupied by states to which people can find concise descriptions (in any language) is extremely small compared with the entire space of possibilities. The escape from the simpler to the more complex is merely a perceptual distortion of the underlying transition from the narrow to the wider, as people fail to record the much more frequent transitions from the complex to the complex.

The credibility of inferred models, like thermodynamical transformations, depends on the cardinality of the space of descriptions. While low cardinality is a necessary quality of the space of simple descriptions the converse is not generally true. The positive correlation between the two may have resulted in our tendency to regard the simpler as the more trustworthy, but cannot be relied upon for testing credibility unless the procedure of theory selection is examined. It would, therefore, be more appropriate to connect credibility with the nature of the selection procedure rather than with properties of its final product. When the former is not explicitly known, as is the case with human communication, simplicity merely serves as a rough indicator for the type of processing that took place prior to discovery.

ACKNOWLEDGMENT

The author acknowledges with thanks the kind hospitality of the Department of Applied Mathematics at The Weizmann Institute of Science, Rehovot, Israel, where this work was performed. The support granted by the National Science Foundation, Division of Computer Research under Grants No. MCS75-18734 and MCS74-12208 A01 made this work possible.

REFERENCES

1. W. V. O. Quine, "On Simple Theories of a Complex World." *Synthese*, **15**, 1963, pp. 103–106.
2. S. Toulmin, *Foresight and Understanding*. Indiana University Press, Bloomington, 1961.
3. C. G. Hempel, "Recent Problems of Induction." In: *Probabilities, Problems and Paradoxes*, edited by S. A. Luckenback, Dickinson Publishing Company, Encino, California, 1972, pp. 161–182.
4. N. Goodman, *Fact, Fiction and Forecast*, 2nd Ed. Bobbs-Merrill, Indianapolis, 1965.
5. J. G. Kemeney, "The Use of Simplicity in Induction." In: *Probability, Confirmation and Simplicity*, edited by Foster and Martin, The Odyssey Press, New York, 1966, pp. 301–321.
6. M. Minsky and S. Papert, *Perceptrons*. MIT Press, Cambridge, Mass., 1969.
7. T. M. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition." *IEEE Transactions on Electronic Computers*, **EC-14**, 1965, pp. 326–334.
8. P. M. Spira, "On Time-Hardware Complexity Tradeoffs for Boolean Functions." *Proceedings of Fourth Hawaii International Conference on System Sciences*, pp. 525–527.
9. J. S. Savage, *The Complexity of Computing*. John Wiley, New York, 1976, p. 26.
10. N. Pippenger, "Information Theory and the Complexity of Boolean Functions." *Mathematical Systems Theory*, **10**, 1977, pp. 124–162. Also: *16th Ann. IEEE Symp. on Found. Comp. Sci.*, Berkeley, 1975, pp. 113–118.
11. L. A. Sholomov, "On Functionals Characterizing the Complexity of a System of Undetermined Boolean Functions." *Systems Theory Research* (Problemy Kibernetiki) **19**, 1970, pp. 123–141.
12. H. Reichenbach, *Experience and Prediction*. The University of Chicago Press, Chicago, 1938.
13. J. V. Uspensky, *Introduction to Mathematical Probability*. McGraw-Hill, New York, 1937, Chapter VI.
14. V. N. Vapnik and A. Ya. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities." In: *Theory of Probability and Its Applications*, **XVI**, 1971, pp. 264–280.
15. L. P. Devroye and T. J. Wagner, "A Distribution-free Performance Bound in Error Estimation." *IEEE Trans. on Information Theory*, **IT-22**, No. 5, September, 1976, pp. 586–588.

Judea Pearl was born in Tel-Aviv, Israel, on 4 September 1936. He received the B.S. degree in electrical engineering from Technion-Israel Institute of Technology, Haifa, Israel, in 1960; the M.S. degree from Newark College of Engineering, Newark, New Jersey, in 1961; the M.S. degree in physics from Rutgers University, New Brunswick, New Jersey; and the Ph.D. degree in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, New York, in 1965.

During 1960–61 he was engaged in medical electronic research at New York University, New York, and taught mathematics at Newark College of Engineering. In 1961 he joined RCA Laboratories, Princeton, New Jersey, where he conducted research on micromagnetic memories, thin-film transistors, low-noise electron tubes and super-conductive parametric and storage devices. In 1965 he performed an experiment which first proved the existence of the Magnus force in superconductors, and was a corecipient of the RCA Research Award for the development of a superconductive parametric amplifier. In 1966 he became Director of Advanced Memory Devices at Electronic Memories, Inc., Hawthorne, California, heading the development of plated wire memories. In 1969 he joined the school of Engineering and Applied Science, University of California, Los Angeles, where he is a Professor of Engineering. His present interests lie in signal processing, pattern recognition, artificial intelligence and decision theory. He has published over 30 technical papers in his fields of interest.

Dr. Pearl is a member of the Association of Computing Machinery (SIGART) and the Institute of Electrical and Electronic Engineering.