Inferring Probabilistic Theories from Data

Edwin P.D. Pednault
Knowledge Systems Research Department
AT&T Bell Laboratories
Holmdel, NJ 07733

ABSTRACT

When formulating a theory based on observations influenced by noise or other sources of uncertainty, it becomes necessary to decide whether the proposed theory agrees with the data "well enough." This paper presents a criterion for making this judgement. The criterion is based on a gambling scenario involving an infinite sequence of observations. In addition, a rule derived from the idea of minimal-length representations is presented for selecting an appropriate theory based on a finite set of observations. A proof is briefly outlined demonstrating that the theories selected by the rule obey the success criterion given a sufficient number of observations.

1. INTRODUCTION

Much of the work in inductive inference has considered the problem of formulating deterministic theories from error-free observations (e.g., see review articles by Angluin and Smith [1983], Dietterich and Michalski [1983], and Kearns *et al.* [1987]). However, the real world often presents us with data influenced by noise or other sources of uncertainty, or with situations for which a deterministic model is inappropriate. In the deterministic case, any theory that does not absolutely agree with the observations can be ruled out. In the presence of uncertainty, on the other hand, one must consider the degree to which a theory accounts for the observations. This complicates the inductive inference problem, since one cannot simply choose the theory that best fits the data. For a finite set of data, it is possible to select a theory that fits "too well." An example would be selecting a polynomial of high enough degree so that it passes through every point in a set of data points. This amounts to fitting the theory to the noise rather than to the underlying relationships, thereby producing a rather poor model of the data [e.g., Tukey 1977]. If we extend the selection problem to an infinite set of data, an exact fit is impossible, since otherwise the model would be deterministic. In that case, no matter what theory we might propose, there exists another that more closely agrees with the observations. The problem is to judge when the fit is "good enough" and an appropriate theory has been obtained.

This paper considers the problem of devising a criterion for judging whether a proposed theory is an appropriate model for a set of data. The analysis assumes that one is dealing with theories for predicting future events based on past observations, and that the "best" theory is the one with the greatest predictive power. A gambling scenario is used to measure predictive power. Given a predictive theory, one can imagine using it to place bets on future events. If the predictions are accurate, the bets will be won and money will be made. The more accurate the predictions, the greater the return. The relative predictive power of two or more theories can therefore be assessed by comparing the amounts that each wins. The theory that wins the most money *in the long run and to within a constant factor* is deemed to have a suitable level of predictive power and, hence, is an appropriate model for the observations. The qualification of considering the long term is important, since greater predictive power implies less hedging of bets. By comparing the long-term winnings, we avoid the possibility of highly unlikely events from eliminating a theory with greater predictive power. Predictive power is therefore treated as an asymptotic property (i.e., it is measured with respect to an infinite set of observations). The constant factor takes into account the ability to find increasingly better fits to an infinite set of observations.

Complementing this asymptotic analysis, a rule is presented for selecting an appropriate theory based on a finite set of observations. In addition, a proof is briefly outlined demonstrating that the theories thus selected obey the success criterion described above given a sufficient number of observations. The selection rule is based on the minimum description length principle that has been suggested by several authors [Rissanen 1978, 1983; Segen 1980, 1985; Barron and Cover 1983, 1985; Sorkin 1983]. According to the latter, the theory one should select given observations $x_1 \cdots x_n$ is the one that minimizes the following sum:

$$\ell(T) + \ell(x_1 \cdots x_n \mid T) \qquad (1)$$

where $\ell(T)$ is the length in bits of a machine-readable representation of theory T and $\ell(x_1 \cdots x_n \mid T)$ is the number of bits needed to encode the observations with respect to T. The quantity $\ell(T)$ effectively measures the complexity of T, while $\ell(x_1 \cdots x_n \mid T)$ measures the degree to which T accounts for the observations, with fewer bits indicating a better fit. The sum of these two quantities defines the number of bits needed to represent the observations. When minimizing this sum, the $\ell(T)$ term counterbalances the degree-of-fit term to prevent one from selecting theories that agree with the available data too closely. Rissanen [1978, 1983] has shown that this selection rule converges when the appropriate theory is a member of a known parametric family of probabilistic models. Barron

[1985] has generalized this result to include stationary, ergodic probabilistic models. Convergence in the general case, however, remains an open problem.

This paper presents a slightly different selection rule for which a general convergence proof has been obtained. The rule is to select theory T if it minimizes

$$\ell(T) + d(x_1 \cdots x_n \,||\, T) \tag{2}$$

where

$$d(x_1 \cdots x_n \,||\, T) \stackrel{\text{def}}{=} \ell(T) + \ell(x_1 \cdots x_n \mid T) - \min_S[\ell(S) + \ell(x_1 \cdots x_n \mid S)] \,. \tag{3}$$

The quantity $d(x_1 \cdots x_n \,||\, T)$ is the number of extra bits needed to represent the observations using theory T as opposed to the theory that yields the minimal representation. It essentially measures the degree to which theory T accounts for the observations *relative to all other theories*. This relative measure avoids certain stumbling blocks encountered when attempting to prove the general convergence of the minimum description-length rule. Due to space limitations, only a brief outline of the convergence proof is presented in this paper.

2. JUDGING PREDICTIVE THEORIES

The gambling scenario for judging the success of a theory assumes that an infinite stream of observations is available in machine-readable form. The stream need only be infinite in the sense that additional observations can always be obtained if so desired. Machine readability is necessary for machine learning.

Bets are made on the binary representation of the observation stream. The bits in the observation stream are revealed one at a time. Bets are placed on each bit immediately before it is revealed. Once revealed, the winners are paid double the amount bet on that outcome.

Without loss of generality, we can assume that each bet consists of a certain amount placed on an outcome of 0 with the rest placed on 1. With 2-to-1 odds, no money need ever be kept aside. Betting an amount a on 0 and an amount b on 1, with an amount c kept aside, is equivalent to betting $(a + \frac{1}{2}c)$ on 0 and $(b + \frac{1}{2}c)$ on 1, with nothing kept aside. In both cases, one is paid an amount $2a + c$ if the outcome is 0, and $2b + c$ if the outcome is 1.

Assuming that no money is kept aside, a betting strategy can be described in terms of a *gambling function*. A gambling function defines the fractional amounts of one's current assets to place on the possible values of the next bit in the observation stream. If p is such a function, then $p(x_1)$ is the fraction to bet on the first bit having the value x_1, while $p(x_n \mid x_1 \cdots x_{n-1})$ is the fraction to bet on the n'th bit having the value x_n given that the first $(n-1)$ bits were $x_1 \cdots x_{n-1}$. Notice that gambling functions are subject to the following constraints:

$$p(x_1) \geq 0, \quad p(0) + p(1) = 1, \quad p(x_n \mid x_1 \cdots x_{n-1}) \geq 0$$
$$p(0 \mid x_1 \cdots x_{n-1}) + p(1 \mid x_1 \cdots x_{n-1}) = 1 \,. \tag{4}$$

For the purposes of machine learning, we must restrict our attention to computable gambling functions. Since gambling functions define real numbers in the interval $[0,1]$, a computable gambling function is one for which a computer program exists that can approximate these real numbers to arbitrary accuracy:

Definition 1. A gambling function p is said to be *computable to arbitrary accuracy* if and only if there is a computer program $\hat{p}$ that takes as input an integer α and a finite binary sequence $x_1 \cdots x_n$ and produces as output a rational number written as $\hat{p}_\alpha(x_n \mid x_1 \cdots x_{n-1})$ such that

$$\left| p(x_n \mid x_1 \cdots x_{n-1}) - \hat{p}_\alpha(x_n \mid x_1 \cdots x_{n-1}) \right| \leq 2^{-\alpha} \,.$$

The integer α corresponds to the number of bits of accuracy to which p is to be computed. Thus,

$$\lim_{\alpha \to \infty} \hat{p}_\alpha(x_n \mid x_1 \cdots x_{n-1}) = p(x_n \mid x_1 \cdots x_{n-1}) \,.$$

As a notational convenience when dealing with a program $\hat{p}$ for estimating a gambling function, we will write $p(x_n \mid x_1 \cdots x_{n-1})$ to mean $\lim_{\alpha \to \infty} \hat{p}_\alpha(x_n \mid x_1 \cdots x_{n-1})$. Also, as a terminological convenience, we will refer to programs for estimating gambling functions as *computable gambling functions*, if this can be done without ambiguity (keeping in mind that a program may define a gambling function, but is not itself one).

The results presented in this paper assume that programs for estimating gambling functions represent theories about the observation stream. The representation may either be direct (i.e., the program *is* the theory) or indirect (i.e., the theory is compiled into a program). In either case, theories are compared in terms of their corresponding gambling functions.

Let $\{b_i\}_{i=1}^{\infty} = b_1 b_2 b_3 \cdots$ be the observed sequence of bits. Using gambling functions, the capital that remains after gambling on the first n bits of this sequence is given by

$$C_n = C_0 2^n p(b_1 \cdots b_n)$$

where C_0 is the amount of initial capital, and where

$$p(b_1 \cdots b_n) = p(b_1)\, p(b_2 \mid b_1) \cdots p(b_n \mid b_1 \cdots b_{n-1}) \,.$$

By convention, we will assume that $C_0 = 1$, so that $C_n = 2^n p(b_1 \cdots b_n)$.

Ideally, we would like to construct a gambling function p^* capable of predicting the observed sequence exactly; that is, p^* should satisfy

$$p^*(b_1) = p^*(b_n \mid b_1 \cdots b_{n-1}) = 1 \tag{5}$$

Such a p^* would maximizes our earnings (i.e., $C_n = 2^n$). However, because we are restricted to computable gambling functions, and because the set of computer programs can be placed into one-to-one correspondence with the set of natural numbers, there are only countably many gambling functions to choose from. On the other hand,

there are uncountably many observation streams $\{b_i\}_{i=1}^{\infty}$, since these sequences can be placed into one-to-one correspondence with the set of real numbers. Consequently, there are observation streams for which no *computable* ideal gambling function exists. In fact, there are uncountably many such streams! For most observation streams we must resort to computable gambling functions that converge to values between 0 and 1.

In those cases in which the ideal gambling function is not computable, we might hope to construct a computable gambling function $\hat{p}^*$ that always wins at least as much money as any other computable gambling function; i.e.,

$$\forall \hat{q} \quad \forall n \geq 1 \quad p^*(b_1 \cdots b_n) \geq q(b_1 \cdots b_n) \; . \qquad (6)$$

However, no such gambling function exists. To see why, first note that for any $\hat{p}^*$ we might propose, it is possible to construct a series of programs $\{\hat{q}^k\}_{k=1}^{\infty}$ such that $\hat{q}^k$ predicts the first k bits of the observation sequence exactly and then places the same bets as $\hat{p}^*$ on all subsequent bits. By construction,

$$p^*(b_1 \cdots b_n) = q^k(b_1 \cdots b_n) \cdot p^*(b_1 \cdots b_k) \qquad \text{for } n \geq k.$$

Furthermore, since we are considering the case in which p^* is not ideal, there must be a value for k such that $p^*(b_1 \cdots b_k) < 1$. For this value of k, it will be the case that

$$\forall n \geq k \quad p^*(b_1 \cdots b_n) < q^k(b_1 \cdots b_n)$$

which contradicts Condition 6. Hence, there is no $\hat{p}^*$ that always wins at least as much money as any other computable gambling function when the ideal gambling function is not computable.

In general, the best we can do is to find a computable gambling function $\hat{p}^*$ that wins at least as much money as any other computable gambling function *to within a constant factor*; i.e.,

$$\forall \hat{q} \quad \exists C_{\hat{q}} > 0 \quad \forall n \geq 1 \quad p^*(b_1 \cdots b_n) \geq C_{\hat{q}} \cdot q(b_1 \cdots b_n) \quad (7)$$

For example, if $\hat{q} = \hat{q}^k$ as defined above, a suitable value for $C_{\hat{q}^k}$ would be $C_{\hat{q}^k} = p(b_1 \cdots b_k)$. The factor $C_{\hat{q}}$ can be interpreted as the initial capital available to bettors using $\hat{q}$. A value of $C_{\hat{q}}$ that is less than 1 thus represents a handicap placed on $\hat{q}$.

Although Condition 7 provides us with a definition of a suitable gambling function, it cannot be used for selecting one. The reason is that the criterion requires knowledge of the complete (i.e., infinite) observation stream. In practice, we have no choice but to converge asymptotically on an appropriate $\hat{p}^*$. As each bit in the observation stream is revealed, a guess must be made as to what $\hat{p}^*$ should be. If $\hat{p}^k$ is the guess made after seeing the first k bits, this process results in a sequence of computable gambling functions $\{\hat{p}^k\}_{k=1}^{\infty}$. To converge asymptotically on an optimal $\hat{p}^*$ (or a set of optimal $\hat{p}^*$'s) is to produce a sequence $\{\hat{p}^k\}_{k=1}^{\infty}$ for which all guesses beyond a certain point in the sequence are optimal gambling functions according to Condition 7. This is analogous to *identification in the limit* and *behaviorally correct identification* in the case of deterministic theories [e.g., see review article by Angluin and Smith 1983]. Notice that Condition 7 provides no guidance as to how to select each $\hat{p}^k$. For example, restricting the criterion to a finite segment of the observation stream does not help, since this produces

$$\forall \hat{q} \quad \exists C_{\hat{q}} > 0 \quad \forall 1 \leq k \leq n \quad p^*(b_1 \cdots b_k) \geq C_{\hat{q}} \cdot q(b_1 \cdots b_k)$$

which is satisfied by all $\hat{p}^*$'s for which $p^*(b_1 \cdots b_k)$ is nonzero. Some other criterion must be employed, such as the selection rules discussed in the introduction.

3. ENCODING OBSERVATIONS

To employ the selection rules discussed in the introduction, it is necessary to devise a way of encoding a sequence of observations relative to a gambling function. This can be done by noticing that gambling functions, as defined by Equation 4, satisfy the definition of a probability mass function for binary sequences. We can therefore employ information-theoretic techniques such as Shannon coding [e.g., Gallager 1968] to encode an observed sequence. Shannon coding minimizes the average length of the codeword, assuming a random binary sequence drawn according to the probability mass function implied by the gambling function. The length in bits of the codeword for an observed sequence $b_1 \cdots b_n$ is given by

$$\left\lceil \log_2 \left(\frac{1}{p(b_1 \cdots b_n)} \right) \right\rceil \; = \; -\lfloor \log_2 p(b_1 \cdots b_n) \rfloor$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x and $\lfloor x \rfloor$ is the largest integer less than or equal to x. Thus, the likeliest sequences have the shortest coding lengths, while the least likely have the longest. The number of bits needed to encode $b_1 \cdots b_n$ using $\hat{p}$ is therefore given by

$$\ell(\hat{p}) - \lfloor \log_2 p(b_1 \cdots b_n) \rfloor \qquad (8)$$

where $\ell(\hat{p})$ is the length (in bits) of program $\hat{p}$. For convenience, all coding lengths will be measured in bits and, hence, all logarithms will be in base two. Equation 8 thus corresponds to Equation 1 given in the introduction. The minimum description length rule is to find the gambling function $\hat{p}$ that minimizes this sum.

To arrive at the modified selection rule discussed in the introduction, $d(x_1 \cdots x_n \parallel \hat{p})$ will be defined as follows:

$$d(x_1 \cdots x_n \parallel \hat{p}) \stackrel{\text{def}}{=} \ell(\hat{p}) - \log p(x_1 \cdots x_n) - \min_{\hat{q}} [\ell(\hat{q}) - \log p(x_1 \cdots x_n)] \; . \qquad (9)$$

$d(x_1 \cdots x_n \parallel \hat{p})$ is essentially the number of extra bits needed to encode $x_1 \cdots x_n$ using $\hat{p}$ instead of the computable gambling function $\hat{q}$ that yields the shortest encoding. The floor brackets are removed for mathematical convenience; consequently, this function actually approximates the number of extra bits to within an error of ± 1.

The function $d(x_1 \cdots x_n \parallel \hat{p})$ has the interesting property that, for any given sequence $\{x_i\}_{i=1}^{\infty}$ and any given

computable gambling function $\hat{p}$, either $d(x_1 \cdots x_n \parallel \hat{p})$ has an upper bound or it increases without bound as $n \to \infty$:

Theorem 1. For any binary sequence $\{x_i\}_{i=1}^{\infty}$ and any computable gambling function $\hat{p}$, either

(1) $\exists \beta \quad \forall n \geq 1 \quad d(x_1 \cdots x_n \parallel \hat{p}) \leq \beta$, or

(2) $\forall \beta \quad \exists N \quad \forall n \geq N \quad d(x_1 \cdots x_n \parallel \hat{p}) > \beta$.

Thus, $d(x_1 \cdots x_n \parallel \hat{p})$ cannot become arbitrarily large and then arbitrarily small again (i.e., behavior as exhibited by $|n \sin n|$ is excluded). The proof centers upon a demonstration that the value of $d(x_1 \cdots x_n \parallel \hat{p})$ establishes a lower bound on $d(x_1 \cdots x_{n+m} \parallel \hat{p})$ for $m \geq 1$. This lower bound is a monotonically increasing function of $d(x_1 \cdots x_n \parallel \hat{p})$, which implies that either $d(x_1 \cdots x_n \parallel \hat{p})$ has an upper bound or it increases without bound.

Another interesting property of $d(x_1 \cdots x_n \parallel \hat{p})$ is that, if $d(x_1 \cdots x_n \parallel \hat{p}^*)$ has an upper bound, then $\hat{p}^*$ satisfies Condition 7. This can be seen by first noticing that

$$\exists \beta \quad \forall n \geq 1 \quad d(b_1 \cdots b_n \parallel \hat{p}^*) \leq \beta \tag{10}$$

is equivalent to

$$\exists \beta \quad \forall n \geq 1 \quad \forall \hat{q} \quad \begin{pmatrix} \ell(\hat{p}^*) - \log p^*(b_1 \cdots b_n) \\ -\ell(\hat{q}) + \log q(b_1 \cdots b_n) \end{pmatrix} \leq \beta \, .$$

This latter condition implies

$$\forall \hat{q} \quad \exists \beta \quad \forall n \geq 1 \quad p^*(b_1 \cdots b_n) \geq 2^{\ell(\hat{p}^*) - \ell(\hat{q}) - \beta} q(b_1 \cdots b_n)$$

which can be shown to be equivalent to Condition 7 by equating $C_{\hat{q}}$ in Condition 7 with $2^{\ell(\hat{p}^*) - \ell(\hat{q}) - \beta}$. Condition 10 above can therefore be used as an alternative criterion for selecting optimal gambling functions.

4. CONVERGING TO AN OPTIMAL $\hat{P}^*$

As discussed earlier, an optimal gambling function must be arrived at asymptotically by making guesses as to what the function should be as each bit in the observation stream is revealed. If $\hat{p}^k$ is the guess made after seeing the first k bits, this process results in a sequence of computable gambling functions $\{\hat{p}^k\}_{k=1}^{\infty}$. To converge asymptotically on an optimal $\hat{p}^*$ (or a set of optimal $\hat{p}^*$'s) is to produce a sequence $\{\hat{p}^k\}_{k=1}^{\infty}$ for which all guesses beyond a certain point in the sequence are optimal gambling functions. Using Condition 10 as the optimality criterion, we therefore want to construct a sequence of gambling functions such that

$$\exists K \quad \forall k > K \quad \exists \beta \quad \forall n \geq 1 \quad d(b_1 \cdots b_n \parallel \hat{p}^k) \leq \beta \, . \tag{11}$$

The problem of selecting an appropriate $\hat{p}^*$ is thus reduced to the problem of choosing an appropriate $\hat{p}^k$ at each step.

One rule for choosing $\hat{p}^k$ that immediately comes to mind is to select the gambling function that minimizes $d(b_1 \cdots b_k \parallel \hat{p}^k)$. As it turns out, this is equivalent to minimizing the description length, since $d(b_1 \cdots b_k \parallel \hat{p}^k)$ achieves a minimum of zero when $\hat{p}^k$ minimizes

$$\ell(\hat{p}^k) - \log p^k(b_1 \cdots b_k) \, .$$

Unfortunately, it is not clear whether this rule always converges on a set of optimal gambling functions in the sense of Condition 11. If there exists a $\hat{p}$ for which $d(b_1 \cdots b_n \parallel \hat{p})$ is bounded and it happens to be the case that the number of distinct programs in the sequence $\{\hat{p}^k\}_{k=1}^{\infty}$ is finite, then it is relatively easy to show using Theorem 1 that the sequence does indeed converge. For example, it can be shown using Barron's analysis [Barron 1985] that this will occur if we restrict our attention to stationary ergodic processes. To generalize this result to the case in which the number of distinct $\hat{p}^k$'s is infinite, it is necessary to rule out the case in which each distinct $\hat{p}^k$ appears only a finite number of times and $d(b_1 \cdots b_n \parallel \hat{p}^k)$ diverges for every $\hat{p}^k$. A proof that this case can be ruled out has not yet been found, however.

A somewhat different rule can be obtained by modifying the minimum description-length rule so as to ensure that the number of distinct programs in the sequence $\{\hat{p}^k\}_{k=1}^{\infty}$ is finite whenever an optimal gambling function exists. This is accomplished by choosing the $\hat{p}^k$ that minimizes

$$\ell(\hat{p}^k) + d(b_1 \cdots b_k \parallel \hat{p}^k) \, . \tag{12}$$

For this selection rule, the following theorem holds:

Theorem 2. Suppose that there exists a computable gambling function $\hat{p}^*$ satisfying Condition 10. Let $\hat{p}^k$ be chosen so as to minimize Equation 12. Then the following statements are true:

(1) There are a finite number of distinct programs in the sequence $\{\hat{p}^k\}_{k=1}^{\infty}$.

(2) Every $\hat{p}^k$ is optimal for k sufficiently large (i.e., $\{\hat{p}^k\}_{k=1}^{\infty}$ satisfies Condition 11).

The existence of an optimal $\hat{p}^*$ places an upper bound on the length of each $\hat{p}^k$, thus ensuring that the number of distinct $\hat{p}^k$'s is finite. $\hat{p}^*$ also places an upper bound on $d(b_1 \cdots b_k \parallel \hat{p}^k)$. Since the number of distinct $\hat{p}^k$'s that can possibly diverge is finite, it follows from Theorem 1 that there will be a value of n after which $d(b_1 \cdots b_n \parallel \hat{p}^k)$ exceeds this bound for all $\hat{p}^k$'s that diverge. All $\hat{p}^k$'s beyond this point must therefore satisfy Condition 11. Minimizing Equation 12 thus produces a sequence of computable gambling functions $\{\hat{p}^k\}_{k=1}^{\infty}$ that converges asymptotically to a finite set of optimal gambling functions if an optimal gambling function exists.

5. SUMMARY AND DISCUSSION

A criterion has been presented for judging whether a proposed predictive theory is an appropriate model for an infinite set of data. In addition, a rule was presented for selecting an appropriate theory based on a finite set of observations. A proof was briefly outlined demonstrating that the theories thus selected obey the appropriateness criterion given a sufficient number of observations.

While the convergence of this rule is a pleasing result, there are barriers to its practical implementation. If the language for describing theories permits one to define the notion of a Turing machine, then the selection rule will be undecidable, owing to the halting problem of Turing machines. Even if this is not the case, the number of theories that must be compared when applying the rule may be impractically large. To apply the rule in practice, one must therefore introduce restrictions and/or approximations. It would be a worthwhile enterprise, for example, to characterize the kinds of models that can be learned in polynomial time, much as is being done by Valiant and others in concept learning [e.g., see review article by Kearns *et al.* 1987]. Nonetheless, from the standpoint of uncovering the fundamental principles of inductive inference, the rule presented in this paper and its accompanying analysis provide a mathematical basis for exploring the theoretical limits of what can be learned independent of the amount of computation involved.

From this theoretical standpoint, the analysis raises an intriguing philosophical issue. Although it was assumed in the introduction that we would be considering noisy data, at no point was this assumption made in the analysis. The need to consider nondeterministic models arose due to the fact that not all observation sequences have computable generating functions. There are countably many computer programs (i.e., they may be placed in one-to-one correspondence with the natural numbers) but uncountably many infinite binary sequences (i.e., they may be placed in one-to-one correspondence with the real numbers). Consequently, it is impossible to associate each binary sequence with a computable generating function that predicts the sequence exactly. One of three possibilities therefore exist for any given observation sequence:

(1) The sequence has a computable generating function and, hence, can be predicted exactly.

(2) A computable generating function does not exist; however, a computable probabilistic model can be constructed that predicts the sequence as well as any other computable model.

(3) A computable generating function does not exist and, for every computable probabilistic model, there exists another that is asymptotically more accurate in its predictions.

This raises the following question: if either Case 2 or 3 holds for a particular sequence, was that sequence generated by a "random" process? From a mathematical standpoint, the sequence exists as an entity in an abstract space. There is also a well-defined generating function that predicts the sequence exactly, it is just that this function is not computable. Does the fact that it is not computable necessarily imply that the sequence arose from a random process? Could it not have been predetermined in some sense? Is the apparent randomness a property of the thing being observed (i.e., ontological), or is it due to a fundamental limit on the kind of knowledge one can possess of that thing (i.e., epistemological)? Do random processes truly exist in the universe, as some proponents of Quantum Mechanics would have us believe, or is Quantum Mechanics merely the best theory we can come up with given the limitations of mind and machine? While the analysis presented in this paper does not purport to resolve these issues, I hope it will at least provoke some lively debate.

ACKNOWLEDGEMENTS

The research presented here was inspired by numerous discussions with Tom Cover on information theory, Kolmogorov complexity, logical smoothing, and gambling. The choice of a gambling scenario for comparing candidate theories was influenced by Tom's ideas on using gambling as a basis for motivating probability theory. I would like to thank Marla and Corinne Babcock, John Gabbe, Alan Ginsberg, Lawrence O'Gorman, and Jakub Segen for their comments on earlier drafts of the paper.

REFERENCES

[Angluin and Smith, 1983] D. Angluin and C.H. Smith. Inductive inference: theory and methods. *Computing Surveys*, Vol. 15, No. 3, pp 237–269 (September 1983).

[Barron and Cover, 1983] A.R. Barron and T.M. Cover. Convergence of logically simple estimates of unknown probability densities. Presented at the *1983 International Symposium on Information Theory*, St. Jovite, Quebec, Canada (1983).

[Barron, 1985] A.R. Barron. Logically smooth density estimation. Technical Report 56, Department of Statistics, Stanford University, Stanford, California (1985).

[Dietterich and Michalski, 1983] T.G. Dietterich and R.S. Michalski. A comparative review of selected methods for learning from examples. In *Machine Learning: An Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell, and T.M. Mitchell (eds.), pp 41–81 (Tioga Publishing, Palo Alto, California, 1983).

[Gallager, 1968] R.G. Gallager. *Information Theory and Reliable Communication* (John Wiley and Sons, New York, New York, 1968).

[Kearns *et al.*, 1987] M. Kearns, M. Li, L. Pitt, and L. Valiant. Recent results on boolean concept learning. *Proc. 4th International Workshop on Machine Learning*, Irvine, California, pp 337–352 (June 1987).

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description. *Automatica*, Vol. 14, pp 465–471 (1978).

[Rissanen, 1983] J. Rissanen. A universal prior of integers and estimation by minimum description length. *Annals of Statistics*, Vol. 11, pp 416–431 (1983).

[Segen, 1980] J. Segen. *Pattern-Directed Signal Analysis: Unsupervised Model Inference, Applications to EEG and Speech.* Ph.D. Thesis, Dept. of Electrical Engineering, Carnegie-Mellon University, Pittsburgh, PA (1980).

[Segen, 1985] J. Segen. Learning concept descriptions from examples with errors. *Proc. IJCAI-85*, Los Angeles, California, pp 634–636 (August, 1985).

[Sorkin, 1983] R. Sorkin. A quantitative occam's razor. *International Journal of Theoretical Physics*, Vol. 22, pp 1091–1103 (1983).

[Tukey, 1977] J.W. Tukey. *Exploratory Data Analysis* (Addison-Wesley, Reading, Massachusetts, 1977).