Predictability and Randomness

Lenhart K. Schubert
University of Rochester

Abstract. Algorithmic theories of randomness can be related to theories of probabilistic sequence prediction through the notion of a predictor, defined as a function which supplies lower bounds on initial-segment probabilities of infinite sequences. An infinite binary sequence z is called unpredictable iff its initial-segment "redundancy" $n+\log p(z(n))$ remains sufficiently low relative to every effective predictor p. A predictor which maximizes the initial-segment redundancy of a sequence is called optimal for that sequence. It turns out that a sequence is random iff it is unpredictable. More generally, a sequence is random relative to an arbitrary computable distribution iff the distribution is itself an optimal predictor for the sequence. Here "random" can be taken in the sense of Martin-Löf by using weak criteria of effectiveness, or in the sense of Schnorr by using stronger criteria of effectiveness. Under the weaker criteria of effectiveness it is possible to construct a universal predictor which is optimal for all infinite sequences. This predictor assigns nonvanishing limit probabilities precisely to the recursive sequences. Under the stronger criteria of effectiveness it is possible to establish a law of large numbers for sequences random relative to a computable distribution, which may be useful as a criterion of "rationality" for methods of probabilistic prediction. A remarkable feature of effective predictors is the fact that they are expressible in the special form first proposed by Solomonoff. In this form sequence prediction reduces to assigning high probabilities to initial segments with short and/or numerous encodings. This fact provides the link between theories of randomness and Solomonoff's theory of prediction.

Preface

This article makes available an extended study of the theoretical relationship between predictability and randomness, for many years available only as a technical report in the Computing Science Department of the University of Alberta (TR77-2, September 1977, now no longer available). The typography in the TR was poor, as text formatting had not come of age yet. In essence the article shows that definitions of (non)randomness for infinite sequences in terms of computational nonrandomness tests, predictability, and "compressibility" through encodings as programs are equivalent.

I conducted the research leading to the report from 1975-1977, unaware except in the last months of the work that a 1970 article in a Soviet journal had reported many of "my" theorems, without proofs (Zvonkin & Levin, 1970). I accordingly annotated those theorems with (Leonid) Levin's name, before submitting the work for publication. The preempted theorems were limited to the semicomputable characterization of randomness, while my manuscript also covered the recursively computable characterization by C.P. Schnorr. However, while the manuscript was under review, a treatment of the latter characterization appeared in an article by that author.* The reviewer mentioned the possibility of extending the article to a more complete survey, but I found it difficult to contemplate presenting my hard-won results as mostly a survey of results by others. Thus, the report languished as a TR, even though all the proofs were new, and a few unpublished propositions remained. Certain other researchers in this area later suggested to me that the missing proofs in the Zvonkin & Levin survey should really have entitled me to publication and co-discoverer status for at least the semicomputable results. In any event, the existence of the arXiv system has made it possible to make the original version easily accessible.

The apparently new results in the submitted article were Theorems 2-4 (on the extent to which semicomputable measures actually allow probabilistic prediction), Th. 6 (which seems to slightly strengthen the previously known result that there is no recursive universal distribution – "probabilities" assigned by "predictors" don't have to add up to 1), Th. 7 (about a nearly optimal *additive* "predictor" – a nontrivial result), Th. 11 (a kind of law of large numbers, which the reviewer said was implicit in some of Schnorr's published work, a comment that I did not succeed in confirming), and Th. 13, with Corollaries 1 & 2 (though Levin had proved very closely related results).**

I include the symbol glossary that prefaced the TR, even though the meanings of the symbols are mostly clear from the text.

---

*Claus-Peter Schnorr, & P. Fuchs, "General Random Sequences and Learnable Sequences," *J. Symb. Logic 42*(3), pp. 329-340 (1977); also,

C.P. Schnorr, "A survey of the theory of random sequences', in R.E. Butts and J. Hintikka (eds.), *Basic Problems in Methodology and Linguistics*, Dordrecht: D. Reidel, pp. 193–210 (1977).

**In 1988 I communicated these points, along with the TR, to Professor Levin, by then at Boston University.

Symbol Glossary

Symbol	Meaning
0,1	unit strings; or numbers (clear from context)
N	{0,1,2,...}
$\mathbb{R}$	the nonnegative reals
$\mathbb{R}^+$	the positive reals
R	the real interval [0,1]
B	the numbers in R with finite radix-2 representations
Q	the rational numbers in [0,1]
X	the 2-element alphabet {0,1}
X^*	the concatenation closure of X
X^∞	the semi-infinite binary sequences
upper case Latin letters other than B, N, R, Q, X	subsets of X^* or $X^* \times \mathbb{R}$; or, procedure variables
dom	domain
$f(S)$	$\{f(x) \mid x \in S\}$
$f^{-1}(S)$	$\{x \mid f(x) \in S\}$
log	base 2 logarithm
p, p', p'', p_i	predictors or conditional predictors
$p^*(x)$	surplus probability of $x =_{df} p(x) - p(x0) - p(x1)$
p_f	the Solomonoff predictor determined by a process f
pf	prefix-free
re	recursively enumerable
$rp(x)$	$\lvert x\rvert + \log p(x)$
x and other lower case letters near the end of the Latin alphabet	binary sequence
$x(n)$	prefix of length n of binary sequence x
x^-	finite binary sequence x with last digit complemented
$x(n)^-$	$x(n)$ with nth digit complemented
δ	an element of $\mathbb{R}^+$
λ	Church's lambda operator
Λ	the null sequence
μ	a measure on subsets of X^∞
σ	$\sigma S = \sum_{x \in S} 2^{-\lvert x\rvert}$
ϕ	partial recursive function from N to N or from $X^* \times N$ to Q
$\emptyset$	the empty set
$\times$	Cartesian product
$\sqsubseteq$	is a prefix of
$\sqsubset$	is a proper prefix of
$\sqsupseteq$	is an extension of
$\sqsupset$	is a proper extension of

$(\ )$	open interval; syntactic delimiters
$[\ ]$	closed interval; assertion delimiters (numerical value 1 or 0 corresponding to true or false when used as arithmetic expression)
$\langle\ \rangle$	ordered pair
$\succcurlyeq$	is an encoding of
$\gg$	is a reduced encoding of

1. Introduction

The decade beginning in 1963 saw the development of two types of computational theories for infinite sequences. Theories of the first type are concerned with the algorithmic distinction between random and nonrandom sequences, while theories of the second type are concerned with prediction of infinite sequences or inductive discovery of programs for them.

Relatively little attention has been paid to the connections between these two lines of development, even though the existence of such connections has always been apparent. Indeed, von Mises' (1919) original proposal for characterizing random sequences involved a notion similar to prediction, viz., *a priori* selection of digits from an infinite binary sequence. Von Mises' proposal, taken up by Wald (1937) and Church (1940) among others, did not lead to a satisfactory characterization of random sequences (see the critique of Ville, 1939). The later work of Kolmogorov (1965), Martin-Löf (1966), Chaitin (1966), and Schnorr (1971) at last yielded several apparently successful approaches to this problem. However, none of these approaches turned explicitly upon any proper notion of sequence prediction (although the stakes wagered in Schnorr's "gambling strategies" could be viewed as implicit predictions).

On the other hand, the work on inductive inference was not directly concerned with the definition of randomness. Solomonoff (1964) proposed several classes of methods for predicting sequences probabilistically, and Willis (1970) showed that one of these classes contains approximations to all recursive sequential probability distributions. Cover (1974), like Schnorr (1971), investigated sequential gambling schemes. He explicitly related them to prediction schemes and devised an interesting variant of one of Solomonoff's universal prediction schemes. Subsequently Solomonoff (1976) proved convergence and other desirable properties for his original universal predictor.

The studies most directly concerned with the relationship between prediction and randomness are those of Chaitin and Levin. Chaitin (1975) defined randomness in terms of Solomonoff-like probabilities, and has asserted (Chaitin, 1977) that his definition is equivalent to that of Martin-Löf (1966). At first sight Chaitin's probabilities seem unsuitable for infinite sequence prediction: the probability assigned to an initial segment may exceed the probabilities of shorter initial segments. However, it seems clear in retrospect that Chaitin's probabilities could have been used as a basis for the present study. The probability of a sequence as discussed herein apparently corresponds to the sum of Chaitin's probabilities over all finite extensions of the sequence.

The Soviet[†] mathematician L. A. Levin made major contributions to the unification of the theories of randomness and prediction in a series of papers sparked by Levin's association with Kolmogorov (Zvonkin & Levin, 1970, and Levin, 1973, 1976). Levin introduced the notion of a semicomputable measure, which can be viewed intuitively as a

[†]at the time

method of probabilistic sequence prediction. The theorems of Levin in Zvonkin and Levin (1970) show, in effect, that semicomputable measures are expressible in the form proposed by Solomonoff (1964), although Levin did not explicitly attach this interpretation to his results. Furthermore, he established analogous results for computable measures (similar to the results obtained independently by Willis (1970)). Subsequently Levin (1973) made the crucial connection between semicomputable measures and randomness, stating a theorem to the effect that a sequence is random in the sense of Martin-Löf iff it is irredundant with respect to every semicomputable measure; indeed, he found more generally that a sequence is Martin-Löf random *relative* to any given computable distribution iff its redundancy as measured by any semicomputable measure is no greater than its redundancy relative to the given distribution (apart from a constant). Later Levin (1976) further generalized these results to sequences which are random relative to arbitrary (not necessarily computable) measures, and related them to information theory.

The results of the present paper were obtained before the author became aware of Levin's work. The central concern is with sequence prediction in the sense of prior ("subjective") probability assignments to initial segments of infinite sequences. The objective is to relate this notion of prediction to definitions of randomness due to Martin-Löf and Schnorr on the one hand and to Solomonoff's ideas about prediction on the other. Several of the main results of Secs. 3 and 5 are contained in the cited papers of Levin. The presentation of new proofs is justified in part by the differences in approach (e.g., the construction of an optimal predictor in Th. 8 without reduction of predictors to Solomonoff's form) and in part by the fact that Levin did not publish proofs for all of his results (e.g., the connection between predictability and Martin-Löf randomness, Th. 5). The present paper is more explicitly concerned with sequence prediction than Levin's studies; the terminology and techniques reflect this concern.

A topic not treated here is the extrapolation of recursive sequences or inductive discovery of programs for such sequences (e.g., Gold, 1967, and Blum & Blum, 1973). Although nonprobabilistic extrapolation of recursive sequences can be viewed as a special case of probabilistic prediction, the results herein are of too general a nature to shed any new light on this special case.

Sec. 2 introduces the formal notation and some basic concepts. An incrementable predictor (cf. Levin's semicomputable measure) is defined as a lower bound on a sequential probability distribution which is approachable from below. Thus the class of incrementable predictors contains all recursive methods of prediction, as well as certain nonrecursive methods. Alternative intuitive interpretations of predictors are considered, and some computability properties of prediction schemes based on incrementable predictors are examined.

In Sec. 3 it is shown that any infinite binary sequence z is Martin-Löf random iff its initial-segment redundancy $n + \log p(z(n))$ is bounded relative to every incrementable predictor p. Actually this is established as a corollary of the fact that z is Martin-

Löf random *relative* to distribution p iff p is *optimal* for z, where optimality means maximization of initial-segment redundancy. An optimal *universal* predictor is then constructed, i.e., one which maximizes the initial-segment redundancy of every infinite binary sequence. This predictor assigns nonvanishing limit probabilities precisely to the recursive sequences.

In Sec. 4 attention is restricted to recursive predictors. It is shown that an infinite sequence is Schnorr-random iff its initial-segment redundancy does not grow "noticably" (in a suitable sense) relative to any recursive predictor. Again this is a corollary of a result about Schnorr-randomness *relative* to a recursive predictor, viz., that z is Schnorr-random relative to p iff p is "weakly optimal" for z. A related fact is that a recursive predictor maximizes the initial-segment redundancy of a sequence only if the conditional probabilities it assigns to events occurring in that sequence agree with the frequencies of those events. For example, for about 70% of the cases where a 1-digit is predicted with 70% conditional probability, a 1-digit actually occurs. Thus an optimal predictor exhibits a type of consistency which seems desirable in any sequential inductive method.

In Sec. 5 it is shown that the class of incrementable predictors coincides with one of Solomonoff's classes of predictive methods based on program lengths. Also some variants of Willis' 1970) and Levin's (Zvonkin & Levin, 1970) results about the reduction of recursive predictors to Solomonoff's form are presented.

2. Predictors

The following basic notation and terminology will be used. N is the set of natural numbers including 0, $\mathbb{R}$ is the set of nonnegative real numbers, $\mathbb{R}^+$ is $\mathbb{R} - \{0\}$, R is the real interval [0,1], Q is the set of rational numbers in R, B is the set of numbers in R with finite radix-2 representations, $X = \{0,1\}$, X^* is the set of finite binary sequences including the null sequence Λ, and X^∞ is the set of infinite binary sequences. $|x|$ is the length (number of digits) of a sequence $x \in X^*$. If $|x| = n$ then x is said to be an n-sequence. The notation $x \sqsubseteq y$ (or $y \sqsupseteq x$) expresses that x is a prefix of y (y is an extension of x), where $x \in X^*$ and $y \in X^* \cup X^\infty$. Similarly $x \sqsubset y$ (or $y \sqsupset x$) expresses that x is a proper prefix of y, i.e., $x \sqsubseteq y$ and $x \neq y$. A prefix of length n of a finite or infinite sequence x is denoted by $x(n)$ ($x(n)$ is undefined if $n > |x|$). The concatenation of two sequences x and y is written as xy. Similarly $\{xy|x \in S, y \in T\}$ is written as ST. Set concatenations involving singletons (e.g., $\{x\}X^*$) are shortened by omitting braces of the singletons (e.g., xX^*). Also n-fold self-concatenation of a sequence x or set of sequences S is written as x^n or S^n respectively. Note that $\emptyset S = S\emptyset = \emptyset$. A set $S \subset X^*$ is *prefix-free* (pf) iff $S \cap (SXX^*) = \emptyset$, i.e., it contains no proper extension of any of its members. Such a set is also called an *instantaneous code* (Abramson, 1963).

A *predictor* is a total function $p : X^* \to R$ such that $p(x) \geq p(x0) + p(x1)$ for all $x \in X^*$. Thus p corresponds either to a subadditive measure μ on X^∞ such that μxX^∞

$= p(x)$ or to an additive measure μ on $X^* \cup X^\infty$ such that $\mu x(X^* \cup X^\infty) = p(x)$, for all $x \in X^*$. Intuitively $p(x)$ may be regarded as the prior probability of x or as a lower bound on its prior probability (see below). The difference $p^+(x) = p(x) - p(x0) - p(x1)$ is called the *surplus* probability of x. A predictor satisfying $p(\Lambda) = 1$ and $p(x) = p(x0) + p(x1)$ for all $x \in X^*$ is called a *sequential probability distribution* (Martin-Löf, 1966), or a distribution, for short.[1]

A total function $f : X^* \to R$ is *incrementable* iff there is a recursive function $g : X^* \times N \to Q$ which is nondecreasing in its second argument such that
$f(x) = \lim_n g(x, n)$ for all $x \in X^*$;
i.e., $f(x)$ is approachable from below, with each increase in n supplying a nonnegative rational increment in the approximation $g(x, n)$ to $f(x)$.[2] When a recursive function g and a function f are related as above, g is said to *underlie* f.

One of the primary concerns in this paper will be the class of incrementable predictors.[3] The importance of this class of predictors lies in its relationship to the class of Martin-Löf random sequences on the one hand (Sec. 3) and to the class of processes on the other (Sec. 5).

The following two simple facts about predictors are noteworthy.

Theorem 1. (a) Every incrementable distribution is recursive.
(b) Every recursive predictor can be increased to a recursive distribution.

Proof. (a) If p is any incrementable distribution then $p(\Lambda)$ is trivially computable. Assume for induction on n that $p(x)$ is computable for every n-sequence x. Now from $p(x0) = p(x) - p(x1)$ it is seen that $p(x0)$ is approachable from above, since $p(x)$ is computable by assumption and $p(x1)$ is approachable from below. But $p(x0)$ is also approachable from below, so that $p(x0)$ is computable; similarly for $p(x1)$.
(b) For any recursive predictor p, a distribution p' such that $p'(x) \geq p(x)$ for all $x \in X^*$ can be defined as follows: Let $p'(\Lambda) = 1, p'(x0) = p'(x) - p(x1)$, and $p'(x1) = p(x1)$ for all $x \in X^*$. Then it is easily verified by induction on sequence length that p′ meets the requirements of the theorem. □

In what sense and under what conditions does a predictor allow sequence prediction? If the predictor is a recursive distribution the answer is straightforward. Consider any *nonterminating* process which generates a succession of binary digits; then $p(xy)/p(x)$

[1] Solomonoff (1964) used the term "normalized probability evaluation methods" for computable distributions. Schnorr (1971) defined randomness in terms of martingales, where a martingale $f : X^* \to \boldsymbol{R}^+$ satisfies $f(x) = (f(x0) + f(x1))/2$ for all $x \in X^*$. Thus if $f(x) \leq 2^{|x|}$ for all $x \in X^*$, then $2^{-|x|} f(x)$ defines a sequential probability distribution. See Sec. 4.

[2] It is assumed that procedures which accept rational numbers as inputs or generate them as outputs utilize some effective encoding of the rational numbers, e.g., integer pairs $\langle m, n \rangle$ such that $m/n = q$. Instead of the rational numbers a more restricted set such as B (the numbers with finite radix-2 representations), or a less restricted set such as the computable numbers in R could have been used.

[3] These correspond exactly to Levin's semicomputable measures (Zvonkin & Levin, 1970), apart from the inessential condition $p(\Lambda) = 1$, i.e., $\mu X^* \cup X^\infty = 1$, on any semicomputable measure μ.

can be regarded as the conditional probability that x will be followed by y, given that x has occurred (replace any ratio 0/0 by 0). Thus a "prediction" of a sequence continuation is analogous to a weather forecast, say, which attributes a probability to some future weather condition (e.g., "60% chance of rain tomorrow").[4]

Arbitrary predictors, however, admit two intuitive interpretations, corresponding to the two measure-theoretic interpretations mentioned above. In the first interpretation a predictor supplies *lower bounds* on prior probabilities of initial output sequences generated by a *nonterminating* process. Corresponding upper and lower bounds on conditional probabilities are supplied in Th. 2. These are approachable from above and below respectively, whenever the given predictor is incrementable (Th. 3).

In the second interpretation a predictor supplies prior probabilities on nonempty output sequences of a process *which may or may not terminate*. The surplus probability of a sequence $x \in XX^*$ is then the probability that the process will generate x and halt. As in the case of distributions, $p(xy)/p(x)$ is the probability that y will follow x, given that x has been generated, but with no guarantee that a continuation of length $|y|$ will be generated at all. These conditional probabilities need not be approachable from below, even if the given predictor is incrementable (Th. 4).

The following theorem gives the sharpest possible bounds on conditional probabilities implicit in the values of a predictor, when these are interpreted as lower bounds on initial-segment probabilities in a nonterminating process (first interpretation). For any $x \in XX^*$, x^- denotes the sequence obtained by changing the last digit of x to its complement. Thus $v(i)^-$ is $v(i)$ with the ith digit complemented. A sum over no terms (in particular, a sum from a higher to a lower summation index) is taken to be 0. As before, occurrences of 0/0 are to be replaced by 0.

Theorem 2. If p is any predictor and p' is any distribution such that $\forall y \in X^* : p'(y) \geq p(y)$, then $\forall v \in X^* : \forall w \in XX^*$:

$$\frac{p(vw)}{1 - \sum_{i=1}^{|v|} p(v(i)^-)} \leq \frac{p'(vw)}{p'(v)} \leq \frac{1 - \sum_{i=1}^{|vw|} p((vw)(i)^-)}{1 - \sum_{i=1}^{|v|} p(v(i)^-)}.$$

Furthermore, these are the sharpest possible bounds derivable from p in the sense that $\forall v \in X^* : \forall w \in XX^*$: $\exists$ distributions p', p'': $\forall x \in X^* : \; p'(x), p''(x) \geq p(x)$ and

$$\frac{p'(vw)}{p'(v)} = \frac{p(vw)}{1 - \sum_{i=1}^{|v|} p(v(i)^-)} \quad \text{and} \quad \frac{p''(vw)}{p''(v)} = \frac{1 - \sum_{i=1}^{|vw|} p((vw)(i)^-)}{1 - \sum_{i=1}^{|v|} p(v(i)^-)}.$$

Proof. The lower bound on $p'(vw)/p'(v)$ is obtained from the lower bound $p(vw)$ on $p'(vw)$ and upper bound $p'(\Lambda) = \sum_{i=1}^{|v|} p(v(i)^-)$

[4] The "rationality" of such predictions depends on the frequency with which events assigned particular conditional probabilities occur; see discussion preceding Th. 11.

on $p'(v)$, which is easily inferred from the distribution property of p' and the fact that each $p(v(i)^-)$ is a lower bound on $p'(v(i)^-)$, for $i = 1, ..., |v|$. The upper bound on $p'(vw)/p'(v)$ is obtained by noticing that the difference between $p'(v)$ and $p'(vw)$ is at least

$\sum_{i=|v|+1}^{|vw|} p((vw)(i)^-)$,

so that $p'(vw)/p'(v)$ is maximized by choosing $p'(vw)$ as large as possible while keeping $p'(v) - p'(vw)$ to its minimum.

The second part of the theorem is proved by constructing p' such that

$p'(x) = 1 - \sum_{i=1}^{|x|} p(x(i)^-)$, and
$p'(x^-) = p(x^-)$ for all $x \sqsubset vw$;
$p'(vw) = p(vw)$; and
$p'(vw^-) = 1 - \sum_{i=1}^{|vw|-1} p((vw)(i)^-) - p(vw)$.

Then p' is easily seen to be a distribution bounded below by p for all x and x^- such that $x \sqsubseteq vw$; its extension to other $x \in X^*$ is straightforward. Evidently $p'(vw)/p'(v)$ equals the lower bound of the theorem. Similarly p'' is constructed such that

$p''(x) = 1 - \sum_{i=1}^{|x|} p(x(i)^-)$, and
$p''(x^-) = p(x^-)$ for all $x \sqsubseteq vw$.

Again p'' is easily seen to be a distribution bounded below by p for all x and x^- such that $x \sqsubseteq vw$; its definition is easily completed. Evidently $p''(vw)/p''(v)$ equals the upper bound of the theorem. □

Theorem 3. If the predictor p of Th. 2 is incrementable, then the upper and lower bounds on conditional probability of that theorem are approachable from above and below respectively.

Proof. Since p is approachable from below, the given lower bound on $p'(vw)/p'(v)$ is approachable from below. Inspection of the upper bound indicates that whenever $p(x^-)$ is incremented for some $x \sqsubseteq vw$, the numerator will decrease while the denominator will either decrease by the same amount or remain unchanged. In neither case does the upper bound increase, since the numerator is at most as large as the denominator. □

Thus the conditional predictor determined by an incrementable predictor under the first interpretation is itself incrementable. This is not the case under the second interpretation, i.e., the probability $p'(x, y)$ that y follows x is not approachable from below, as the following theorem shows. The theorem also deals with the non-incrementability, under the second interpretation of predictors, of the probability $p''(x, y)$ that y follows x, given that the process does *not* terminate prematurely (i.e., after generating $xy' \sqsubset xy$). This probability would be of interest if it were known that a continuation of length $\geq |y|$ had been generated, but its digits were still unknown. Solomonoff's (1964, 1976) and Cover's (1974) conditional probabilities are of this type.

Theorem 4. If p is a predictor then p′ defined as

$p'(x,y) = p(xy)/p(x)$ for all $x, y \in X^*$
and p'' defined as
$p''(x,\Lambda) = p(x)/p(x)$,
$p''(x,w) = p(xw)/(p(xw) + p(xw^-))$, and
$p''(x,yw) = p''(x,y)\,p''(xy,w)$ for all $x, y \in X^*$ and $w \in X$,
are not in general approachable from below.

Proof. Let $\phi_i : N \to N, i = 0, 1, 2, ...$ be a recursive enumeration of the partial recursive functions. Let
$p(0^m) = 2^{-\min\{n | n \geq m,\ \phi_n \text{ defined}\}}$ for all $m \in N$, and
$p(x) = 0$ for all $x \in X^*1X^*$.
Clearly p is a predictor. Since $\{n | \phi_n(n) \text{ defined}\}$ is recursively enumerable (re), p is incrementable. It is easily seen that $p(0^{m+1})/p(0^m) > .5$ iff $\phi_m(m)$ is undefined. But $p(0^{m+1})/p(0^m) = p'(0^m, 0)$. Hence if p' were approachable from below one could eventually verify that $p(1^{m+1})/p(1^m) > .5$ whenever this is the case. But then one could recursively enumerate $\{m | \phi_m(m) \text{ undefined}\}$, which is well-known to be impossible (e.g., see Rogers, 1967). Hence p' is not approachable from below.

Now let p be redefined as follows:
$p(0^m 1) = 2^{-m-1}$ if $\phi_m(m)$ is defined,
$= 0$ otherwise, for all $m \in N$;
$p(0^m) = \sum_{i=0}^{\infty} p(0^{m+i}1)$ for all $m \in N$; and
$p(x) = 0$ for all $x \in X^* - \{0\}^*\{\Lambda, 1\}$.

Again p is easily shown to be an incrementable predictor, with the property that $\forall m \in N$:
$p(0^{m+1})/(p(0^{m+1}) + p(0^m 1) > .5$ iff $\phi_m(m)$ is undefined.
But this ratio is $p''(0^m, 0)$ so that if p'' were approachable from below, $\{m |\, \phi_m(m) \text{ undefined}\}$ would be re. Hence p'' is not approachable from below. □

This result indicates that some incrementable predictors are far from "effective" as methods of prediction, particularly under the second interpretation of predictors. Therefore it is important to study narrower classes of predictors, such as the recursive predictors (Sec. 4).

3. Quasipredictability and Martin-Löf Randomness

The *redundancy* of a sequence $x \in X^*$ relative to a predictor p is defined as $rp(x) = |x| + \log\ p(x)$, with logarithms taken to base 2 and $\log 0 = -\infty$. This can be thought of as the maximum possible information of x, viz. $|x|$, less (the upper bound on) its actual information, viz. $-\log p(x)$. For a sequence $z \in X^\infty$ the redundancy relative to p is defined as $rp(z) = \limsup_n rp(z(n))$. Note that every sequence has zero redundancy relative to the uniform distribution $p(x) = 2^{-|x|}$.

An important related notion is that of optimality. A predictor is optimal for a given

infinite sequence if it "reveals the regularities (redundancies)" of that sequence essentially as well as any other predictor. Formally, an incrementable predictor p is *optimal for z* $z \in X^\infty$, iff $rp'(z(n)) - rp(z(n))$ is bounded above for every incrementable predictor p'.

A sequence $z \in X^\infty$ is *quasipredictable* iff its redundancy is ∞ relative to some incrementable predictor. The prefix "quasi" indicates that prediction with an incrementable predictor is not fully effective.

Quasipredictability will now be related to Martin-Löf randomness.

A *Martin-Löf (M-L) sequential test* is a set $V \subset X^* \times N$ with the following 4 properties, where V_m denotes $\{x | \langle x, m\rangle \in V\}$:

(a) Effectiveness: V is re.

(b) Nestedness: $V_{m+1} \subseteq V_m$ for all $m \in N$.

(c) Numerosity: the number of n-sequences in $V_m \leq 2^{n-m}$ for all $m, n \in N$.

(d) Monotonicity: $x \in V_m \Rightarrow xy \in V_m$ for all $x, y \in X^*$.

For motivation of these properties see Martin-Löf (1966).[5] Intuitively, $\langle x, m\rangle \in V$ means that the test V rejects the randomness hypothesis at significance level 2^{-m} for all infinite sequences beginning with x.

The *critical level* $m_V(x)$ of a sequence $x \in X^*$ relative to a M-L sequential test V is $\max\{m | \langle x, m\rangle \in V\}$, where $\max \emptyset = 0$ (this latter condition in effect extends V_0 to X^*). A sequence $z \in X^\infty$ is *M-L random* iff there is no M-L sequential test V such that $\lim_n m_V(z(n)) = \infty$.

More generally, a *M-L sequential test for p* where p is any recursive distribution, is defined as above except that the numerosity condition becomes

$\sum_{y \in X^n \cap V_m} p(y) \leq 2^{-m}$ for all $m, n \in N$.[6]

Accordingly a sequence $z \in X^\infty$ is *M-L random relative to p* where p is any recursive distribution, iff there is no M-L test V for p such that $\lim_n m_V(z(n)) = \infty$. Note that "M-L random" is the same as "M-L random relative to the uniform distribution".[7]

Intuitively, one would expect that if a distribution is optimal for a sequence $z \in X^\infty$, i.e., if it reveals all the regularities (redundancies) of z, then z should appear to behave randomly relative to the probability assignments of the distribution. This is indeed the case.

Theorem 5 (Levin). For any recursive distribution p and any $z \in X^\infty$, z is M-L

[5] Alternative definitions can be found in Zvonkin & Levin (1970) and Schnorr (1971, 1973).

[6] Martin-Löf originally used strict inequality to facilitate the construction of a universal test (because equality of computable reals is not effectively confirmable). However, extension of V_0 to X^* is then no longer possible, and in any case the construction of a universal test is still possible with non-strict inequality (see Zvonkin & Levin, 1970).

[7] It should be mentioned that the notion of randomness relative to recursive distributions does not allow for randomness relative to Bernoulli distributions of the form $p(\Lambda) = 1, p(x0) = p(x)(1-r), p(x1) = p(x)r$ for any (not necessarily computable) $r \in R$. For a treatment of Bernoulli sequences see Martin-Löf (1966) and Schnorr (1971). For randomness tests relative to arbitrary distributions see Levin (1976).

random relative to p iff p is optimal for z.

Proof. $\Rightarrow$: Suppose that $rp'(z(n)) - rp(z(n))$, i.e., $\log p'(z(n)) - \log p(z(n))$, is unbounded for some incrementable predictor p'. Let

$V = \{\langle x, m\rangle \mid x \in X^* \ \& \ m \in N \ \& \ \exists y \sqsubseteq x : p'(y) > 2^m p(y)\}$.

V will be shown to be a M-L test for p such that z is not M-L random relative to p.

Nestedness and monotonicity are obviously satisfied by V. V is clearly re since p' is incrementable and p is recursive. The numerosity condition can be verified by considering a partitioning of the n-sequences in V_m into groups such that group i consists of all n-sequences extending some y_i with $p'(y_i) > 2^m p(y_i)$. Then

$$1 \geq \sum_i p'(y_i) > 2^m \sum_i p(y_i) = 2^m \sum_{\substack{|x|=n \\ x \sqsupseteq y_i}} p(x) = 2^m \sum_{\substack{|x|=n \\ x \in V_m}} p(x),$$

$$\text{so that} \quad 2^m \sum_{\substack{|x|=n \\ x \in V_m}} p(x) < 2^m.$$

Now from the definition of V, $\langle z(n), m\rangle \in V$ if $\log p'(z(n)) > m + \log p(z(n))$. But $\limsup_n [\log p'(z(n)) - \log p(z(n))] = \infty$, hence $\lim_n m_V(z(n)) = \infty$ and z is not M-L random relative to p.

$\Leftarrow$: Assume without loss of generality that $p(z(n))$ does not vanish for any n (otherwise $\log p'(z(n)) - log\ p(z(n))$ will certainly be unbounded for any nonvanishing p'). Let V be a M-L test for p such that $\lim_n m_V(z(n)) = \infty$. Let p'(x) be defined for all $x \in X^*$ as

$p'(x) = \lim_n q(x, n)$, where

$q(x, n) = \sum_{\substack{y \sqsupseteq x \\ |y|=n}} m_V(y)p(y)$ for $n > |x|$.

p' will be shown to be an incrementable predictor such that $\log p'(z(n)) - \log\ p(z(n))$ is unbounded. First observe that q(x,n) is nondecreasing in n, as $m_V(y0), m_V(y1) \geq m_V(y)$, so that

$$\begin{aligned} \sum_{\substack{y \sqsupseteq x \\ |y|=n+1}} m_V(y)p(y) &= \sum_{\substack{y \sqsupseteq x \\ |y|=n}} [m_V(y_0)p(y0) + m_V(y1)(p(y1)] \\ &\geq \sum_{\substack{y \sqsupseteq x \\ |y|=n}} m_V(y)p(y). \end{aligned}$$

Furthermore $q(x, n) \leq q(\Lambda, n) \leq 1$ for all x, n since

$$\begin{aligned} q(\Lambda,n) &= \sum_{|y|=n} m_V(y)p(y) = \sum_{m\in N} m \sum_{\substack{|y|=n \\ m_V(y)=m}} p(y) \\ &= \sum_{m\in N} m \left[\sum_{y\in X^n\cap V_m} p(y) - \sum_{y\in X^n\cap V_{m+1}} p(y) \right] \\ &= \sum_{m=1}^{\infty} \sum_{y\in X^n\cap V_m} p(y) \leq \sum_{m=1}^{\infty} 2^{-m} = 1. \end{aligned}$$

Thus $\lim_n q(x,n)$ exists and is ≤ 1 for all x. From $q(x,n) = q(x0,n) + q(x1,n)$ for all $n > |x|$, it follows that $p'(x) = p'(x0) + p'(x1) \geq 0$ so that p' is a predictor (in fact, an additive predictor). From the fact that V is re and that for all $x \in X^*$ and $r \in R$

$$p'(x) > r \;\Rightarrow \exists n : q(x,n) > r$$

it is easy to see that p is incrementable. Now since $q(x,n)$ is nondecreasing in n,

$$p'(z(n)) \geq q(z(n),n) = m_V(z(n))p(z(n)).$$

Hence if $m_V(z(n))$ is unbounded, so is $p'(z(n))/p(z(n))$ and hence also $\log p'(z(n)) - \log p(z(n))$. □

Remark. It would have been possible to use

$$q(x,n) = \sum_{\substack{y\sqsupseteq x \\ |y|=n}} f(m_V(y))p(y)$$

in the proof, where f is any unbounded nondecreasing recursive function from N to $\mathbb{R}$ such that

$$\sum_{m\in N} f(m)2^{-m-1} \leq 1;$$

the reason is that $q(\Lambda,n)$ then becomes

$$\begin{aligned} &f(0) + \sum_{m=1}^{\infty} \left[(f(m) - f(m-1)) \sum_{y\in X_n\cap V_m} p(y) \right] \\ &\leq f(0) + \sum_{m=1}^{\infty} (f(m) - f(m-1))2^{-m} = \sum_{m=0}^{\infty} 2^{-m-1}. \end{aligned}$$

A particularly interesting corollary of Th. 5 results from specializing p to the uniform distribution. This provides the link between quasipredictability and M-L randomness.

Corollary 1. A sequence $z \in \mathrm{X}^\infty$ is M-L random iff it is not quasipredictable.

In other words, the M-L random sequences are those for which the uniform predictor is optimal.

The next corollary supplies a predictor which is *universal* in the weak sense that it assigns infinite redundancy to all nonrandom sequences. A predictor which is optimal for all sequences (and hence certainly universal) is given in Th. 8.

Corollary 2 (universal predictor). There is an incrementable predictor p such that for any $z \in X^\infty$ and any incrementable predictor p'

$$rp'(z) = \infty \;\Rightarrow\; rp(z) = \infty.$$

A suitable p is given by

$$p(x) = \lim_n 2^{-n} \sum_{\substack{y \sqsupseteq x \\ |y|=n}} m_U(y),$$

where U is a universal M-L sequential test (Martin-Löf, 1966).

Corollary 3. If p is a universal predictor then $p(x) > 0$ for all $x \in X^*$.

Proof. For any $x \in X^*$, consider the sequence $z = xy$, where $y \in X^\infty$ is some fixed recursive sequence. Clearly there is an incrementable predictor p' such that $p'(z(n)) = 1$ for all $n \in N$. Hence z is nonrandom and hence $rp(z) = \infty$. But if $p(x)$ were 0, $rp(z(n))$ would be $-\infty$ for all $n \geq |x|$. □

Universal predictors are not recursive. This is the analogue of the fact that there is no recursive universal M-L sequential test.

Theorem 6. There is no recursive universal predictor.

Proof. By Cor. 3 of Th. 5 it is only necessary to prove that any nonvanishing recursive predictor p is not universal. Now if p is recursive then arbitrarily tight upper and lower bounds on $p(x)$ can be effectively computed for any $x \in X^*$. This fact allows the construction of a recursive sequence $y \in X^\infty$ whose redundancy is bounded relative to p, showing that p is not universal. Specifically, the $(n+1)$st digit of y is chosen as follows. Increasingly tight upper and lower bounds on $p(y(n)), p(y(n)0),$ and $p(y(n)1)$ are computed. $y(n+1)$ is assigned the value $y(n)0$ if the inequality

$$p(y(n)0) < p(y(n))2^{2^{-n}-1},$$

is first confirmed, or the value $y(n)1$ if

$$p(y(n)1) < p(y(n))2^{2^{-n}-1}$$

is first confirmed. One of these inequalities will be confirmed eventually since the multiplier of $p(y(n))$ on the right-hand side exceeds $1/2$ for all $n \in N$, and $p(y(n)0), p(y(n)1)$ cannot both exceed $p(y(n))/2$. Thus

$$p(y(n+1)) < p(y(n))2^{2^{-n}-1}$$

for all n and hence

$$p(y(n)) < p(\Lambda) \prod_{i=0}^{n-1} 2^{2^{-i}-1} = p(\Lambda)2^{2-2^{-n+1}-n}.$$

Consequently the redundancy of y satisfies

$$n + \log P(y(n)) < \log p(\Lambda) + 2 - 2^{-n+1}$$

and so is bounded above. Thus p is not universal. □

Corollary. There is no incrementable universal distribution.

Proof. Immediate from Th. 1(a) and Th. 6. □

Note that this implies that for any additive universal predictor $p(\Lambda)$ is not computable.

The predictor of Cor. 2, Th. 5 was seen to be universal in that it assigns infinite redundancy to all infinite nonrandom sequences. However, it is nonoptimal, i.e., it does not in general maximize initial-segment redundancy. The fact that any universal M-L sequential test maximizes the initial-segment critical level of any infinite sequence, apart from a constant (Martin-Löf, 1966), suggests that it should be possible to derive an optimal predictor from such a M-L test. The following theorem does not quite succeed in confirming this intuition. The predictor exhibited falls short of maximizing redundancies by a term logarithmic in the redundancy. Its special interest lies in the fact that it is additive. In Th. 8 the existence of a truly optimal (but nonadditive) universal predictor will be established without reliance on the properties of M-L tests.

Theorem 7. There is an additive incrementable predictor p such that for any incrementable predictor p' and any real $c > 1$

$$rp'(x) - rp(x) < c \log rp'(x)$$

for all $x \in X^*$ such that $rp'(x)$ is sufficiently large (i.e., larger than some constant dependent on p' and c). A suitable p is given by

$$p(x) = \lim_n 2^{-n} \sum_{\substack{y \sqsupseteq x \\ |y|=n}} f(m_U(y)),$$

where $\log f(m) = m - \log(m+2) - 2\log\log(m+5)$ for all $m \in N$,
and U is any universal M-L sequential test.

Proof. It can be verified that f is nondecreasing and satisfies

$$\textstyle\sum_{m \in N} f(m) 2^{-m-1} \leq 1.$$

By the proof of Th. 5 and the remark following it, p is an additive incrementable predictor. Given any incrementable predictor p', let

$$V = \{\langle x, m\rangle | x \in X^* \;\&\; m \in N \;\&\; \exists y \sqsubseteq x : p'(y) > 2^{m-|y|}.\}$$

For this M-L sequential test it is known that if $\log p'(x) > m - |x|$ then $m_V(x) \geq m$. Also, since U is universal there is an integer a such that $m_U(x) \geq m_V(x) - a$ for all $x \in X^*$. Clearly $p(x) \geq 2^{-|x|} f(m_U(x))$ or,

$$\log p(x) \geq -|x| + m_U(x) - \log(m_U(x)+2) - 2\log\log(m_U(x)+5).$$

Thus $\log p'(x) > m - |x|$ implies $m_U(x) \geq m - a$ which in turn implies

$$\log p(x) \geq -|x| + m - a - \log(m-a+2) - 2\log\log(m-a+5)$$

provided that $m - a \geq 0$ (in view of the fact that $f(i)$ is nondecreasing in i for $i \in N$);

or, $|x| + \log p'(x) > m \geq a$ implies

$|x| + \log p(x) \geq m - [a + \log(m - a + 2) + 2 \log\log(m - a + 5)]$;

or, choosing m such that $m+1 \geq |x|+\log p'(x) > m$ and assuming that $|x|+\log p'(x) > a$,

$\log p'(x) - log\ p(x)$

$< 1 + a + \log(|x| + \log p'(x) - a + 2) + 2 \log\log(|x| + \log p'(x) - a + 5)$

$< c \log rp'(x)$ for all sufficiently large $rp'(x)$

and the theorem follows. □

Essentially the following lemma says that the class of incrementable predictors is re, and that the function underlying a predictor can be chosen to be subadditive (like the predictor itself). This fact facilitates the construction of an optimal predictor (Th. 8) and the reduction of predictors to Solomonoff's form (Th. 12).

Lemma 1. There is a re class of recursive functions $\{h_i : X^* \times N \to Q,\ i \in N\}$ each of whose members h_i underlies a predictor p_i and satisfies

$h_i(x, n) \geq h_i(x0, n) + h_i(x1, n)$

for all $x \in X^*$, $n \in N$. Furthermore $\{p_i | i \in N\}$ is the class of all incrementable predictors.

Proof. It will first be shown that the class of incrementable *functions* is re. From any partial recursive function $\phi : X^* \times N \to Q$ a recursive function $g : X^* \times N \to Q$ can be obtained uniformly effectively, such that $\lambda n\, g(x, n)$ is nondecreasing and

$\lim_n g(x, n) = \sup \phi(x, n) | n \in N$ for all $x \in X^*$, where $\sup \emptyset = 0$.

Given a procedure for ϕ, $g(x, n)$ can be computed by simulating the computation of $\phi(x, 0), ..., \phi(x, n)$ for n steps each and returning the largest of the outputs obtained (0 if none are obtained). If ϕ is already a recursive nondecreasing function, then $\lim_n g(x, n)$ will clearly be the same as $\lim_n \phi(x, n)$. Since the class of partial recursive functions is re, it follows that the class of incrementable functions is re.

The members of this class can now be further modified to yield only (and all) incrementable predictors. For any incrementable g, let

(1) $h(x, 0) = 0$,
(2) $h(\Lambda, n) = g(A, n)$
(3) $h(x0, n) = \min\{g(x0, n), h(x, n) - h(x1, n - 1)\}$ if $n > 0$, and
(4) $h(x1, n) = \min\{g(x1, n), h(x, n) - h(x0, n)\}$, for all $x \in X^*$, $n \in N$.

Assume for induction on $|x|$ that $h(x, n) \geq h(x, n - 1)$ for all $n > 0$ and all x of length $\leq k$. Then the inductive step requires proving $h(x0, n) \geq h(x0, n - 1)$ and $h(x1, n) \geq h(x1, n - 1)$, i.e.,

(5) $g(x0, n) \geq h(x0, n - 1)$,
(6) $h(x, n) - h(x1, n - 1) \geq h(x0, n - 1)$,
(7) $g(x1, n) \geq h(x1, n - 1)$, and
(8) $h(x, n) - h(x0, n) \geq h(x1, n - 1)$.

Since g is nondecreasing and since $g(x, n) \geq h(x, n)$ for all x, n (by inspection of (1)-(4)),

therefore (5) and (7) hold. From (4)
(9) $h(x, n-1) \geq h(x0, n-1) + h(x1, n-1)$
for all x and $n > 0$, and together with the induction assumption this implies (6). (8) is immediate from (3). The basis of the induction is $h(\Lambda, n) = g(\Lambda, n) \geq g(\Lambda, n-1) = h(\Lambda, n-1)$. Thus h is nondecreasing and hence (9) holds in the limit as $n \to \infty$, i.e., h underlies a predictor.

Whenever g underlies a predictor p, h underlies the same predictor. This is proved by assuming for induction that for all x of length $\leq k$ and all $r \in R$, if $\exists n \in N : g(x, n) > r$ then $\exists n \in N : h(x, n) > r$. Suppose that $\exists r \in R : \exists n \in N : g(x0, n) > r$. Then since g underlies predictor p, $\exists n' \in N : g(x, n') > r + p(x1)$, and hence $\exists n'' \in N : h(x, n'') > r + p(x1)$, by the induction assumption. Hence by (3) either $h(x0, n'') = g(x0, n'') > r$ (assuming w.l.g. that $n'' \geq n$) or $h(x0, n'') = h(x, n'') - h(x1, n'-1) > r + p(x1) - g(x1, n''-1) \geq r$. The argument for $h(x1, n)$ is similar, and equation (2) starts the induction. Since the transformation from g to h is uniformly effective, and h is subadditive (see equation (9)), the proof of the lemma is complete. □

The construction of the following optimal predictor is modelled on Martin-Löf's (1966) costruction of a universal test. An alternative construction follows as an easy corollary of the reduction of incrementable predictors to Solomonoff predictors (see corollary of Th. 12); this was the approach used by Levin (in Zvonkin & Levin, 1970). However, the required reduction is itself nontrivial, so that an approach not dependent upon it is of some interest.

Theorem 8 (Levin, optimal universal predictor). There is an incrementable predictor p such that for any incrementable p' there is a constant c satisfying

$$rp'(x) - rp(x) \leq c \text{ for all } x \in X^*.$$

Proof. With p_i, $i \in N$, defined as in Lemma 1, the optimal universal predictor is given by

$$p(x) = \sum_{i=0}^{\infty} 2^{-i-1} p_i(x) \text{ for all } x \in X^*.$$

Then $p(x) = \lim_n h(x, n)$ for all $x \in X^*$, where

$$h(x, n) = \sum_{i=0}^{-i-1} h_i(x, i).$$

Clearly h is recursive and nondecreasing and

$$p(x0) + p(x1) \leq p(x) \leq \sum_{i=0}^{\infty} 2^{-i-1} = 1 \text{ for all } x \in X^*.$$

and hence p is an incrementable predictor. Furthermore if p' is any incrementable predictor then there is an $i \in N$ such that $p' = p_i$, so that $p(x) \geq 2^{-i-1} p'(x)$ for all $x \in X^*$. The theorem follows with $c = i + 1$. □

Corollary (Levin). For any optimal predictor p and recursive distribution p', the infinite sequences z with absolutely bounded $rp(z(n)) - rp'(z(n))$ are the sequences which are M-L random relative to p'. In particular, the infinite sequences whose redundancy relative to p is absolutely bounded are the M-L random sequences.

Proof. $rp(z(n)) - rp'(z(n))$ is bounded above for any sequence $z \in X^\infty$ which is M-L random relative to p', by Th. 5. Clearly it is also bounded below for if it were not then $rp'(z(n)) - rp(z(n))$ would not be bounded above, in contradiction with Th. 8. □

This section will be concluded with a proof of the fact that any optimal universal predictor assigns nonvanishing limit probabilities precisely to the recursive sequences. This interesting result was previously obtained by de Leeuw et al (1956) in a paper on probabilistic machines, and was given its present interpretation by Levin (in Zvonkin & Levin, 1970). Levin's own proof depended on properties of Loveland's "uniform complexity" (Loveland, 1970). For a closely related result see also Chaitin (1976).

Theorem 9 (de Leeuw et al., Levin). If p is an optimal universal predictor then $z \in X^\infty$ is recursive iff $\lim_n p(z(n)) > 0$.

Proof. $\Rightarrow$: For any recursive $z \in X^\infty$ define $p'(z(n)) = 1$ for all $n \in N$ and $p'(x) = 0$ for $x \not\sqsubset z$. By Th. 5 there exists a constant c such that $rp'(z(n)) - rp(z(n)) = -\log p(z(n)) \leq c$ for all $n \in N$, or $p(z(n)) \geq 2^{-c}$ for all $n \in N$.

$\Leftarrow$: Let $z \in X^\infty$ be any sequence such that for some $c \in N$ $p(z(n)) > 2^{-c}$ for all $n \in N$. Thus $\lim_n p(z(n)) \geq 2^{-c}$. There are fewer than 2^c such infinite sequences, i.e., with limiting probability $\geq 2^{-c}$. Let $z(k)$ be the shortest prefix of z which is not a prefix of any of the other sequences with limiting probability $\geq 2^{-c}$; let l be the smallest integer such that all finite sequences of length $> l$ with probability $> 2^{-c}$ are prefixes of infinite sequences with limiting probability $\geq 2^{-c}$ (it is not hard to see that such an l exists – note that if there were infinitely many finite sequences whose probabilities exceed 2^{-c} but which are not prefixes of a finite number of infinite sequences, then $p(\Lambda) = \infty$ would hold); finally, let $m = \max\{k, l\}$. Then $z(n)$ can be computed for any $n > m$ by enumerating pairs $\langle x, q\rangle \in S$, where S underlies p, until a pair is obtained such that $|x| = n$, $z(m) \sqsubset x$, and $q > 2^{-c}$; this x is $z(n)$. □

Note that another way of stating Th. 9 is that if p is any optimal universal predictor then $z \in X^\infty$ is recursive iff there is a $c \in N$ such that $rp(z(n)) \geq n - c$ for all $n \in N$, i.e., the recursive sequences are those whose redundancy as a function of initial-segment length n is approximately n.

4. Predictability and Schnorr Randomness

Schnorr (1971) drew attention to the fact that the criteria of effectiveness employed by Martin-Löf in defining sequential tests are too weak by the standards of constructive mathematics. In particular, the measure $\mu V_m X^\infty$ is not in general a recursive function of m for a M-L sequential test V. Putting the criticism another way, Schnorr pointed out that M-L tests classify as nonrandom certain sequences whose nonrandomness cannot be effectively observed, in any reasonable sense of effective observation.

He therefore proposed several alternative ways of strengthening the criteria of effec-

tiveness for distinguishing between random and nonrandom sequences, and showed that these lead to equivalent characterizations of randomness (Schnorr, 1971, 1973). One proposal is to require $\mu V_m X^\infty$ to be a recursive function of m. The random sequences are taken to be those with bounded critical level relative to all such tests. An equivalent proposal is contained in the definition of Schnorr randomness which follows.

A *growth function* is a recursive nondecreasing unbounded function $g : N \to N$. A sequence $z \in X^\infty$ is *Schnorr (S-) random* iff there does not exist a M-L sequential test V, a recursive lower bound f of m_V, and a growth function g such that $\limsup_n f(z(n))/g(n) > 0$.

Evidently this definition expresses a specific thesis about what it means for the nonrandomness in a sequence to be effectively observable, viz., one must be able to confirm effectively that the upward excursions of the critical level are bounded below by some growth function. (Note incidentally that one could equally well use 1 or any other positive constant in place of 0 on the right-hand side of the inequality).[8]

It has already been noted that there is no universal recursive predictor, and hence certainly no optimal universal recursive predictor. Nevertheless the concept of optimality, i.e., maximization of initial-segment redundancy, is of as much interest in connection with recursive predictors as in connection with incrementable predictors.

In Sec. 3 an optimal predictor was required to maximize initial-segment redundancy apart from a constant. Adherence to Schnorr's criteria of effectiveness calls for a slightly weaker notion of optimality. A recursive distribution p is said to be *weakly optimal for z* where $z \in X^\infty$, iff for every recursive distribution p' and every growth function g, $\limsup_n (rp'(z(n)) - rp(z(n)))/g(n) \leq 0$. Thus no recursive distribution reveals the redundancy of z noticably better than a predictor which is weakly optimal for z.

For example, corresponding to any computable $r \in R$, the predictor $p(x) = r^{n(x)}(1-r)^{|x|-n(x)}$, where $n(x) =$ the number of ones in x, is weakly optimal for all Bernoulli sequences with success probability r, when the following notion of S-randomness relative to a recursive distribution is substituted in Martin-Löf's definition of Bernoulli sequences.

A sequence $z \in X^\infty$ to be *S-random relative to p* where p is a recursive distribution, iff there is no M-L sequential test V for p, recursive lower bound f of m_V and growth function g such that $\limsup_n f(z(n))/g(n) > 0$.

Once again a direct connection between optimality and randomness can be established, much as in Th. 5.

First it should be noted that the numerosity condition of a M-L test V for p can be

[8]Schnorr's contention that the S-random sequences best correspond to the intuitively random sequences is not universally accepted. Indeed Müller (1972) proposes *weaker* criteria of effectiveness than Martin-Löf, citing the existence of M-L random sequences which are limiting recursively computable as a defect of Martin-Löf's conception of randomness.

restated as

$\sum_{y\in Y\cap V_m} \leq 2^{-m}$

for every finite pf $Y \subset X^*$. Also the following fact will be used.

Lemma 2. Let V be a M-L sequential test for a distribution p, and let g be a growth function. Then for any $x \in X^*$, any integer $k \geq |x|$, and any finite pf $Y \subset xX^{k-|x|}X^*$,

$\sum_{y\in Y\cap V_{g(|y|)}} g(|y|)p(y) \ \leq \ (g(k)+1)2^{-g(k)}$.

Proof. By the numerosity condition the total probability of sequences in $Y \cap V_{g(k)}$ is at most $2^{-g(k)}$, that of sequences in $Y \cap V_{g(k)+1}$ at most $2^{-g(k)-1}$, etc. Hence the above sum is at most

$g(k)2^{-g(k)-1} + (g(k)+1)2^{-g(k)-2} + ... = (g(k)+1)2^{-g(k)}$. □

Theorem 10. For any recursive distribution p and any $z \in X^\infty$, z is S-random relative to p iff p is weakly optimal for z.

Proof. The proof parallels that of Th. 5.
⇒: Suppose that p is not weakly optimal for z, i.e., there is a recursive distribution p' and a growth function g such that $\limsup_n [rp'(z(n)) - rp(z(n))]/g(n) > 0$. The test V is defined as in Th. 5, and its required properties established as before. Thus $\exists c > 0$: for infinitely many n: $\log p'(z(n)) - \log p(z(n)) > cg(n)$, and hence for infinitely many n: $m_V(z(n))/g(n) > 0$. Since m_V is recursive, this implies that z is not S-random relative to p.
⇐: Suppose that V is a M-L test for p such that for some recursive lower bound f of m_V and some growth function g, $\limsup_n f(z(n))/g(n) > 1$.

From this point on the situation is more complicated than in Th. 5, because $\lim_n q(x,n)$ need not be recursive, even with f replacing m_V in the definition of q. The limit operation must somehow be cut short, without violating (sub-) additivity.

Let $p'(x) = p_1(x) + p_2(x) + p_3(x)$ for all $x \in X^*$, where p_1, p_2, and p_3 are defined recursively as follows:

$p_1(\Lambda) = 0, \ \ p_2(\Lambda) = (g(1)+1)2^{-g(1)}, \ \ p_3(\Lambda) = 1 - p_2(\Lambda),$

and for all $u \in X$ and $x \in X^*$

$p_1(xu) = \max_{\text{pf } Y\subseteq xu(X^*-X^{k-|x|-1}X^*)} \sum_{\substack{y\in Y \\ f(y)\geq g(|y|)}} g(|y|)p(y),$

$p_2(xu) = (g(k)+1)2^{-g(k)}$, and

$p_3(xu) = p_3(x)/2 - (g(k)+1)2^{-g(k)} + [p_1(x) + p_2(x) - p_1(x0) - p_1(x1)]/2,$

where k = least integer $\geq |x|+2$ such that $(g(k)+1)2^{-g(k)} < p_3(x)/2$. p' will be shown to be a recursive distribution, assuming without loss of generality that $g(n) \geq 2$ for all $n \in N$.

The properties $p'(\Lambda) = 1$, additivity, and recursiveness are easily verified assuming that a suitable k exists for each $x \in X^*$. To prove the correctness of the latter assumption

by induction, assume that for all x of length l or less

(a) $p_1(x),\ p_2(x)$, and $p_3(x)$ are well-defined; and

(b) $p_3(x) > 0$.

Consider any particular x of length l. Denote the corresponding value of k by k' (if $x = \Lambda$, use $k' = 1$). Since g is a growth function, $(g(k)+1)2^{-g(k)} \to 0$ as $k \to \infty$. Also $p_3(x) > 0$, hence the new value of k required in the definition of $p'(xu), u \in X$, exists. Denote it by k''. Thus $p_1(xu)$, $p_2(xu)$, and $p_3(xu)$ are well-defined. From the definition of p_3, observe that $p_3(xu) > 0$ if $p_1(x) + p_2(x) \geq p_1(x0) + p_1(x1)$. But

$$
\begin{aligned}
& p_1(x0) + p_1(x1) \\
&= \max_{\text{pf } Y \subseteq xX(X^* - X^{k''-|x-1}X^*)} \sum_{\substack{y \in Y \\ f(y) \geq g(|y|)}} g(|y|)p(y) \\
&\leq \max_{\text{finite pf} \subseteq xX^*} \sum_{\substack{y \in Y \\ f(y) \geq g(|y|)}} g(|y|)p(y) \\
&\leq \max_{\text{pf } Y \subseteq x(X^* - X^{k'-|x|}X^*)} \sum_{\substack{y \in Y \\ f(y) \geq g(|y|)}} g(|y|)p(y) \\
&\quad + \max_{\text{finite pf } \subseteq xX^{k'-|x|}X^*} \sum_{\substack{y \in Y \\ f(y) \geq g(|y|)}} g(|y|)p(y) \\
&\leq p_1(x) + (g(k')+1)2^{-g(k')} \quad \text{by Lemma 2} \\
&= p_1(x) + p_2(x),
\end{aligned}
$$

so that the inductive step is complete. The induction starts at $l = 0$. In this case (a) certainly holds and $p_3(\Lambda) \geq 1/4$, since $g(1) \geq 2$ by assumption.

Now the definition of p_1 ensures that $p'(x) \geq g(|x|)p(x)$ whenever $f(x) \geq g(|x|)$. But $f(z(n)) > g(n)$ for infinitely many n, hence $p'(z(n)) \geq g(n)p(z(n))$ for infinitely many n, i.e., $\limsup_n [rp'(z(n)) - rp(z(n))]/log\ g(n) \geq 1$, so that p is not weakly optimal for z. □

In conformity with Schnorr's notion of effectively observable growth, a sequence $z \in X^\infty$ is defined to be *predictable* iff there exists a recursive predictor p and a growth function g such that $\limsup_n rp(z(n))/g(n) > 0$; i.e., the redundancy of z grows "noticably". Note that p may as well be taken to be a recursive distribution, by Th. 1(b). Also, since $(p(x) + 2^{-|x|})/2$ defines a positive recursive distribution, $p(x)$ may be taken to be positive for all $x \in X^*$. This leads to the following

Corollary. A sequence $z \in X^\infty$ is S-random iff it is not predictable.

Proof. Let p in Th. 10 be the uniform distribution. □

An alternative proof of the corollary is easily obtained from one of Schnorr's (1971) re-

sults about martingales. Such a proof is given below, partly because of the significance of the corollary and partly because of the inherent interest of Schnorr's result. Martingales (first used by Ville, 1939) describe the capital of a gambler who bets on the occurrence of 0 and 1 as next digit in a sequence and subsequently wins the amount wagered on the digit which actually follows and loses the amount wagered on its complement. Formally, a *martingale* is a total function $f : X^* \to \mathbb{R}^+$ such that $f(x) = (f(x0) + f(x1))/2$ for all $x \in X^*$. Schnorr showed that a sequence $z \in X^\infty$ is S-random iff there does not exist a recursive martingale f and a growth function g such that $\limsup_n f(z(n))/g(n) > 0$. Again this embodies the previous notion of effectively observable growth, in this case of the gambler's capital.

Alternative proof of corollary. $\Rightarrow$: Suppose that there is a positive recursive distribution p and a growth function g such that $\limsup_n rp(z(n))/g(n) > 0$. Then f defined by $f(x) = 2^{|x|}p(x)$ for all $x \in X^*$ is a recursive martingale and $\limsup_n f(z(n))/2^{g(n)} = \limsup_n 2^{rp(z(n))}/2^{g(n)} > 0$. Hence z is not S-random.

$\Leftarrow$: Suppose that there is a recursive martingale f and a growth function g such that $\limsup_n f(z(n))/g(n) > 1$. Then p defined by $p(x) = 2^{-|x|}f(x)/f(\Lambda)$ for all $x \in X^*$ is a recursive distribution and $\limsup_n rp(z(n))/\log g(n) = \limsup_n(\log\ f(x) - \log\ f(\Lambda))/\log g(n) > 0$. Hence z is predictable. □

It is tempting, in view of Ths. 5 & 10, to identify optimal methods of prediction with "rational" methods of prediction, i.e., to stipulate as a general requirement for any "rational" method of probabilistic prediction that any sequence of observations to which the method is applied should appear to behave randomly relative to the probability assignments of the method. This seems to be Levin's view (Zvonkin & Levin, 1970, and Levin, 1973, 1976), since he identifies the probabilities of the optimal semicomputable measure with intuitive prior probabilities. However, this requirement is surely too strong, since no universal (and hence no generally optimal or weakly optimal) computable predictor exists; i.e., no predictive method exists which is both effective and "rational" in so strong a sense.

Therefore it seems appropriate to admit as "rational" those methods of prediction relative to which any sequence passes some, but not necessarily all, tests for randomness. In particular, it may be sufficient to require all sequences to satisfy a law of large numbers relative to the predictive method. Then it becomes possible to construct effective methods of prediction which are "rational".

Th. 10 is thus best viewed as providing a method for comparing "rational" predictors as to their predictive "power": a predictor is powerful to the extent that sequences pass a variety of randomness tests (in addition to those criterial to "rationality") relative to the pedictor. The second part of the proof of Th. 10 also indicates how a predictive method might be improved if a sequence is known which does not pass some randomness test relative to the method.

What kind of law of large numbers can be formulated relative to an arbitrary recursive

predictor? The law should express that the frequencies of the predicted events conform with the probabilities assigned to them. Specifically, the frequency of occurrence of events which are assigned conditional probabilities within some particular interval should lie in that same interval. The next theorem shows that this law is satisfied by relatively S-random sequences.

First a *rational-computable* distribution is defined as one whose values are rational numbers which can be found uniformly effectively for all arguments. Note that if p is a recursive distribution then there is a rational-computable approximation p' to p which attributes the same redundancy as p to all $x \in X^*$, apart from a constant. $p'(x)$ need only lie within distance $2^{-|x|}p(x)$ of $p(x)$ for all $x \in X^*$; this guarantees that $\ln p(x) - \ln p'(x) < 2$ for all $x \in X^*$. Hence if p is weakly optimal for a sequence, so is p'. Thus the restriction to rational-computable distributions in the following theorem detracts little from its interest. It is assumed that an assertion in square brackets has numerical value 1 if the assertion is true and value 0 otherwise.

Theorem 11. Let p be a rational-computable distribution which is weakly optimal for some $z \in X^\infty$. Let r, s be rational numbers such that $.5 < r \leq s < 1$ and

$$\liminf_n \sum_{i=0}^n \sum_{w \in X} [r \leq p(z(i))w)/p(z(i)) \leq s]/g(n) > 0$$

for some growth function g, i.e., the frequency of next-digit predictions with probabilities in $[r, s]$ is nonnegligible. Then the proportion of *confirmed* predictions with probabilities in this range satisfies

$$\lim\{ \begin{matrix} \inf_n \\ \sup_n \end{matrix} \} \frac{\sum_{i=1}^n [r \leq p(z(i+1))/p(z(i)) \leq s]}{\sum_{i=0}^n \sum_{y \in X} [r \leq p(z(i)y)/p(z(i)) \leq s]} \{ \begin{matrix} \geq r \\ \leq s \end{matrix} \}.$$

Proof. In view of Th. 10 one might attempt a proof by formulating a suitable M-L test for p. Instead the following argument proceeds directly from the assumption that p is weakly optimal for z.

Suppose contrary to the theorem that the lim inf of the above ratio $< r - a$ for some real $a > 0$. As before let x^- denote the sequence obtained by changing the last digit of $x \in XX^*$ to its complement. Let b be any rational number such that $0 < b < a$, and define p' by

$p'(\Lambda) = 1$ and for all $x \in X^*$, $u \in X$
$p'(xu)/p'(x) = p(xu)/p(x) - b$ if $p(xu)/p(x) \in [r, s]$,
$p'(xu)/p'(x) = p(xu)/p(x) + b$ if $p(xu^-)/p(x) \in [r, s]$,
$p'(xu)/p'(x) = p(xu)/p(x)$ otherwise,

where any occurrences of $0/0$ are replaced by 0. Clearly p' is a recursive distribution (note that $p(xu)/p(x) \in [r, s]$ is decidable if p is rational-computable).

By supposition, for infinitely many $n \in N : \exists k, l \in N :$

$k = \sum_{i=0}^{n-1}[r \leq p(z(i+1))/p(z(i)) \leq s]$,
$l = \sum_{i=0}^{n-1}[r \leq p(z(i+1)!13)/p(z(i)) \leq s] > 0$, and
$k/(k+l) < r - a$.

For any such n denote the k values of $p(z(i+1))/p(z(i))$ in $[r,s]$ by $p_1, ..., p_k$ and the l values of $p(z(i+1))/p(z(i))$ in $[1-s, 1-r]$ by $q_1, ..., q_l$ (observe that $p(z(i+1))/p(z(i)) \in [1-s, 1-r]$ iff $p(z(i+1)!13)/p(z(i)) \in [r,s]$). The corresponding values of $p'(z(i+1))/p'(z(i))$ are $p_1 - b, ..., p_k - b$ and $q_1 + b, ..., q_l + b$ respectively. Since the conditional probabilities of the digits of z determined by p and p' for $i \leq n$ are otherwise identical and nonzero,

$$\begin{aligned}
&\ln p'(z(n) - \ln p(z(n)) \\
&> \sum_{i=1}^{k} \frac{b}{p_i - b} + \sum_{i=1}^{l} \frac{b}{q_i + b} \\
&\geq -\frac{kb}{r-b} + \frac{lb}{1-r+b} = b\frac{(k+l)(r-b) - k}{(r-b)(1-r+b)} \\
&> b\frac{(k+1)(a-b)}{(r-b)(1-r+b)} \text{ since } (k+1)r - k > (k+1)a.
\end{aligned}$$

But $k + l$, as a function of n, is bounded below by $g(n)$ times some constant for all sufficiently large n, hence

$\limsup_n [rp'(z(n)) - rp(z(n))]/g(n) > 0$,

in contradiction with the assumption that p is weakly optimal for z. The proof of the second part of the theorem is entirely analogous to that of the first. □

It may be possible to strengthen the theorem in various ways, for example by specifying the rate of convergence towards the two limits (more precisely, by specifying the critical levels associated with deviations of the proportion of confirmed predictions from lower bound r or upper bound s). It may also be possible to formulate the theorem so as to apply to arbitrary subsequences extracted from infinite sequences by application of recursive selection rules, or to apply to prediction of more general types of events, such as arbitrary cylinder sets of sequences. However, the theorem in its present form sufficiently illustrates the conformity between the predictions of an optimal predictor and the occurrences of the predicted events.

5. Solomonoff Predictors

In this section incrementable predictors will be related to one of the classes of sequence predictors proposed by Solomonoff (1964) in his pioneering work on inductive inference for infinite sequences (see also Willis, 1970, Zvonkin & Levin, 1970, Chaitin, 1975, Cover, 1974, Leung-Yan-Cheong & Cover, 1975, and Solomonoff, 1976 for closely related studies).

Actually Solomonoff considered four methods of predicting sequences probabilistically. In the first three methods the probability $p(y)$ of a sequence $y \in X^*$ is obtained by summing terms of the type $2^{-|x|}$, where x is an encoding or program from which y can be generated on a fixed machine. This formalizes the intuitive idea that the highest prior probability should go to sequences with short and/or numerous encodings. The three methods differ in the types of machines considered and in other relatively minor respects. Solomonoff conjectures that they are essentially equivalent.

Here a machine-independent formulation of Solomonoff's second method will be used. The formulation is based on Schnorr's notion of a *process* (or *monotone function* defined as a partial recursive function $f : X^* \to X^*$ such that $f(x) \sqsubseteq f(xy)$ for all x, xy in the domain of f (Schnorr, 1971, 1973; also Zvonkin & Levin, 1970). Thus processes map extensions of inputs into extensions of outputs. x is said to be an *encoding* of y relative to process f, abbreviated as $f(x) \succcurlyeq y$, whenever f maps x, but no proper prefix of x, into an extension of y. In symbols, $f(x) \succcurlyeq y$ iff $x \in f^{-1}(yX^*) - f^{-1}(yX^*)XX^*$. A *Solomonoff predictor* is now defined as a function $p : X^* \to R$ such that for some process f

$$p(y) = \sum_{f(x) \succcurlyeq y} 2^{-|x|} \text{ for all } y \in X^*,$$

where a sum over no terms is 0, as before. The symbol p_f denotes the Solomonoff predictor determined by any process f. The notation σS will be used as an abbreviation for $\sum_{x \in S} 2^{-|x|}$.

Thus $p_f(y) = \sigma\{x | f(x) \succcurlyeq y\}$. Note that for any pf set $S \subset X^*$, $\mu S X^\infty = \sigma S$, where μ is the uniform measure on X^∞.

The equivalence of Solomonoff predictors and incrementable predictors will now be established. Thus in considering methods of probabilistic sequence prediction, one can in principle restrict attention to methods which attribute high probability to sequences with short and/or numerous encodings, just as Solomonoff suggested.

In the following it will sometimes be helpful to think of any set S $\subseteq$X* as a set of nodes of a binary tree rooted at Λ, with a pf set containing leaf nodes only. At other times it will be useful to think of a sequence $x = x_1 x_2 ... x_n$ ($x_i \in X$) as the real *interval*

$$[\textstyle\sum_{i=1}^{n} x_i 2^{-i} \ , \ \sum_{i=1}^{n} x_i 2^{-i} + 2^{-n}),$$

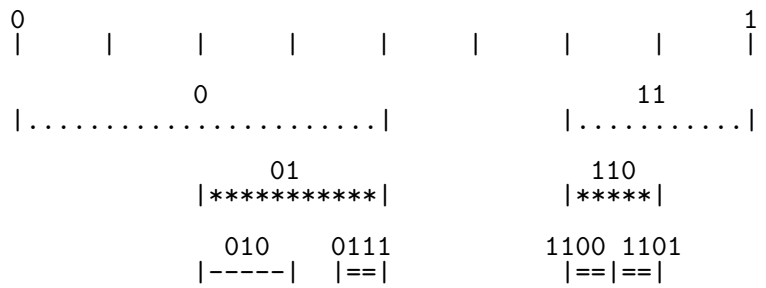
i.e., the least interval containing all numbers in B whose (finite) radix-2 representation begins with x. Then pf sets of sequences correspond to sets of disjoint intervals, and extensions of sequences correspond to subintervals.

Theorem 12 (Levin). A function $p : X^* \to R$ is an incrementable predictor iff it is a Solomonoff predictor.

Proof. $\Leftarrow$: For any $y \in X^*$, sequences mapped by a process f into extensions of $y0$ or $y1$ are certainly mapped into extensions of y. Hence $p_f(y0) + p_f(y1) \leq p_f(y) \leq 1$. p_f is incrementable because f is re and for any finite process $f' \subseteq f$, $p_{f'}(y)$ is rational and

$\leq p_f(y)$ for all $y \in X^*$.

$\Rightarrow$: As previously indicated[2], the ranges of the functions underlying incrementable predictors might equally well have been confined to B (instead of Q). Also, by Lemma 1 the function h underlying an incrementable predictor can be chosen to be subadditive, i.e., $h(y,n) \geq h(y0,n)+h(y1,n)$ for all $y \in X^*$, $n \in N$. The elements of a process f such that $p_f = p$ can now be generated as follows. At stage n of the generation procedure a finite set of elements is added to the process for each y of length 0, 1, ..., n (in that order), so as to increase the Solomonoff probability of y from $h(y,n-1)$ to $h(y,n)$. Each new element $\langle x,y\rangle$ added to the process is chosen so that no extension of x is as yet in the domain of the process, and x properly extends some x' where $\langle x', y(|y|-1)\rangle$ was added to the process earlier (the latter condition is omitted for $y = \Lambda$). The required sets of additions to increase the Solomonoff probabilities of $y0$ and $y1$ from $h(y0,n-1)$ and $h(y1,n-1)$ to $h(y0,n)$ and $h(y1,n)$ respectively always exist because of the subadditivity of h (a detailed argument using induction on n and $|y|$ is easily supplied). The sets are always finite because the required probability increments are in B. The construction is illustrated in Fig. 1, using the interval representation of certain sequences in the domain of the process being constructed. □



Graphical symbolism:

|.....| mapped into proper prefixes of y
|*****| mapped into y
|-----| mapped into $y0$
|=====| mapped into $y1$

Fig. 1. Construction of process f corresponding to given incrementable predictor. At the point shown $p_f(y) = 3/8$, $p_f(y0) = 1/8$, $p_f(y1) = 3/16$.

A process f is said to be <u>*universal*</u> iff for every process f' : $\exists x \in X^*$: $\lambda y\, f(xy) = f'$. Levin (in Zvonkin & Levin, 1970) and also Schnorr (1973) proved the existence of a universal process.

<u>*Corollary.*</u> If f is a universal process then p_f is an optimal universal predictor.

<u>*Proof.*</u> Let p be an optimal universal predictor and f' a process such that $p_{f'} = p$. Then

$\exists x \in X^* : \lambda y\, f(xy) = f'$. Hence

$$p_{f'}(z) = \sigma\{y|\, f(xy) \succcurlyeq z\} = 2^{|x|}\sigma\{xy|f(xy) \succcurlyeq z\} \leq 2^{|x|}p_f(z)$$

for all $z \in X^*$, and the corollary follows.□

An optimal Solomonoff predictor, as defined in the corollary, is not quite the same as Cover's universal prediction scheme (Cover, 1974). Cover's scheme could be obtained by retaining only certain terms of p_f, namely those contributed by encodings x such that x is the *shortest* argument for which f assumes value $f(x)$. Optimal Solomonoff predictors do appear to be essentially the same as Solomonoff's own measures P'_M (Solomonoff, 1976) apart from the normalization term employed by Solomonoff. To prove this one would have to relate processes to Solomonoff's computational model, which permits finite and infinite inputs and outputs, as well as finite outputs generated by nonterminating computations.

Willis (1970) called a distribution p *binary-computable* iff it is a recursive mapping into B. As in the case of rational-computable distributions it is easily seen that any recursive distribution p can be approximated by a binary-computable distribution which attributes the same redundancy as p to all $x \in X^*$, apart from an arbitrarily small constant.

A class of processes will now be characterized which corresponds to the class of binary-computable distributions. This leads to a machine-independent formulation of one of Willis' main results (a closely related result is proved in Zvonkin & Levin, 1970).

A process f is called *endless* iff the set $f(z(n)|n \in N)$ is infinite for every $z \in X^\infty$.

Theorem 13. p is a binary-computable distribution iff there is an endless process f such that $p = p_f$.

Proof. ⇒: A procedure for generating a process can be used, similar to that in the proof of Th. 12. Corresponding to each y of length 0, 1, 2, ... (considered in that order), a set of elements is added to the process such that the Solomonoff probability of y becomes $p(y)$. Because of the distribution property, i.e., $p(y) = p(y0) + p(y1)$, the construction of f and the proof that $p_f = p$ present no difficulty. Now clearly the minimal length of sequences in the successive subdomains $f^{-1}(\Lambda), f^{-1}(X), \ldots, f^{-1}(X^n), \ldots$ is strictly increasing as a function of n, and $X^\infty = f^{-1}(\Lambda)X^\infty = f^{-1}(X)X^\infty = \ldots$. Hence f is endless.

⇐: For a given endless process f, $p(y)$ can be computed for any $y \in X^*$ by enumerating elements of f until a finite subprocess $f' \subseteq f$ is obtained such that

$$\textstyle\sum_{|x|=|y|} p_{f'}(x) = \sum_{|x|=|y|} p_f(x) = 1.$$

At that point $p_f(y)$ will be available as a finite sum of nonpositive powers of 2. To see that the required f' always exists, note first that $f^{-1}(y)$ is finite for all y. For if it were not for some y, then by König's infinity lemma (see e.g., Knuth, 1968) there would be an infinite sequence z all of whose prefixes have extensions in $f^{-1}(y)$. Hence by the definition of an endless process there would be an infinite set of prefixes of z mapped by

f onto an infinite set of output sequences. But prefixes of z can only be mapped into the finite set of prefixes of y (by the process property); thus $f^{-1}(\{y\})$ cannot be infinite. It follows that for any sufficiently large n, no element of $X^nX^* \cap \text{dom}\, f$ will be mapped into sequences of length $\leq |y|$. But since dom f contains arbitrarily long prefixes of every infinite sequence, hence $(X^nX^* \cap \text{dom}\, f)X^\infty = X^\infty$ for all $n \in N$. Thus the Solomonoff probability of the sequences generated by f on this subdomain is 1, and for n sufficiently large these sequences are all of length $\geq |y|$. Furthermore, the set $X^nX^* \cap$ dom f, made pf by removal of sequences which properly extend other sequences in the set, is finite; otherwise there would be a $z \in X^\infty$ none of whose prefixes are in the set, again by König's lemma. Thus the subprocess $f' \subseteq f$ with pf subdomain $X^nX^* \cap$ dom f and with n sufficiently large possesses the required properties. □

Corollary 1. p is a binary-computable distribution iff there exists an endless recursive process f such that $p = p_f$.

Proof. It need only be shown that for every endless process f there exists an endless recursive process f' such that $p_{f'} = p_f$. Such an f' is easily obtained by a slight modification of the process construction mentioned in the first part of the proof of Th. 13 (the construction is applicable because p_f is a binary-computable distribution). In addition to the process elements generated in that construction, $\langle x, y\rangle$ is added to the process whenever $\langle x0, yu\rangle$ and $\langle x1, yv\rangle$ have previously been added, where $\{u, v\} \subseteq X$. These additions do not affect the Solomonoff probabilities, and are easily seen to extend the domain of the process to X^*. □

Willis also showed that if p is a binary computable distribution then there is a machine (of the type he considered) whose *shortest* encoding for any output sequence y determines the highest-order digit of $p(y)$. Relative to certain machines, therefore, sequence prediction on the basis of the shortest encoding is nearly as accurate as prediction on the basis of all encodings.[9]

An analogous but somewhat stronger result can be proved to the effect that a process exists corresponding to p in which *each* digit of $p(y)$ is determined by exactly one encoding of y. To do so, however, the notion of encoding used so far needs to be modified, as Willis' result would be patently false on the basis of that notion. For consider the predictor $p(1^n) = 1$ for all $n \in N$, all other values being zero. Although a process f can be constructed such that $\min\{x|\, f(x) \geqslant 1^n\}$ grows arbitrarily slowly with n, this minimum must nevertheless grow unboundedly and hence the fractional contribution of any minimal encoding of 1^n to $p_f(1^n)$ must approach 0 as $n \to \infty$.

This difficulty in reformulating Willis' result disappears if the following "liberalized" notion of encoding is used. x is said to be a *reduced encoding* of y (symbolically, $f(x) \gg y$), iff there is a finite pf $S \subset X^*$ such that $\sigma S = 1$ and $f(xS) \subset yX^*$, and no such S exists

[9] Solomonoff, Willis, and Chaitin have all commented on the relationship between sequence prediction and "scientific" prediction. Willis' result about the efficacy of the shortest encoding seems related to the efficacy of the simplest (shortest) theory in scientific prediction.

for any proper prefix of x.[10] This is still a reasonable notion of encoding, since it is possible to generate y given $|y|$ and a reduced encoding of y. Furthermore, encodings could be replaced by reduced encodings in the definition of Solomonoff probabilities, i.e.,

$$\sigma\{x|\, f(x) \gg y\} = \sigma\{x|\, f(x) \succ y\} = p_f(y).$$

In the following corollary "$\exists_1$" denotes "there exists exactly one".

Corollary 2. p is a binary computable distribution iff there exists an endless recursive process f such that for all $y \in X^*$

$$p(y) = p_f(y) = \sum_{n=0}^{\infty}[\exists_1 x : |x| = n \;\&\; f(x) \gg y]2^{-n}.$$

Proof. Again only a slight modification of the construction in the first part of Th. 13 is needed. The modification ensures that the "intervals" chosen for $f^{-1}(\{y0, y1\})$ (see Fig. 1) finitely partition the "intervals" previously chosen for $f^{-1}(\{y\})$ in such a way that to each digit of $p(y0)$ or $p(y1)$ contributing 2^{-i} to the probability, there corresponds a set of adjacent intervals whose union represents some i-sequence. That such a partitioning always exists can be proved by induction on $|y|$. □

6. Concluding Remarks

It has been shown that the notion of a predictor provides a common basis for the study of randomness and the study of probabilistic sequence prediction. The random sequences are those which are irredundant with respect to all effective predictors, and all effective predictors are obtained by assigning high probabilities to sequences with short and/or numerous encodings with respect to some effective process. It was also suggested that a minimal constraint on any "rational" method of prediction is that all sequences obey a law of large numbers relative to it, while the requirement that all sequences should appear to behave randomly relative to it is too strong.

A new proof of the existence of an optimal incrementable predictor was given. The fact that this predictor is not computable detracts from its "practical" interest. Perhaps more interesting than the optimal predictor itself is its method of construction. Since this is based on the recursive enumerability of the class of predictors under consideration, a similar construction is possible for more restricted classes of predictors, e.g., the predictors derived from the primitive recursive functions. Thus there will be predictors which are optimal within "practical" classes of functions, whenever the weighted sum of predictors stays within the class under consideration. It should not be hard to prove (or ensure) that such predictors are also "rational".

An open question is whether a process can be found corresponding to any incrementable predictor p such that the highest-order digit of $p(x)$, $x \in X^*$, is determined by the *shortest* reduced encoding of x relative to the process. An affirmative answer

[10] A correspondence can be established between processes and Willis' concrete model of computation (Willis, 1970) such that "$f(x) \gg y$" becomes equivalent to "x is an $|x|$-program for y".

would give the analogue of Th. 13, Cor. 2 for incrementable predictors. Another set of questions concerns the classification of randomness tests according to the growth in redundancy of sequences rejected by such tests. Such a classification should be easily obtainable from Schnorr's classification of randomness tests according to the growth of martingales (Schnorr, 1971b). An entirely different set of questions arises if the *difficulty* of predicting sequences which are predictable to some degree is investigated. Some of Schnorr's (1971b) work on complexity-based degrees of randomness pertains to these questions.

Acknowledgements The author is indebted to Dr. Amram Meir of the Department of Mathematics and to Dr. Arthur Wouk of the Department of Computing Science of the University of Alberta for steering him away from a faulty version of Th. 13. The research was supported by the National Research Council of Canada under Operating Grant A8818.

References

ABRAMSON, N. (1963), *Information Theory and Coding*, McGraw-Hill, New York, N.Y.

BLUM, L., and BLUM, M. (1973), "Inductive inference: a recursion theoretic approach", Memo. No. ERL-M386, Electronics Res. Lab., Univ. of Calif., Berkeley; also, "Towards a mathematical theory of inductive inference," *Inform. Contr. 28* (1975), 125-155.

CHAITIN, G. J. (1966), "On the length of programs for computing finite binary sequences," *J. Ass. Comp. Mach. 13*, 547-569.

CHAITIN, G. J. (1969), "On the length of programs for computing finite binary sequences: statistical considerations," *J. Ass. Comp. Mach. 16*, 145-159.

CHAITIN, G. J. (1970), "On the difficulty of computations," *IEEE Trans. Inf. Theory IT-16*, 5-9.

CHAITIN, G. J. (1974), "Information-theoretic computational complexity," *IEEE Trans. Inf. Theory IT-20*, 10-15.

CHAITIN, G. J. (1975), "A theory of program size formally identical to information theory," *J. Assoc. Comp. Mach. 22*, 329-340.

CHAITIN, G. J. (1976), "Information-theoretic characterizations of recursive infinite strings," *Theoret. Comput. Sci. 2*, 45-48.

CHAITIN, G. J. (1977), "Algorithmic information theory," *IBM J. Res. Develop. 21*, 350-359.

CHURCH, A. (1940), "On the concept of a random sequence," *Bull. Amer. Math. Soc. 46*, 130-135.

COVER, T. M. (1974), "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Statistics Dept. Rep. 12, Stanford Univ., Stanford, CA.

DE LEEUW, K., MOORE, E. F., SHANNON, C. E., and SHAPIRO, N. (1956), "Computability by Probabilistic Machines," Automata Studies, Princeton Univ., Princeton, N. J.

GOLD, E. M. (1967), "Language identification in the limit," *Inform. Contr. 10*, 447-474.

KNUTH, D. E. (1968), *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, Addison-Wesley, Reading, Mass., p 381.

KOLMOGOROV, A. N. (1965), "Three approaches to the quantitative definition of information," *Inf. Transmission 1*, 3-11; also *Int. J. Comp. Math. 2* (1968), 157-168.

LEUNG-YAN-CHEONG, S. K., and COVER, T. M. (1975), "Some inequalities between Shannon entropy and Kolmogorov, Chaitin, and extension complexities," Tech. Report no. 16, Statistics Dept., Stanford Univ., Stanford, CA.; to appear in *IEEE IT*.

LEVIN, L. A. (1973), "On the notion of a random sequence," *Soviet Math. Doklady 14*, 1413-1416.

LEVIN, L. A. (1976), "Uniform tests of randomness," *Soviet Math. Doklady 17*, 337-340.

LOVELAND, D. W. (1970), "A variant of the Kolmogorov notion of complexity," *Inform. Contr. 15*, 510-526.

MARTIN-LÖF, P. (1966), "The definition of random sequences," *Inform. Contr. 9*, 602-619.

MÜLLER, D. W. (1972), "Randomness and extrapolation," *Proc. 6th Berkeley Symp. on Math. Statistics and Probability*, June 21 - July 18, 1970, Le Cam, L. M., Neyman, J., and Scott, E. L. (eds), Univ. Calif. Press, Berkeley and Los Angeles, CA., 1-31.

ROGERS, H. (1967), *Theory of Recursive Functions and Effective Computability,* McGraw-Hill, New York, N. Y.

SCHNORR, C. P. (1971a), "A uniform approach to the definition of randomness," *Math. Systems Theory 5*, 9-28.

SCHNORR, C. P. (1971b), "Zufälligkeit und Wahrscheinlichkeit," *Lecture Notes in Mathematics, Vol. 218*, Springer, Berlin - Heidelberg - New York.

SCHNORR, C. P. (1973), "Process complexity and effective random tests," *J. Computer and Systems Sciences 7*, 376-388.

SOLOMONOFF, R. J. (1964), "A formal theory of inductive inference," *Inform. Contr. 7*, 1-22, 224-254.

SOLOMONOFF, R. J. (1976), "Complexity based induction systems: comparisons and convergence theorems," Report RR-329, Rockford Research, Cambridge, Mass.

VILLE, J. (1939), *Etude Critique de la Notion Collectif,* Gauthier-Villars, Paris.

VON MISES, R. (1919), "Grundlagen der Wahrscheinlichkeitstheorie," *Math. Z. 5*, 52-99.

WALD, A. (1937), "Die Widerspruchsfreiheit des Kollektivbegriffs in der Wahrscheinlichkeitsrechnung," *Ergebnisse eines math. Koll. 8*, 38-72.

WILLIS, D. G. (1970), "Computational complexity and probability constructions," *J. Ass. Comp. Mach. 17*, 241-259.

ZVONKIN, A. K., and LEVIN, L. A. (1970), "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms," *Russian Math. Surveys 25*, 83-124.