Complexity-Based Induction Systems: Comparisons and Convergence Theorems

R. J. SOLOMONOFF, MEMBER, IEEE

Abstract—**In 1964 the author proposed as an explication of *a priori* probability the probability measure induced on output strings by a universal Turing machine with unidirectional output tape and a randomly coded unidirectional input tape. Levin has shown that if $\tilde{P}'_M(x)$ is an unnormalized form of this measure, and $P(x)$ is any computable probability measure on strings, x, then**

$$\tilde{P}'_M(x) > CP(x)$$

where C is a constant independent of x. The corresponding result for the normalized form of this measure, P'_M, is directly derivable from Willis' probability measures on nonuniversal machines. If the conditional probabilities of P'_M are used to approximate those of P, then the expected value of the total squared error in these conditional probabilities is bounded by $-(1/2)\ln C$. With this error criterion, and when used as the basis of a universal gambling scheme, P'_M is superior to Cover's measure b^*. When $H^* \equiv -\log_2 P'_M$ is used to define the entropy of a finite sequence, the equation $H^*(x,y) = H^*(x) + H^*_x(y)$ holds exactly, in contrast to Chaitin's entropy definition, which has a nonvanishing error term in this equation.

I. Introduction

In 1964 [1], we proposed several models for probability based on program size complexity. One of these, P'_M, used a universal Turing machine with unidirectional input and output tapes with the input tape having a random sequence. While the relative insensitivity of the models to the choice of universal machine was shown, with arguments and examples to make them reasonable explicata of "probability," few rigorous results were given. Furthermore, the "halting problem" cast some doubt on the existence of the limits defining the models.

However, Levin [8, Th. 3.3, p. 103] proved that the probability assigned by P'_M to any finite string, $x(n)$, differs by only a finite constant factor from the probability assigned to $x(n)$ by any computable probability measure, the constant factor being independent of $x(n)$.

Manuscript received August 27, 1976; revised November 22, 1977. This work was supported in part by the United States Air Force Office of Scientific Research under Contracts AF-19(628)5975, AF-49(638)-376, and Grant AS-AFOSR 62-377; in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contracts N00014-70-A-0362-0003 and N00014-70-A-0362-0005; and in part by the Public Health Service under NIH Grant GM 11021-01. This paper was presented at the IEEE International Symposium on Information Theory, Cornell University, Ithaca, NY, October 10–14, 1977.

The author is with the Rockford Research, Inc., Cambridge, MA 02138.

Since the measure P'_M is not effectively computable, for practical induction it is necessary to use computable approximations, such as those investigated by Willis [2]. Sections II and III show the relationship of Willis' work on computable probability measures and the machines associated with them to the incomputable measure P'_M and its associated universal machine.

Section IV shows that if the conditional probabilities of P'_M are used to approximate those of any computable probability measure, then the expected value of the total squared error for these conditional probabilities is bounded by a constant. This superficially surprising result is shown to be consistent with conventional statistical results.

Section V deals with Chaitin's [3] probability measure and entropy definitions. These are based on Turing machines that accept only prefix sets as inputs, and are of two types: conditional and unconditional. His unconditional probability is not directly comparable to P'_M, since it is defined for a different kind of normalization. Leung-Yan-Cheong and Cover [4] used a variant of his conditional probability that appears to be very close to P'_M, but there is some uncertainty about the effect of normalization.

Section VI discusses Cover's [5] b^*, a probability measure based on Chaitin's unconditional entropy. P'_M is shown to be somewhat better than b^* with respect to mean-square error. Also, if used as the basis of a gambling system, it gives larger betting yields than b^*.

In Section VII $H^* \equiv -\log_2 P'_M$ is considered as a definition of the entropy of finite sequences. H^* is found to satisfy the equation

$$H^*(x,y) = H^*(x) + H^*_x(y)$$

exactly, whereas Chaitin's entropy definition requires a nonvanishing error term.

For ergodic ensembles based on computable probability measures, $E(H^*(X(n)))/n$ is shown to approach H, the entropy of the ensemble. The rate of approach is about the same as that of $E(H^c(X(n)/n))/n$ and perhaps faster than that of $E(H^c(x(n)))/n$ where $H^c(X(n)/n)$ and $H^c(X(n))$ are Chaitin's conditional and unconditional entropies, respectively.

0018-9448/78/0700-422$00.75

II. P'_M and Willis' Probability Measures

The various models proposed as explications of probability [1] were initially thought to be equivalent. Later [6] is was shown that these models form two equivalence classes: those based on a general universal Turing machine and those based on a universal Turing machine with unidirectional input and output tapes and a bidirectional work tape. We will call this second type of machine a "universal UIO machine."

One model of this class [1, Section 3.2, pp. 14–18] uses infinite random strings as inputs for the universal UIO machine. This induces a probability distribution on the output strings that can be used to obtain conditional probabilities through Bayes' theorem.

Suppose M is a (not necessarily universal) UIO machine with working symbols 0 and 1. If it reads a blank square on the input tape (e.g., at the end of a finite program), it always stops. We use $x(n)$ to denote a possible output sequence containing just n symbols, and s to denote a possible input sequence.

We say "s is a code of $x(n)$ (with respect to M)" if the first n symbols of $M(s)$ are identical to those of $x(n)$. Since the output tape of M is unidirectional, the first n bits of $M(s)$ can be defined even though subsequent bits are not; e.g., the machine might print n bits and then go into an infinite nonprinting loop.

We say "s is a minimal code of $x(n)$" if 1) s is a code of $x(n)$, and 2) when the last symbol of s is removed, the resultant string is no longer a code of $x(n)$. All codes for $x(n)$ are of the form $s_i a$, where s_i is one of the minimal codes of $x(n)$, and a may be a null, finite, or infinite string. It is easy to show that for each n the minimal codes for all strings of length n form a prefix set.

Let $N(M,x(n),i)$ be the number of bits in the ith minimal code of $x(n)$, with respect to machine M. We set $N(M,x(n),i)=\infty$ if there is no code for $x(n)$ on machine M.

Let $x_j(n)$ be the jth of the 2^n strings of length n. $N(M,x_j(n),i)$ is the number of bits in the ith minimal code of the jth string of length n. For a universal machine M we defined P'_M in [1] by

$$P'_M(x(n)) \triangleq \sum_{i=1}^{\infty} 2^{-N(M,x(n),i)} \Big/ \sum_{j=1}^{2^n} \sum_{i=1}^{\infty} 2^{-N(M,x_j(n),i)}. \tag{1}$$

This equation can be obtained from [1, (7), p. 15] by letting the T of that equation be the null sequence, and letting a be the sequence $x(n)$. The denominator is a normalization factor.

Although P'_M appeared to have many important characteristics of an *a priori* probability, there were serious difficulties with this definition. Because of the "halting problem," both the numerator and denominator of (1) were not effectively computable, and the sums had not been proved to converge.

Another less serious difficulty concerned the normalization. While P'_M satisfies

$$\sum_{j=1}^{2^n} P'_M(x_j(n)) = 1, \tag{2}$$

it does not appear to satisfy the additivity condition

$$P'_M(x(n)) = P'_M(x(n)0) + P'_M(x(n)1). \tag{3}$$

The work of Willis (2), however, suggested a rigorous interpretation of (1) that made it possible to demonstrate the convergence of these sums and other important properties. With suitable normalization, the resultant measure could be made to satisfy both (2) and (3).

Willis avoids the computability difficulties by defining a set of measures based on specially limited machines that have no "halting problem." He calls these machines FOR's (Frames of Reference). One important example of a FOR is the machine M_T, which is the same as the universal UIO machine M except that M_T always stops at time T if it has not stopped already. For very large T, M_T behaves much like a universal UIO machine. Willis' measure is defined by the equation

$$P^R(x(n)) = \sum_i 2^{-N(R,x(n),i)}. \tag{4}$$

The sum over i is finite, since for finite n a FOR has only a finite number of minimal codes. This measure differs from that of (1) in being based on a nonuniversal machine, and in being unnormalized in the sense of (2) and (3). Usually

$$\sum_{j=1}^{2^n} P^R(x_j(n)) < 1.$$

Let us define $\tilde{P}'_M$ to be the numerator of (1). It can be obtained from Willis' measure by using M_T and letting T approach infinity:

$$\tilde{P}'_M(x(n)) \equiv \lim_{T\to\infty} \sum_i 2^{-N(M_T,x(n),i)}. \tag{5}$$

Theorem 1: The limit in (5) exists.

Proof: The minimal codes for sequences of length n form a prefix set, so by Kraft's inequality,

$$\sum_i 2^{-N(M_T,x(n),i)} \leqslant 1.$$

Furthermore, this quantity is an increasing function of T, since as T increases, more and more codes for $x(n)$ can be found. Since any monotonically increasing function that is bounded above must approach a limit, the theorem is proved.

For certain applications and comparisons between probability measures, it is necessary that they be normalized in the sense of (2) and (3). To normalize P'_M, define

$$\begin{aligned} P'_M(x(n)) &\triangleq \tilde{P}'_M(x(n))C(x(n)) \\ &\triangleq \tilde{P}'_M(x(n)) \prod_{i=0}^{n-1} \frac{\tilde{P}'_M(x(i))}{\tilde{P}'_M(x(i)0)+\tilde{P}'_M(x(i)1)}. \end{aligned} \tag{6}$$

Here $C(x(n))$ is the normalization constant, and n is any positive integer.

We will now show that P'_M satisfies (2) and (3) for $n \geqslant 1$.

It is readily verified from (6) that P'_M satisfies (3) for $n \geqslant 1$. To show (2) is true for $n \geqslant 1$, first define $\tilde{P}'_M(x(0)) \triangleq 1$, $x(0)$ being the sequence of zero length.Then from (6)

$$P'_M(0) = \tilde{P}'_M(0) \frac{\tilde{P}'_M(x(0))}{\tilde{P}'_M(0) + \tilde{P}'_M(1)},$$

$$P'_M(1) = \tilde{P}'_M(1) \frac{\tilde{P}'_M(x(0))}{\tilde{P}'_M(0) + \tilde{P}'_M(1)},$$

so $P'_M(0) + P'_M(1) = P'_M(x(0)) = 1$, and thus (2) is true for $n = 1$. (3) implies that if (2) is true for n, then it must be true for $n+1$. Since (2) is true for $n = 1$, it must be true for all n. Q.E.D.

III. The Probability Ratio Inequality for P'_M

In this section we will develop and discuss an important property of P'_M. First we define several kinds of probability measures.

The term *computable probability measure* (cpm) will be used in Willis' sense [2, pp. 249–251]. Loosely speaking, it is a measure on strings, satisfying (2) and (3), which can be computed to within an arbitrary nonvanishing error ϵ in finite time.

Paraphrasing Willis, we say a probability measure P on finite strings is *computable* if it satisfies (2) and (3) and there exists a UIO machine with the following properties: a) it has two input symbols (0 and 1) and a special input punctuation symbol, b (blank); b) when the input to the machine is $x(n)b$, its output is the successive bits of a binary expansion of $P(x(n))$. If $P(x(n)) = 0$, the machine prints 0 and halts in a finite time.

If the machine can be constructed so that it always halts after printing only a finite number of symbols, then P is said to be a *2-computable probability measure* (2-cpm).

Levin [8, p 102, Def. 3.6] has defined a *semi-computable probability measure* (scpm) $\tilde{P}_Q$, and has shown it to be equivalent to

$$\tilde{P}_Q(x(n)) \triangleq \lim_{T \to \infty} \sum_i 2^{-N(Q_T, x(n), i)} \tag{7}$$

where Q is an arbitrary (not necessarily universal) UIO machine. From (5) it is clear that $\tilde{P}'_M$ is a semi-computable measure in which Q is universal.

A *normalized semicomputable probability measure* (nscpm) is a measure that is obtainable from a scpm by a normalization equation such as (6). It satisfies (2) and (3).

A simple kind of probability measure is the binary Bernoulli measure in which the probability of the symbol 1 is p. If p is a terminating binary fraction such as $3/8$, then the measure is a 2-cpm. If p is a computable real number such as $1/2$ or $1/3$ or $(1/2)\sqrt{2}$, then the measure is a cpm. If p is an incomputable real or simply a random number between 0 and 1, then the measure is not a cpm. Neither is it a scpm nor a nscpm. Since computable numbers are denumerable, almost all real numbers are incomputable, and so this type of incomputable probability measure is quite common. The most commonly used probabilistic models in science—i.e., continuous probabilistic functions of incomputable (or random) parameters—are of this type. Though none of the theorems of the present paper are directly applicable to such measures, we will outline some relevant results that have been obtained through further development of these theorems.

While $\tilde{P}'_M$ is a semi-computable probability measure, we will show as a corollary of Theorem 2 that it is not a cpm. Moreover, P'_M is a nscpm, but it is *not* a scpm.

All 2-cpms are cpms. All cpms are scpms. All cpms are nscpms. However, scpms and ncpms have no complete inclusion relation between them, since, as we have noted, P'_M is a nscpm but not a scpm, and $\tilde{P}'_M$ is a scpm but not a nscpm. Schubert [14, p. 13, Th. 1(a)] has shown that all probability measures that are both scpms and nscpms must be cpms. It is easy to draw a Venn diagram showing these relations.

Theorem 2: Given any universal UIO machine M and any computable probability measure P there exists a finite positive constant k such that for all $x(n)$

$$P'_M(x(n)) \geqslant 2^{-k} P(x(n)). \tag{8}$$

Here $x(n)$ is an arbitrary finite string of length n, and k depends on M and P but is independent of $x(n)$.

We will first prove Lemma 1:

Lemma 1: Given any universal UIO machine and any 2-computable probability measure P' there exists a finite positive constant k' such that for all $x(n)$

$$P'_M(x(n)) \geqslant 2^{-k'} P'(x(n)). \tag{9}$$

Lemma 1 is identical to Theorem 2, but applies only for 2-computable probability measures. Its proof will be similar to that of Willis' Theorem 16 [2, p. 256]

Proof of Lemma 1: From Willis ([2, p. 252, Theorem 12], but also see [4, Lemma of the last Theorem] for a more transparent proof), we note that there constructively exists a FOR R_0 such that for all $x(n)$

$$P^{R_0}(x(n)) = \sum_i 2^{-N(R_0, x(n), i)} = P'(x(n)). \tag{10}$$

Since R_0 is a FOR, it has only a finite number of minimal codes for $x(n)$, and they are all effectively computable. Since M is universal, it has minimal codes for $x(n)$ that are longer than those of R_0 by an additive constant k. This may be seen by considering the definition of "minimal code." If σ is a minimal code for R_0 and $R_0(\sigma) = x(n)$, then $M(S\sigma) = x(n)$, S being the simulation instructions from R_0 to M. If σ' is σ with the last symbol removed, then since σ is a *minimal* code, $R_0(\sigma') \neq x(n)$, implying $M(S\sigma') \neq x(n)$, so $S\sigma$ must be a minimal code for $x(n)$ with respect to M. Thus,

$$N(M, x(n), i) = N(F_0, x(n), i) + k \tag{11}$$

where k is the length of the M simulation instructions for R_0. As a result,

$$\sum_i 2^{-N(M_T,x(n),i)} \geq \sum_i 2^{-N(R_0,x(n),i)-k} = 2^{-k}P'(x(n)) \quad (12)$$

for large enough T. If it takes at most $T_{x(n)}$ steps for M to simulate the R_0 minimal code executions resulting in $x(n)$, then "large enough T" means $T \geq T_{x(n)}$. We have the inequality sign in (12) because M_T may have minimal codes for $x(n)$ in addition to those that are simulations of the R_0 codes.

From (12), (5), and Theorem 1,

$$\tilde{P}'_M(x(n)) \geq 2^{-k}P'(x(n)). \quad (13)$$

In (6) we note that the normalization constant $C(x(n))$ is the product of factors

$$\frac{\tilde{P}'_M(x(i))}{\tilde{P}'_M(x(i)0)+\tilde{P}'_M(x(i)1)}.$$

Appendix A shows that each of these factors must be ≥ 1. As a result, $P'_M \geq \tilde{P}'_M$, and from (13) we have $P'_M(x(n)) \geq 2^{-k}P'(x(n))$, which proves Lemma 1. To prove Theorem 2, we first note [2, p. 251] that if P is any computable probability measure and ϵ is a positive real <1, then there exists a 2-computable probability measure P' such that for all finite strings $x(n)$,

$$P(x(n))(1-\epsilon) \leq P'(x(n)) \leq P(x(n))(1+\epsilon).$$

Starting with our P, let us choose $\epsilon = 1/2$ and obtain a corresponding P' such that

$$P' \geq \frac{1}{2}P. \quad (14)$$

From Lemma 1 we can find a k' such that

$$P'_M \geq 2^{-k'}P' \geq 2^{-k'-1}P \quad (15)$$

so, with $k = k'+1$, Theorem 2 is proved.

Corollary 1 to Theorem 2: Let $[s_i]$ be the set of all strings such that for all x

$$M(s_i x) = R_0(x),$$

i.e., s_i is a code for the M simulation of R_0. Let $[s'_i]$ be any subset of $[s_i]$ that forms a prefix set. If $|s'_i|$ is the number of bits in the string s'_i, then for all $x(n)$

$$P'_M(x(n)) \geq \sum_i 2^{-|s'_i|}P(x(n)). \quad (16)$$

The summation is over all members of the prefix set $[s'_i]$. The proof is essentially the same as that of Theorem 2. Q.E.D.

To obtain the best possible bound on P'_M/P, we would like to choose the prefix set so that

$$\sum_i 2^{-|s'_i|}$$

is maximal. It is not difficult to choose such a subset, given the set $|s_i|$.

Willis [2, p. 256, Th. 17] has shown that if P is any cpm, then there constructively exists another cpm P' such that for any finite $k>0$ there exists a $x(n)$ for which

$$P'(x(n)) > kP(x(n)).$$

From this fact and from Theorem 2, it is clear that P'_M cannot be a cpm.

Levin [8, p. 103, Th. 3.3] has shown that if $\tilde{P}_Q(x(n))$ is any semicomputable probability measure, then there exists a finite $C>0$ such that for all $x(n)$,

$$\tilde{P}'_M(x(n)) \geq C\tilde{P}_Q(x(n)).$$

From this it follows that, since the normalization constant of P'_M is always ≥ 1,

$$P'_M(x(n)) \geq C'\tilde{P}_Q(x(n)), \quad (16)$$

giving us a somewhat more powerful result than Theorem 2. Note, however, that in (16) $\tilde{P}_Q$ is restricted to be a semicomputable probability measure, rather than a normalized semicomputable probability measure—a constraint which will limit its applicability in the discussions that follow.

To what extent is P'_M unique in satisfying the probability ratio inequality of (8)? In Sections V and VI we will discuss other measures, also based on universal machines, that may have this property. T. Fine notes [13] that if P is known to be a member of an effectively enumerable set of probability measures $[P_i]$, then the measure

$$P' = \sum_i a_i P_i \qquad \left(\text{with } a_i > 0,\ \sum_i a_i = 1\right)$$

also satisfies

$$P' = \sum_i a_i P_i \geq 2^{-k}P_j, \qquad \text{where } k = -\lg a_j,$$

and lg denotes logarithm to base 2. Under these conditions the solution to (8) is not unique. However, while the set of all computable probability measures is enumerable, it is not effectively enumerable, so this solution is not usable in the most general case.

One interpretation of Theorem 2 is given by the work of Cover [5]. Suppose P is used to generate a stochastic sequence, and one is asked to make bets on the next bit of the sequence at even odds. If P is known and bets are made starting with unity fortune so as to maximize the expected value of the logarithm of one's fortune, then the value of one's fortune after n bits of the sequence $x(n)$ have occurred is $2^n P(x(n))$. On the other hand, if it is only known that P is a cpm, and P'_M instead of P is used as a basis for betting, the yield will be $2^n P'_M(x(n))$. The ratio of yield using P'_M to that using the best possible information is then $P'_M(x(n))/P(x(n))$, which as we have shown is $\geq 2^{-k}$.

Cover also shows that if P is used in betting, then for large n the geometric-mean yield per bet is almost certainly $2^{(1-H)}$, where H is the asymptotic entropy per symbol (if it exists) of the sequence generator. If we do not know P, and use P'_M as a basis for betting, our mean yield becomes $2^{-k/n}2^{(1-H)}$. The ratio of the geometric yield *per bet* of P'_M to that of P is $2^{-k/n}$. For large n, this ratio approaches unity.

The bets in these systems depend on the conditional probabilities of P and P'_M. That bets based on P give the maximum possible log yield, and that bets based on P'_M have almost as large a yield as P, suggests that their conditional probabilities are very close. Theorem 3 shows that this is usually true.

IV. Convergence of Expected Value of Total Square Error of P'_M

We will show that if P is any computable probability measure, then the individual conditional probabilities given by P'_M tend to converge in the mean-square sense to those of P.

Theorem 3: If P is any computable probability measure, then

$$\boldsymbol{E}_P\left(\sum_{i=0}^{n-1}\left(\delta_i^n-\delta_i^{n'}\right)^2\right)\triangleq\sum_{j=1}^{2^n}P\left(x_j(n)\right)\cdot\sum_{i=0}^{n-1}\left({}^j\delta_i^n-{}^j\delta_i^{n'}\right)^2\leqslant k\ln\sqrt{2}\ . \quad (17)$$

Notation:

$\boldsymbol{E}_P$	expected value with respect to P,
$x_j(n)$	jth sequence of length n,
${}^j\delta_i^n$, ${}^j\delta_i^{n'}$	conditional probabilities, given the first i bits of $x_j(n)$, that the next bit will be zero for P and P'_M respectively,
δ_i^n	random ${}^j\delta_i^n$, where j corresponds to the $x_j(n)$ randomly chosen by the measure P.

The proof is based on two lemmas.

Lemma 1: If $0\leqslant x\leqslant 1$, then

$$R(x,y)\triangleq x(\lg x-\lg y)+(1-x)(\lg(1-x)-\lg(1-y))\geqslant\frac{2}{\ln 2}(x-y)^2.$$

Lemma 2: Let

$$A_n\triangleq\sum_{j=1}^{2^n}P(x_j(n))\sum_{i=0}^{n-1}R\left({}^j\delta_i^n,{}^j\delta_i^{n'}\right) \quad (18)$$

and

$$B_n\triangleq\sum_{j=1}^{2^n}P(x_j(n))(\lg P(x_j(n))-\lg P'_M(x_j(n))). \quad (19)$$

Then for $n\geqslant 1$, $A_n=B_n$.

To prove Theorem 3, we take the expected value of the lg of both sides of (8) and obtain

$$k\geqslant B_n.$$

From Lemma 2,

$$k\geqslant A_n. \quad (20)$$

From (18), (20), and Lemma 1,

$$k\geqslant\frac{2}{\ln 2}\sum_{j=1}^{2^n}P(x_j(n))\sum_{i=0}^{n-1}\left({}^j\delta_i^n-{}^j\delta_i^{n'}\right)^2$$

which proves Theorem 3.

The proof of Lemma 1 is elementary and is omitted.

To prove Lemma 2, we will first show that $A_1=B_1$ and then that $A_{n+1}-A_n=B_{n+1}-B_n$, from which the lemma follows by mathematical induction. To show $A_1=B_1$, let $D\equiv P(x_1(1))$, $D'\equiv P'_M(x_1(1))$, and note that $P(x_2(1))=1-D$, $P'_M(x_2(1))=1-D'$, ${}^1\delta_0^1={}^2\delta_0^1=D$, and ${}^1\delta_0^{1'}={}^2\delta_0^{1'}=D'$. Then from (18) and (19)

$$\begin{aligned}A_1&=D\cdot R(D,D')+(1-D)\cdot R(D,D')=R(D,D')\\B_1&=D(\lg D-\lg D')+(1-D)(\lg(1-D)-\lg(1-D'))\\&=R(D,D')\\A_1&=B_1.\end{aligned} \quad (21)$$

Next we compute B_{n+1}. B_n was obtained by summing 2^n terms containing probability measures. The corresponding 2^{n+1} terms for B_{n+1} are obtained by splitting each of the 2^n terms of B_n and multiplying by the proper conditional probabilities. Then

$$\begin{aligned}B_{n+1}&=\sum_{j=1}^{2^n}\Big[P(x_j(n))\Big\{{}^j\delta_n^n\Big(\lg\left[P(x_j(n))\cdot{}^j\delta_n^n\right]\\&\quad-\lg\left[P'_M(x_j)\cdot{}^j\delta_n^{n'}\right]\Big)\\&\quad+\left(1-{}^j\delta_n^n\right)\Big(\lg\left[P(x_j(n))\cdot\left(1-{}^j\delta_n^n\right)\right]\\&\quad-\lg\left[P'_M(x_j)\cdot\left(1-{}^j\delta_n^{n'}\right)\right]\Big)\Big\}\Big]\\&=\sum_{j=1}^{2^n}\Big[P(x_j(n))\Big\{{}^j\delta_n^n\left(\lg P(x_j(n))\right.\\&\quad\left.-\lg P'_M(x_j(n))+\lg{}^j\delta_n^n-\lg{}^j\delta_n^{n'}\right)\\&\quad+\left(1-{}^j\delta_n^n\right)\left(\lg P(x_j(n))-\lg P'_M(x_j(n))\right.\\&\quad\left.+\lg\left(1-{}^j\delta_n^n\right)-\lg\left(1-{}^j\delta_n^{n'}\right)\right)\Big\}\Big]\\&=\sum_{j=1}^{2^n}P(x_j(n))\left(\lg cp(x_j(n))-\lg P'_M(x_j(n))\right)\\&\quad+\sum_{j=1}^{2^n}P(x_j(n))\Big[{}^j\delta_n^n\left(\lg{}^j\delta_n^n-\lg{}^j\delta_n^{n'}\right)\\&\quad+\left(1-{}^j\delta_n^n\right)\left(\lg\left(1-{}^j\delta_n^n\right)-\lg\left(1-{}^j\delta_n^{n'}\right)\right)\Big]\\&=B_n+\sum_{j=1}^{2^n}P(x_j(n))R\left({}^j\delta_n^n,{}^j\delta_n^{n'}\right).\end{aligned} \quad (22)$$

To obtain $A_{n+1}-A_n$, we have

$$A_{n+1}=\sum_{j=1}^{2^n}P(x_j(n))\sum_{i=0}^{n-1}R\left({}^j\delta_i^n,{}^j\delta_i^{n'}\right)+\sum_{j=1}^{2^n}P(x_j(n))R\left({}^j\delta_n^n,{}^j\delta_n^{n'}\right),$$

since

$$R\left({}^j\delta_n^n,{}^j\delta_n^{n'}\right)=R\left(1-{}^j\delta_n^n,1-{}^j\delta_n^{n'}\right),$$

and so

$$A_{n+1}=A_n+\sum_{j=1}^{2^n} P(x_j(n))R({}^j\delta_n^n,{}^j\delta_n^{n'}). \tag{23}$$

From (22) and (23), $A_{n+1}-A_n=B_{n+1}-B_n$, which completes the proof.

Corollary 1 to Theorem 3: If P' and P are probability measures (not necessarily recursive) satisfying the additivity and normalization (2) and (3) and

$$P'(x_i(n))\geqslant 2^{-k(n)}P(x_i(n)),$$

then

$$E_P\sum_{i=0}^{n-1}(\delta_i^n-\delta_i^{n'})^2\triangleq\sum_{j=1}^{2^n}P(x(n))\sum_{i=0}^{n-1}({}^j\delta_i^n-{}^j\delta_i^{n'})^2$$

$$\leqslant k(n)\ln\sqrt{2}\,.$$

The notation is the same as in Theorem 3 except that ${}^j\delta_i^{n'}$ is the conditional probability for P' rather than P'_M. The proof is essentially the same as that of Theorem 3.

This corollary is often useful in comparing probability measures, since the only constraint on its applicability is that $P'(x_i(n))>0$ for all $x_i(n)$ of a given n, where $i=1,2,\cdots,2^n$.

Ordinary statistical analysis of a Bernoulli sequence gives an expected squared error for the probability of the nth symbol proportional to $1/n$ and a total squared error proportional to $\ln n$. This is clearly much larger than the constant $k\ln\sqrt{2}$ given by Theorem 3. The discrepancy may be understood by observing that the parameters that define the Bernoulli sequence are real numbers, and as we have noted, probability measures that are functions of reals are almost always incomputable probability measures. Since Theorem 3 applies directly only to computable probability measures, the aforementioned discrepancy is not surprising.

A better understanding is obtained from the fact that the cpms to which Theorem 3 applies constitute a denumerable (but not effectively denumerable) set of hypotheses. On the other hand, Bernoulli sequences with real parameters are a nondenumerable set of hypotheses. Moreover, Koplowitz [7], Kurtz and Caines [11], and Cover [12] have shown that if one considers only a countable number of hypotheses, the statistical error converges much more rapidly than if the set of hypotheses is uncountable. Accordingly, the discrepancy we have observed is not unexpected.

When the measure P is a computable function of b continuous parameters, Theorems 2 and 3 must be slightly modified. We will state without proof that in this case the constant k in Theorem 2 is replaced by $k(n)=c+Ab\ln n$. Here n is the number of symbols in the string being described, A is a constant that is characteristic of the accuracy of the model, and c is the number of bits in the description of the expression containing the b parameters.

From Corollary 1 of Theorem 3, the expected value of the total squared error in conditional probabilities is

$$(c+Ab\ln n)\ln\sqrt{2}\,.$$

V. Chaitin's Probability Measures and Entropy

Chaitin [3] has defined two kinds of probability measure and two kinds of entropy. Conditional probability is defined by

$$P^c(s/t)\triangleq\sum 2^{-|r|},\qquad (U(r,t^*)=s)$$

where r, s, and t are finite binary strings, and $U(\cdot,\cdot)$ is a universal computer with two arguments. The acceptable first arguments (i.e., those for which the output is defined) form a prefix set for each value of the second argument. Also $|r|$ is the length of the string r, and t^* is the shortest string such that $U(t^*,\Lambda)=t$, where Λ is the null string.

U is "universal" in the sense that if C is any other prefix set computer such that $C(s,t)$ is defined, then there is an s' such that $U(s',t)=C(s,t)$ and $|s'|\leqslant|s|+k$, where k is a constant characteristic of U and C but independent of s and t.

Conditional entropy is defined as

$$H^c(s/t)\triangleq\min|r|\qquad\text{such that } U(r,t^*)=s. \tag{24}$$

Thus H^c is the length of the shortest program for s, given the shortest program for t.

Unconditional probability and entropy are defined by

$$P^c(s)\triangleq\sum 2^{-|r|},\qquad (U(r,\Lambda)=s), \tag{25}$$

$$H^c(s)\triangleq\min|r|,\qquad (U(r,\Lambda)=s). \tag{26}$$

Note that $P^c(\cdot)$ is not directly comparable to $P'_M(\cdot)$. On one hand, $\Sigma_i P^c(x_i)\leqslant 1$, the summation being over all finite strings x_i. On the other hand, $\Sigma_{i=1}^{2^n}P'_M(x_i(n))=1$, so $\Sigma_i P'_M(x_i)=\infty$.

While it is possible to normalize $P^c(\cdot)$ so that it satisfies (2) and (3), we have not been able to demonstrate anything about the relationship of the resultant measure to P'_M. $P^c(s/|s|)$, however, is comparable to P'_M. Leung-Yan-Cheong and Cover have shown [4, proof of the last theorem] that

$$P^c(s/|s|)\geqslant 2^{-k}P(s) \tag{27}$$

where P is any computable probability measure and k is a constant independent of the string s.

It is not difficult to show that

$$P^c(s/|s|)\geqslant 2^{-k'}\lim_{T\to\infty}\sum_i 2^{-N(M_T,s,i)}=2^{-k'}\tilde{P}'_M(s) \tag{28}$$

where k' is a constant independent of s.

To see why (28) is true, suppose r is some minimal program for s with respect to M_T. Then independently of T we can construct a program for s with respect to Chaitin's U that is k' bits longer than r. This program tells U to "simulate M, insert r into this simulated M, and stop when $|s|$ symbols have been emitted." Since U has already been given a program for $|s|$, these instructions are a fixed amount k' longer than r and are independent of T. Since M_T was able to generate s in $\leqslant T$ steps with r as input, these instructions for U are guaranteed to eventually produce s as output.

To be useful for induction, for high gambling yield or for small error in conditional probability, it is necessary

that a probability measure be normalizable in the sense of (2) and (3) and always be $\geqslant 0$. When $P^c(s/|s|)$ is normalized using (6), we have not been able to show that (27) continues to hold.

Fine [13] has suggested a modified method of normalization using a "finite horizon" that may be useful for some applications. First a large integer n is chosen. Then $P^c(\cdot/\cdot)$ is used to obtain a normalized probability distribution for all strings of length n:

$$Q_n^c(s(n)) = P^c(s(n)/n) \Big/ \sum_{(s':|s'|=n)} P^c(s'/n).$$

A probability distribution for strings $s(i)$ with $i \leqslant n$ is obtained by

$$Q_{i,n}^c(s(i)) = \sum_{(s'(n):s(i) \text{ is a prefix of } s'(n))} Q_n^c(s'(n)). \quad (29)$$

This probability distribution satisfies (2) and (3) and is >0 for all finite strings. Also, because of (27),

$$Q_{i,n}^c(s(i)) \geqslant 2^{-k} P(s(i)) \quad (30)$$

for any computable probability measure P. Furthermore the constant k can be shown to be independent of n. From (30) the proof of Theorem 3 holds without modification for $Q_{i,n}^c$.

A difficulty with this formulation is the finite value of n. It must always be chosen so as to be greater than the length of any sequence whose probability is to be evaluated. It is not clear that the distribution approaches a limit as n approaches infinity.

VI. Cover's Probability Measure b^*

Cover [5] has devised a probability measure based on Chaitin's unconditional entropy H^c that is directly comparable to P_M'. Let us define the measure

$$B^*(x(n)) \triangleq \sum_{z \in (0,1)^*} 2^{-H^c(x(n)z)} \quad (31)$$

where the summation is over the set of all finite strings $[z]$.

Cover defines the conditional probability that the finite string $x(n)$ will be followed by the symbol x_{n+1} to be

$$b^*(x_{n+1}|x(n)) \triangleq B^*(x(n)x_{n+1})/B^*(x(n)). \quad (32)$$

We will examine the efficiency of B^* when used as the basis of a universal gambling scheme and obtain a bound for the total squared error of its conditional probabilities when used for prediction. These will be compared with the corresponding criteria for P_M'.

Theorem 4: If P is any probability measure and

$$G(n) \equiv E_P(\lg P(x(n)) - \lg B^*(x(n))),$$

then

$$\lim_{n\to\infty} G(n) = \infty. \quad (33)$$

Lemma 1:

$$\lim_{n\to\infty} \sum_{i=1}^{2^n} B^*(x_i(n)) = 0. \quad (34)$$

Proof: Let us define $W(n) \equiv \sum_{i=1}^{2^n} 2^{-H^c(x_i(n))}$, where the sum is over all strings $x_i(n)$ of length n. Then from (31)

$$\sum_{i=1}^{2^n} B^*(x_i(n)) = \sum_{i=n}^{\infty} W(i). \quad (35)$$

By Kraft's inequality $\sum_{n=1}^{\infty} W(n) \leqslant 1$, so (35), which is the latter part of the summation of $W(n)$, must approach zero as n approaches infinity. Q.E.D.

Lemma 2: Let P_i be a set of nonnegative constants such that $\sum P_i = 1$. Then $\sum P_i \lg B_i$ is maximized, subject to the constraint that $\sum B_i = k$, by choosing $B_i = kP_i$. This is proved by using Lagrange multipliers.

Proof of Theorem 4: Consider a fixed value of n. The smallest value of $G(n)$ occurs when

$$E_P(\lg B^*(x(n))) = \sum_{i=1}^{2^n} P(x_i(n)) \lg B^*(x(n))$$

is a maximum. By Lemma 2, this occurs when

$$B^*(x_i(n)) = P(x_i(n)) \sum_{j=1}^{2^n} B^*(x_j(n)).$$

The minimum value of $G(n)$ is then

$$\sum_{i=1}^{2^n} P(x_i(n)) \left(\lg P(x_i(n)) - \lg \left(P(x(n)) \sum_{j=1}^{2^n} B^*(x_j(n)) \right) \right) = -\lg \sum_{i=1}^{2^n} B^*(x_i(n))$$

which by Lemma 1 approaches infinity as n approaches infinity. Q.E.D.

Theorem 5: If P is any computable probability measure and $F(n)$ is any recursive function from integers to integers such that $\lim_{n\to\infty} F(n) = \infty$, then there exists a constant k such that for all $x(n)$

$$\lg P(x(n)) - \lg B^*(x(n)) < k + F(n). \quad (36)$$

To prove this we will exhibit a specific prefix computer C such that (36) holds when B_C^* is computed with respect to C. For any universal computer, the program lengths for any particular string are at most an additive constant k' longer than those for any other specific computer. As a result, $-\lg B^*$ can only be greater than $-\lg B_C^*$ by no more than the additive constant k'. Therefore proving (36) with respect to any particular prefix computer is equivalent to proving it for a universal computer.

The string $x(n)$ is coded for C in the following way.

(i) We write a prefix code of length k_1 that describes the function $F(\cdot)$.

(ii) We write a prefix code of length k_2 that describes the probability function $P(\cdot)$.

(iii) We write a prefix code for the integer $m = F(n)$. We use a simple code in which m is represented by m 1's followed by a 0.

(iv) The final sequence we write is a Huffman code (which is also a prefix code), for strings of length n', using the probability distribution function $P(\cdot)$. Since each

string has only one code, the shortest code is this unique code. Here n' is the smallest integer such that $F(n')>m$.

We wish to code all strings that are of the form $x(n)z$ where the length of z, $|z|$, is $n'-n$. There are just $2^{n'-n}$ strings of this type for each $x(n)$. The total probability (with respect to $P(\cdot)$) of all such strings is exactly $P(x(n))$, i.e.,

$$\sum_{|z|=n'-n} P(x(n)z)=P(x(n)). \tag{37}$$

The Huffman code for a string of probability P is of length $\lceil -\lg P\rceil$, where $\lceil a\rceil$ is the smallest integer not less than a.

Using our sequence of prefix codes for the string $x(n)z$, we have a total code length of $k_1+k_2+(m+1)+\lceil -\lg P(x(n)z)\rceil$. Then

$$B^*_C(x(n)) \geq \sum_{|z|=n'-n} 2^{-H^c_C(x(n)z)}$$

$$\geq 2^{-k_1-k_2-m-2} \sum_{|z|=n'-n} 2^{1-[-\lg P(x(n)^z)]}$$

where H^c_C is Chaitin's unconditional entropy with respect to machine C. The first inequality follows from (31). From $\lg\ x<1-[-\lg\ x]$ and (37), $2^{-k_1-k_2-m-2}P(x(n))<B^*_C(x(n))$ or $\lg\ P(x(n))-\lg\ B^*_C(x(n))<k_1+k_2+m+2$. Since $m=F(n)$ and $-\lg B^*$ is at most an additive constant greater than $-\lg B^*_C$, the theorem follows directly. Q.E.D.

From Theorems 4 and 5, it is clear that, while $\lg (P/B^*)$ approaches infinity with n, it does so more slowly than any unbounded recursive function of n. In contrast $\lg (P/P'_M)$ is bounded by a constant.

Similarly, if b^* is used in Cover's gambling scheme, the ratio of its yield to the maximum feasible yield is $2^{-k-F(n)}$, in which $F(n)$ approaches infinity arbitrarily slowly. Contrast this with P'_M in which the corresponding ratio is a constant. The expected total square error for b^* is $\ln \sqrt{2}\ (k+F(n))$ in contrast to $k \ln \sqrt{2}$ for P'_M.

A major reason for the deficiency of b^* is its not being normalized in the usual way, i.e.,

$$b^*(0|x(n))+b^*(1|x(n))<1.$$

If we define $b^{*\prime}$ by

$$b^{*\prime}(x_{n+1}|x(n)) \triangleq B^*(x(n)x_{n+1})/B^*(x(n)0)+B^*(x(n)1)$$

then $b^{*\prime}(0|x(n))+b^{*\prime}(1|x(n))=1$. We can define $B^{*\prime}(x(n)) \triangleq \Pi_{i=1}^n b^{*\prime}(x_i|x(i-1))$. Noting from (32) that

$$B^*(x(n))=\prod_{i=1}^n b^*(x_i|x(i-1))$$

it is clear that

$$\frac{B^{*\prime}(x(n))}{B^*(x(n))}=\prod_{i=1}^n \frac{b^{*\prime}(x_i|x(i-1))}{b^*(x_i|x(i-1))}$$

$$=\prod_{i=1}^n \frac{B^*(x(n))}{B^*(x(n)0)+B^*(x(n)1)}>1. \tag{38}$$

This is because from (31) $B^*(x(n))=B^*(x(n)0)+B^*(x(n)1)+2^{-H^c(x(n))}$ for all n. The result is that $B^{*\prime}>B^*$, so (36) is satisfied by $B^{*\prime}$ as well as B^*. However, $B^{*\prime}$ does *not* satisfy (34). On the contrary, for all n,

$$\sum_{i=1}^{2^n} B^{*\prime}(x_i(n))=1. \tag{39}$$

$B^{*\prime}$ is at least as good as B^* in approximating P, but $B^{*\prime}$ is probably *better*, since both $B^{*\prime}$ and P satisfy (39). Though it seems likely that $B^{*\prime}$ is as good as P'_M in approximating computable probability measures, we have not been able to prove this.

VII. Entropy Definitions: K, H^c, and H^*

Kolmogorov's concept of unconditional complexity of a finite string was meant to explicate the amount of information needed to create the string—the amount of programming needed to direct a computer to produce that string as output. His concept of conditional complexity of a finite string x with respect to a string y was the amount of information needed to create x given y.

He proposed that unconditional complexity be defined by

$$K(x(n)) \triangleq \min |r|, \qquad (U(r)=x(n))$$

where U is a universal machine and r is the shortest input to U that will produce the output $x(n)$ and then halt. Conditional complexity is defined by

$$K(x(n)/y(m)) \triangleq \min |r|, \qquad (U(r,y(m))=x(m)).$$

The complexity of the pair of finite strings x and y is defined by $K(x,y)=K(g(x,y))$, $g(x,y)$ being any recursive, information-preserving, nonsingular function from pairs of finite strings to single finite strings.

The entropy equation

$$H(x,y)=H(x)+H_x(y)$$

is of central importance in information theory. Kolmogorov's complexity does not satisfy this equation exactly; rather,

$$K(x,y)=K(y/x)+K(x)+\alpha,$$

and Kolmogorov [9] has shown with the following example that α can be unbounded.

Let $x(n)$ be a random binary string of length n, let ℓ be the integer of which $x(n)$ is the binary expansion, and let $y(\ell)$ be a random string of length ℓ. Then $K(y,x)=\ell+c_1$, $K(y/x)=\ell+c_2$, and $K(x)=n+c_3=\lg\ \ell+c_4$. Here c_1, c_2, c_3, and c_4 are all numbers that remain bounded as $n\to\infty$. From the foregoing, it is clear that $\alpha=K(x,y)-K(y/x)-K(x)=c_5-n$ is unbounded.

On the other hand, Kolmogorov and Levin have shown [8, p. 117, Th. 5.2(b)] that if β is the absolute value of α, then

$$\beta<12|K(xy)|$$

where $|K(\cdot)|$ denotes the length of the string $K(\cdot)$, and xy

is the concatenation of the strings x and y. We see that if x and y are very large, then β is very small relative to them.

Chaitin [3] has shown that his entropy satisfies

$$H^c(x,y)=H^c(x)+H^c(y/x)+k$$

where $H^c(x,y)=H^c(g(x,y))$, $g(x,y)$ being any recursive, information-preserving nonsingular mapping from pairs of finite strings to single finite strings, and k is an integer that remains bounded though x and y may become arbitrarily long.

We now define H^*, a new kind of entropy for finite strings, for which

$$H^*(x,y)=H^*(x)+H_x^*(y)$$

holds exactly. Though H^* is close to the H of information theory, certain of its properties differ considerably from those of Kolmogorov's K and Chaitin's H^c.

Before defining H^*, we will define two associated probability measures, $P'_M(x,y)$ and $P'_{Mx}(y)$. The reasons for these particular definitions and the implied properties of P'_M are discussed in Appendix B. Just as $P'_M(x)$ is the probability of occurrence of the finite string x, $P'_M(x,y)$ is the probability of the *co-occurrence* of both x and y, i.e., the probability that x and y occur simultaneously. The definition is as follows.

If x is a prefix of y, then $P'_M(x,y)=P'_M(y)$.

If y is a prefix of x, then $P'_M(x,y)=P'_M(x)$.

If x is not a prefix of y and y is not a prefix of x, then $P'_M(x,y)=0$ since x and y must differ in certain nonnull symbols, and it is therefore impossible for them to co-occur. This completely defines $P'_M(x,y)$.

$P'_{Mx}(y)$ is the conditional probability of y's occurrence, given that x has occurred. We define

$$P'_{Mx}(y) \triangleq \frac{P'_M(x,y)}{P'_M(x)}. \tag{40}$$

From (40) and the definition of $P'_M(x,y)$, the following is clear.

If x is not a prefix of y, and y is not a prefix of x, then $P'_{Mx}(y)=0$.

If y is a prefix of x, then $P'_{Mx}(y)=1$.

If x is a prefix of y, then $P'_{Mx}(y)=(P'_M(y)/P'_M(x))$, for in this case y is of the form xa and $P'_{Mx}(y)$ is the probability that if x has occurred a will immediately follow. Following Willis [2, Section 4, pp. 249–254] we define

$$\begin{aligned} H^*(x) &\triangleq -\lg P'_M(x) \\ H^*(x,y) &\triangleq -\lg P'_M(x,y) \\ H_x^*(y) &\triangleq -\lg P'_{Mx}(y). \end{aligned} \tag{41}$$

From (40) and (41), we directly obtain the desired result that

$$H^*(x,y)=H^*(x)+H_x^*(y).$$

The properties of $H_x^*(y)$ differ considerably from those of $H^c(y/x)$ and $K(y/x)$. Suppose x is an arbitrary finite string and $y=f(x)$ is some simple recursive function of x —say y is the complement of x, $(0\to1,\ 1\to0)$. Then $H^c(y/x)$ and $K(y/x)$ are bounded and usually small. They are both something like the additional information needed to create y, if x is known. $H_x^*(y)$ has no such significance. If x and y are complements, then $P'_{Mx}(y)=0$ (since neither can be the prefix of the other) and $H_x^*(y)=\infty$.

The differences between the various kinds of entropy may be explained by differing motivations behind their definitions. $P'_M(x)$ was devised in an attempt to explicate the intuitive concept of probability. The definitions of $P'_M(x,y)$ and $P'_{Mx}(y)$ were then derived from that of $P'_M(x)$ in a direct manner.

$H^c(y,x)$ and $K(y/x)$ were devised to explicate the additional information needed to create y, given x. The definitions of $H^c(x)$, $K(x)$, etc., were directly derived from those of $H^c(y/x)$ and $K(y/x)$, respectively.

We will next investigate the properties of H^*, K, and H^c when applied to very long sequences of stochastic ensembles and compare them to associated entropies.

Levin states [8, p. 120, Proposition 5.1] that for an ergodic ensemble,

$$\lim_{n\to\infty} \frac{K(x(n))}{n} = H \quad \text{with Pr 1.} \tag{42}$$

If the ensemble is stationary but not ergodic, the statement is modified somewhat so that H varies over the ensemble. Unfortunately, no proof is given, and it is not stated whether or not the ensemble must have a computable probability measure.

Cover has shown [5] that if (42) is true then it follows that for an ergodic process

$$\lim_{n\to\infty} \frac{1}{n} H^c(x(n)) = H \quad \text{with Pr 1.}$$

Leung-Yan-Cheong and Cover [4, last Theorem] have shown that for any stochastic process definable by a computable probability measure P,

$$H_n \leqslant E_P H^c(X(n)/n) \leqslant H_n + k \tag{43}$$

where H_n is the entropy of the set of strings of length n:

$$H_n \triangleq \sum_{i=1}^{2^n} P(x_i(n)) \lg P(x_i(n)),$$

$$E_P H^c(X(n)/n) \triangleq \sum_{i=1}^{2^n} P(x_i(n)) H^c(x_i(n)/n),$$

and k is a constant that depends on the functional form of P but is independent of n. If P defines an ergodic process, then

$$\lim_{n\to\infty} \frac{1}{n} H_n = H,$$

the entropy of the ensemble. In this case from (43) we obtain

$$\lim_{n\to\infty} \frac{1}{n} E_P H^c(X(n)) = H. \tag{44}$$

Theorem 6: For any stochastic process definable by a computable probability measure P,

$$H_n \leqslant E_P H^*(X(n)) \leqslant H_n + k \tag{45}$$

where

$$E_P H^*(X(n)) \triangleq \sum_{i=1}^{2^n} P(x_i(n)) H^*(x_i(n)),$$

and k is a constant, independent of n, but dependent upon the functional form of P.

To prove this, note that from Theorem 2,

$$-\lg P'_M(x(n)) \leqslant -\lg P(x(n)) + k.$$

Therefore

$$-\sum_{i=1}^{2^n} P(x_i(n)) \lg P'_M(x_i(n)) \leqslant -\sum_{i=1}^{2^n} P(x_i(n)) \lg P(x_i(n)) + k$$

and

$$E_P H^*(X(n)) \leqslant H_n + k. \tag{46}$$

From Lemma 2 of Theorem 4,

$$\sum_{i=1}^{2^n} P(x_i(n)) \lg P'_M(x_i(n))$$

has maximum value when

$$P'_M(x_i(n)) = P(x_i(n)),$$

so

$$-\sum_{i=1}^{2^n} P(x_i(n)) \lg P(x_i(n)) \leqslant -\sum_{i=1}^{2^n} P(x_i(n)) \lg P'_M(x_i(n))$$

and

$$H_n \leqslant E_P H'(X(n)). \tag{47}$$

The theorem follows directly from (46) and (47). As we noted in (44), if P is ergodic,

$$\lim_{n\to\infty} \frac{1}{n} E_P H^c(X(n)) = H.$$

Theorem 7: If

$$F(n) \triangleq E_P(\lg P(X(n)) + H^c(X(n))) = -H_n + E_P(H^c(X(n))), \tag{48}$$

then $\lim_{n\to\infty} F(n) = \infty$.

Lemma 1:

$$\lim_{n\to\infty} \sum_{k=1}^{2^n} 2^{-H^c(x_k(n))} = 0.$$

This lemma is a direct consequence of the Kraft inequality from which

$$\sum_{n=1} \left[\sum_{k=1}^{2^n} 2^{-H^c(x_k(n))} \right] \leqslant 1.$$

To prove the Theorem we first rewrite (48) as

$$F(n) = E_P(\lg P(x(n)) - \lg (2^{-H^c(x(n))})).$$

The theorem is then proved via the arguments used to establish Theorem 4. Q.E.D.

Comparison of Theorem 7 with (43) and (45) suggests that $EH^*(X(n))/n$ and $EH^c(X(n)/n)/n$ approach H more rapidly than does $EH^c(X(n))/n$. A more exact comparison can be made if a bound is known for the rate at which $E(-\lg P(X(n)))/n$ approaches H.

Acknowledgment

We are indebted to G. Chaitin for his comments and corrections of the sections relating to his work. In addition to developing many of the concepts upon which the paper is based, D. Willis has been helpful in his discussion of the definition of H^* and the implied properties of P'_M. We want particularly to thank T. Fine for his extraordinarily meticulous analysis of the paper. He found several important errors in an early version and his incisive criticism has much enhanced both the readability and reliability of the paper.

Appendix A

Let $[\alpha_m]$ be the set of all minimal codes for $x(i)$, and let $[\beta_{mj}]$ for fixed m be the set of all finite (or null) strings such that $\alpha_m\beta_{mj}$ is either a minimal code for $x(i)0$ or for $x(i)1$. Then $[\beta_{mj}]$ for fixed m forms a prefix set, so

$$\sum_j 2^{-|\beta_{mj}|} \leqslant 1. \tag{49}$$

By definition

$$\tilde{P}'_M(x(i)) = \sum_m 2^{-|\alpha_m|}, \tag{50}$$

$$P'_M(x(i)0) + P'_M(x(i)1) = \sum_m \sum_j 2^{-|\alpha_m\beta_{mj}|} = \sum_m 2^{-|\alpha_m|} \sum_j 2^{-|\beta_{mj}|}. \tag{51}$$

From (49), (50), and (51),

$$\tilde{P}'_M(x(i)) \geqslant \tilde{P}'_M(x(i)0) + \tilde{P}'_M(x(i)1).$$

Q.E.D.

Appendix B

Our definitions of $P'_M(x)$, $P'_M(x,y)$, and $P'_{Mx}(y)$ correspond to Willis' definitions of $P^R(x)$, $P^R(x,y)$, and $P^R_x(y)$, respectively. Willis regards $P^R(x(n))$ as a measure on the set of all infinite strings that have the common prefix $x(n)$. This measure on sets of infinite strings is shown to satisfy the six axioms [2, pp. 249, 250], [10, chap. 1 and 2] that form the basis of Kolmogorov's axiomatic probability theory [10].

We can also regard $P'_M(x(n))$ as being a measure on sets of infinite strings in the same way. It is easy to show that the first five postulates hold for this measure. From these five, Kolmogorov [10, Chapter 1] shows that joint probability and conditional probability can be usefully defined and that Bayes' Theorem and other properties of them can be rigorously proved. Our definitions of $P'_M(x,y)$ and $P'_{Mx}(y)$ are obtained from his definitions of joint and conditional probabilities, respectively.

A proof that this measure satisfies the sixth postulate (which corresponds to countable additivity) would make it possible to apply Kolmogorov's complete axiomatic theory of probability to P'_M. While it seems likely that the sixth postulate is satisfied, it remains to be demonstrated.

REFERENCES

[1] R. J. Solomonoff, "A formal theory of inductive inference," *Inform. and Contr.*, pp. 1–22, Mar. 1964, and pp. 224–254, June 1964.

[2] D. G. Willis, "Computational complexity and probability constructions," *J. Ass. Comput. Mach.*, pp. 241–259, Apr. 1970.

[3] G. J. Chaitin, "A theory of program size formally identical to information theory," *J. Comput. Mach.*, vol. 22, no. 3, pp. 329–340, (July 1975).

[4] S. K. Leung-Yan-Cheong and T. M. Cover, "Some inequalities between Shannon entropy and Kolmogorov, Chaitin and extension complexities," *Tech. Rep. 16*, Statistics Dept., Stanford Univ., Stanford, CA, 1975.

[5] T. M. Cover, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Rep. 12, Statistics Dept., Stanford Univ., Stanford, CA, 1974.

[6] R. J. Solomonoff, "Inductive inference research status," RTB-154; Rockford Research Inst., July 1967.

[7] J. Koplowitz, "On countably infinite hypothesis testing," presented at IEEE Sym. Inform. Theory, Cornell Univ., Oct. 1977.

[8] A. K. Zvonkin, and L. A. Levin, "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms," *Russ. Math. Surys.*, vol. 25, no. 6, pp. 83–124, 1970.

[9] A. N. Kolmogorov, "On the algorithmic theory of information," Lecture, Int. Symp. Inform. Theory, San Remo, Italy, Sept. 15, 1967. (Example given is from the lecture notes of J. J. Bussgang. Kolmogorov's paper, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 5, Sept. 1968, pp. 662–664, was based on this lecture, but did not include this example.)

[10] A. N. Kolmogorov, *Foundations of the Theory of Probability*. New York: Chelsea, 1950.

[11] B. D. Kurtz, and P. E. Caines, "The recursive indentification of stochastic systems using an automaton with slowly growing memory," presented at IEEE Sym. Inform. Theory, Cornell Univ., Oct. 1977.

[12] T. M. Cover, "On the determination of the irrationality of the mean of a random variable," *Ann. Statis.*, vol. 1, no. 5, pp. 862–871, 1973.

[13] T. L. Fine, Personal correspondence.

[14] K. L. Schubert, "Predictability and randomness," Tech. Rep. TR 77-2, Dept. of Computer Science, Univ. Alberta, AB, Canada, Sept. 1977.

Block Coding for an Ergodic Source Relative to a Zero-One Valued Fidelity Criterion

JOHN C. KIEFFER

***Abstract*—An effective rate for block coding of a stationary ergodic source relative to a zero-one valued fidelity criterion is defined. Under some mild restrictions, a source coding theorem and converse are given that show that the defined rate is optimum. Several examples are given that satisfy the restrictions imposed. A new generalization of the Shannon–McMillan Theorem is employed.**

I. INTRODUCTION

LET $(A, \mathcal{F})$ be a measurable space. A will serve as the alphabet for our source. For $n=1,2,\cdots$ $(A^n, \mathcal{F}_n)$ will denote the measurable space consisting of A^n, the set of all sequences $(x_1,x_2,\cdots,x_n)$ of length n from A, and $\mathcal{F}_n$, the usual product σ-field. $(A^\infty, \mathcal{F}_\infty)$ will denote the space consisting of A^∞, the set of all infinite sequences $(x_1,x_2,\cdots)$ from A, and the usual product σ-field $\mathcal{F}_\infty$. Let $T_A: A^\infty \to A^\infty$ be the shift transformation $T_A(x_1,x_2,\cdots)=(x_2,x_3,\cdots)$. We define our source μ to be a probability measure on A^∞, which is stationary and ergodic with respect to T_A.

Manuscript received February 14, 1977; revised November 1, 1977.

The author is with the Department of Mathematics, University of Missouri, Rolla, MO 65401

Suppose for each $n=1,2,\cdots$, we are given a jointly measurable distortion measure $\rho_n: A^n \times A^n \to [0,\infty)$. We wish to block code μ with respect to the fidelity criterion $F=\{\rho_n\}_{n=1}^\infty$. Most of the results about block coding a source require a single letter fidelity criterion [1, p. 20]. An exception is the case of noiseless coding [2, Theorem 3.1.1]. In this case, we have $\rho_n(x,y)=0$ if $x=y$ and $\rho_n(x,y)=1$ if $x\neq y$. In this paper we consider a generalization of noiseless coding, where we require each distortion measure ρ_n in F to be *zero-one valued*; that is, zero and one are the only possible values of ρ_n allowed. Such a fidelity criterion F we will call a zero-one valued fidelity criterion.

We will impose throughout the paper the following restriction on our zero-one valued fidelity criterion $F=\{\rho_n\}$.

R1: If $\rho_m(x,y)=0$ and $\rho_n(x',y')=0$, then

$$\rho_{m+n}((x,x'),(y,y'))=0,\ m,n=1,2,\cdots.$$

In the preceding, we mean (x,x') to represent the sequence of length $m+n$ obtained by writing first the terms of x, then the terms of x'. Equivalently, $R1$ says $\rho_{m+n}((x,x'),\ (y,y')) \leqslant \rho_m(x,y)+\rho_n(x',y')$. $R1$ is a con-

0018-9448/78/0700-432$00.75