[24] A. N. Kolmogorov, "On the approximation of distributions of sums of independent summands by infinitely divisible distributions," *Sankhyā*, vol. 25, pp. 159–174, 1963.
[25] A. Renyi, "On the amount of missing information and the Neyman–Pearson lemma," in *Research Papers in Statistics*, David, Ed. New York: Wiley, 1966, pp. 281–288.
[26] G. T. Toussaint, "Some upper bounds on error probability for multiclass pattern recognition," *IEEE Trans. Comput.*, C-20, pp. 943–944, 1971.
[27] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. New York and London: Cambridge Univ. Press, 1934.
[28] D. E. Daykin and C. J. Eliezer, "Generalization of Hölder's and Minkowsky's inequalities," *Proc. Cambridge Phil. Soc.*, vol. 64, pp. 1023–1027, 1968.
[29] L. Kanal, "Patterns in pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 6, pp. 697–722, 1974.
[30] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

# Some Equivalences Between Shannon Entropy and Kolmogorov Complexity

SIK K. LEUNG-YAN-CHEONG, MEMBER, IEEE, AND THOMAS M. COVER, FELLOW, IEEE

***Abstract*—It is known that the expected codeword length $L_{UD}$ of the best uniquely decodable (UD) code satisfies $H(X) \leq L_{UD} < H(X)+1$. Let $X$ be a random variable which can take on $n$ values. Then it is shown that the average codeword length $L_{1:1}$ for the best one-to-one (not necessarily uniquely decodable) code for $X$ is shorter than the average codeword length $L_{UD}$ for the best uniquely decodable code by no more than $(\log_2 \log_2 n)+3$. Let $Y$ be a random variable taking on a finite or countable number of values and having entropy $H$. Then it is proved that $L_{1:1} > H - \log_2 (H+1) - \log_2 \log_2 (H+1) - \cdots - 6$. Some relations are established among the Kolmogorov, Chaitin, and extension complexities. Finally it is shown that, for all computable probability distributions, the universal prefix codes associated with the conditional Chaitin complexity have expected codeword length within a constant of the Shannon entropy.**

## I. Introduction

SHANNON has shown that the minimal expected length $L$ of a prefix code for a random variable $X$ satisfies

$$H(X) \leq L < H(X)+1 \tag{1}$$

where $H$ is the entropy of the random variable. Shannon's restriction of the encoding or description of $X$ to prefix codes is highly motivated by the implicit assumption that the descriptions will be concatenated and thus must be uniquely decodable. Since the set of allowed codeword lengths is the same for the uniquely decodable and instantaneous codes [1], [2], the expected codeword length $L$ is the same for both sets of codes. Shannon's result follows by assigning codeword length $l_i = \lceil \log 1/p_i \rceil$ to the $i$th outcome of the random variable, where $p_i$ is the probability of the $i$th outcome. Thus the entropy $H$ plays a fundamental role and may be interpreted as the minimal expected length of the description of $X$. The intuition behind the entropy $H$ is so compelling that it would be disconcerting if $H$ did not figure prominently in a description of the most efficient coding with respect to other less constrained coding schemes. In particular we have in mind one-to-one (1:1) codes, i.e., codes which assign a distinct binary codeword to each outcome of the random variable, without regard to the constraint that concatenations of these descriptions be uniquely decodable. It will be shown here that $H$ is also a first order approximation to the minimal expected length of one-to-one codes.

Throughout this paper we use $L_{1:1}$ and $L_{UD}$ to denote the average codeword lengths for the best 1:1 code and uniquely decodable code, respectively. Since the class of 1:1 codes contains the class of uniquely decodable codes, it follows that $L_{1:1} \leq L_{U.D.}$. We show that $L_{1:1} \geq H - \log \log n - 3$ where $n$ is the number of values that the random variable $X$ can take on. Perhaps more to the point, we also show that $L_{1:1} \geq H - \log(H+1) - O(\log \log(H+1))$. Thus, to first order, a 1:1 code allows no more compression than a uniquely decodable or prefix code.

As a consequence of the work of Kolmogorov and Chaitin, a notion of the intrinsic descriptive complexity of a finite object has been developed. This is closely related to the work of Shannon in which the complexity of a class of objects is defined in terms of the probability distribution over that class. The complexity measures of Kolmogorov and Chaitin, together with a new complexity measure which we call the extension complexity, have associated with them universal coding schemes. We shall establish that the universal encoding associated with the complexity of Chaitin [3] and Willis [6] has an expected codeword length with respect to any computable probability distribution on the set of possible outcomes which is within a constant of the Shannon entropy, thus connect-

Manuscript received September 16, 1975; revised September 6, 1977. This work was supported in part by the National Science Foundation under Grants GK-33250, ENG-10173, and ENG 76-03684, and in part by the Air Force Office of Scientific Research under Contract F44620-74-C-0068. This paper was previously presented at the IEEE International Symposium on Information Theory, Ithaca, NY, October 13, 1977.

S. K. Leung-Yan-Cheong was with Stanford University. He is now with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.
T. M. Cover is with the Department of Electical Engineering and Statistics, Stanford University, Stanford, CA.

0018-9448/78/0500-0331$00.75 

ing the individual complexity measure of Chaitin and Kolmogorov with the average statistical complexity measure of Shannon.

In Section II, we consider a random variable which can take on only a finite number of values, and we maximize $(L_{UD}-L_{1:1})$. In Section III we derive lower bounds on $L_{1:1}$ in terms of the entropy of a random variable taking values in a countable set. In Section IV we recall the definitions of the Kolmogorov and Chaitin complexities of binary sequences and introduce the notion of an extension complexity. We then derive some relationships among these quantities. Finally, in Section V we show that, for all computable probability distributions, the universal prefix codes associated with the conditional Chaitin complexity have expected codeword length within a constant of the Shannon entropy.

## II. Maximization of $(L_{UD}-L_{1:1})$

Let $X$ be a random variable (RV) taking on a finite number of values, i.e.,

$$X=\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}.$$

With no loss of generality, assume $p_1 \geq p_2 \geq \cdots \geq p_n$. Let $l_i$, $i=1,2,\cdots,n$ be the lengths of the codewords in the best 1:1 code for encoding the RV $X$, where $l_i$ is the length of the codeword assigned to $x_i$.

*Remark:* Unless otherwise stated, all logarithms are to the base 2. The set of available codewords is $\{0,1,00,01,10,11,000,001,\cdots\}$.

It is clear that the best 1:1 code must have $l_1 \leq l_2 \leq l_3 \leq \cdots$. Thus, by inspection, we have precisely $l_1=1$, $l_2=1$, $l_3=2,\cdots,$

$$l_i=\left\lceil \log\left(\frac{i}{2}+1\right)\right\rceil \tag{2}$$

and

$$L_{1:1}=\sum_{i=1}^{n} p_i l_i=\sum_{i=1}^{n} p_i\left\lceil \log\left(\frac{i}{2}+1\right)\right\rceil. \tag{3}$$

We now prove the following theorem which gives an upperbound on $(L_{UD}-L_{1:1})$.

*Theorem 1:*

$$L_{1:1} \geq L_{UD}-\log\log n-3. \tag{4}$$

*Proof:* From (1) we have $L_{UD}<H(X)+1$. Therefore

$$\max\,(L_{UD}-L_{1:1})<1+\max\,(H(X)-L_{1:1}). \tag{5}$$

Noting from (3) that

$$L_{1:1} \geq \sum_{i=1}^{n} p_i \log\left(\frac{i}{2}+1\right), \tag{6}$$

we can write

$$H(X)-L_{1:1} \leq \sum_{i=1}^{n} p_i\left(\log\frac{1}{p_i}-\log\left(\frac{i}{2}+1\right)\right). \tag{7}$$

We then use the method of Lagrange multipliers to maximize the right side of (7). The proof is completed by using (5). Details of the proof are given in Appendix A.

## III. Lower Bounds on $L_{1:1}$ in Terms of the Entropy $H$

The objective in this section is to obtain lower bounds on $L_{1:1}$ in terms of the entropy $H$ of the random variable. As a first step, we consider transformations of 1:1 to UD codes. The random variables considered may take on a countable number of values.

### *Some Possible Transformations from 1:1 to UD Codes*

The aim here is to find efficient means of transforming 1:1 codes to UD codes.

Let $l_1,l_2,\cdots$ be the lengths of the codewords for the best 1:1 code; assume $l_1 \leq l_2 \leq \cdots$.

Let $f$ be any function such that $\sum_i 2^{-f(l_i)} \leq 1$. Then from Kraft's inequality, the set of lengths $\{\lceil f(l_i)\rceil\}$ yields acceptable word lengths for a prefix (or UD) code. If $f$ is integer-valued and $\sum_i 2^{-f(l_i)}>1$, $\{f(l_i)\}$ cannot yield a prefix code.

*Theorem 2:* The following functions represent possible transformations from 1:1 to UD codes.

$$\text{i)}\quad f(l_i)=l_i+a\lceil \log l_i\rceil+\log\left(\frac{2^a-1}{2^a-2}\right), \qquad \text{where } a>1; \tag{8}$$

$$\text{ii)}\quad f(l_i)=l_i+2\lfloor \log\,(l_i+1)\rfloor; \tag{9}$$

$$\text{iii)}\quad f(l_i)=l_i+\lfloor \log l_i+\log\,(\log l_i)+\cdots\rfloor+4. \tag{10}$$

The proof of Theorem 2 follows from verification of the Kraft inequality for $f(l_i)$ and is given in Appendix B.

We now make use of Theorem 2 to prove some lower bounds on $L_{1:1}$ in terms of the entropy $H$.

*Theorem 3:* The expected length $L_{1:1}$ of the best 1:1 code satisfies the following lower bounds

$$\text{i)}\quad L_{1:1} \geq H-a(1+\log\,(H+1))-\log\left(\frac{2^a-1}{2^a-2}\right) \qquad \text{where } a>1; \tag{11}$$

$$\text{ii)}\quad L_{1:1} \geq H-2\log\,(H+2); \tag{12}$$

$$\text{iii)}\quad L_{1:1} \geq H-\log(H+1)-\log\log\,(H+1)-\cdots-6. \tag{13}$$

*Proof:* i) From Theorem 2 i) and the fact that the expected length for a UD code $\geq H(X)$, we can write

$$E\left(l+a\lceil \log l\rceil+c\right) \geq H, \qquad \text{where } a>1,\ c=\log\left(\frac{2^a-1}{2^a-2}\right).$$

Therefore $El+a(1+E\ \log\ l)+c \geq H$ where $El=L_{1:1}$. From Jensen's inequality and the convexity of $-\log l$, we have $El+a+a\ \log\ El+c \geq H$. But $El<H+1$, since $l$ corresponds to the best 1:1 code which is certainly better than the best prefix code, and we know that the expected length for the best prefix code is less than $(H+1)$. Thus

$$El \geq H-a(1+\log\,(H+1))-\log\left(\frac{2^a-1}{2^a-2}\right)$$

ii) From Theorem 2 ii) and the fact that $L_{UD} \geq H$, we have

$$E\left(l+2\lfloor \log\,(l+1)\rfloor\right) \geq H,$$

$$El+2E\,\log\,(l+1) \geq H.$$

By Jensen's inequality, $El+2\log(El+1)\geq H$. But $El<H+1$ as before. Thus

$$El+2\log(H+2)\geq H$$

$$L_{1:1}\geq H-2\log(H+2).$$

iii) From Theorem 2 iii) and the fact that $L_{UD}\geq H$, we have

$$E\left(l+\lfloor\log l+\log(\log l)+\cdots\rfloor+4\right)\geq H. \tag{14}$$

Thus

$$E(l+\log l+\log(\log l)+\cdots+4)\geq H. \tag{15}$$

*Definition:* For convenience we will define the function $\log^* n$ by

$$\log^* n \triangleq \log n+\log\log n+\cdots, \tag{16}$$

stopping at the last positive term. Then

$$E(l+\log^* l+4)\geq H. \tag{17}$$

Although $\log^* l$ is not concave, we prove in Appendix C that there exists a (piecewise-linear) concave function $F^*(l)$ such that $F^*(l)\leq\log^* l<F^*(l)+2$. Thus $E\log^* l\leq EF^*(l)+2\leq F^*(El)+2\leq\log^*(El)+2$ yielding, from (17),

$$El+\log^* El+6\geq H. \tag{18}$$

But $El<H+1$ as before. Therefore

$$L_{1:1}\geq H-\log(H+1)-\log\log(H+1)-\cdots-6. \tag{19}$$

## IV. Some Relations Between Kolmogorov, Chaitin, and Extension Complexities

Let $\{0,1\}^*$ denote the set of all binary finite length sequences, including the empty sequence. For any $x=(x_1,x_2,\cdots)\in\{0,1\}^*\cup\{0,1\}^\infty$, let $x(n)=(x_1,x_2,\cdots,x_n)$ denote the first $n$ bits of $x$.

*Definition:* A subset $S$ of $\{0,1\}^*$ is said to have the *prefix property* if and only if no sequence in $S$ is the proper prefix of any other sequence in $S$.

For example, $\{00,100\}$ has the prefix property, but $\{00,001\}$ does not.

*Definition:* The *Kolmogorov complexity* of a binary sequence $x(n)\in\{0,1\}^n$ with respect to a partial recursive function $A:\{0,1\}^*\times\mathbb{N}\to\{0,1\}^*$ is defined to be

$$K_A(x(n)|n)=\min_{A(p,n)=x(n)} l(p) \tag{20}$$

where $l(\cdot)$ is the length of the sequence $p$, and $\mathbb{N}$ denotes the set of natural numbers.

Here $A$ may be considered to be a computer, $p$ its program, and $x$ its output. We shall use interchangeably the recursive function theoretic terminology and computer terminology. (See, for example, Chaitin [3] for a discussion of the equivalence of the two.)

*Definition:* Let $U:\{0,1\}^*\to\{0,1\}^*$ be a partial recursive function with a prefix domain. Then the *Chaitin complexity* of a binary sequence $x$ with respect to $U$ is given by

$$C_U(x)=\min_{U(p)=x} l(p). \tag{21}$$

We now introduce a new complexity measure that is useful in prediction and inference.

*Definition:* Let $U:\{0,1\}^*\to\{0,1\}^*$ be a partial recursive function with a prefix domain. Then the *extension complexity* of a binary sequence $x$ with respect to $U$ is defined by

$$E_U(x)=\min_{U(p)\supseteq x} l(p) \tag{22}$$

where $U(p)\supseteq x$ means that $U(p)$ is an extension of $x$, or equivalently that $x$ is a prefix of $U(p)$.

*Definition:* Given a complexity measure $C_B^*:\Omega\to N$ where $\Omega$ is countable and $B$ is a partial recursive function, we say that $C^*$ is *universal* if there exists a partial recursive function $U_0$ such that for any other partial recursive function $A$, there exists a constant $c$ such that for all

$$\omega\in\Omega,\quad C^*_{U_0}(\omega)\leq C^*_A(\omega)+c. \tag{23}$$

It has been shown [3], [4] that the Kolmogorov and Chaitin complexity measures are universal. The same result can be shown to hold for the extension complexity measure. Thus from now on we will assume that the complexities are measured with respect to some fixed appropriate universal function, and the subscripts will be dropped. We shall denote the Chaitin, Kolmogorov, and extension complexities of a binary sequence $x\in\{0,1\}^*$ by $C(x)$, $K(x|l(x))$, and $E(x)$, respectively.

*Theorem 4:* There exist constants $c_0$ and $c_1$ such that for all $x\in\{0,1\}^*$,

$$\begin{aligned}E(x)+c_0\leq C(x)&\leq E(x)+\log l(x)\\&\quad+\log\log l(x)+\cdots+c_1\\&=E(x)+\log^* l(x)+c_1.\end{aligned} \tag{24}$$

*Proof:* The first inequality follows directly from the definitions of $E(x)$ and $C(x)$. To prove the second inequality, note that the Chaitin complexity program $p'$ can be constructed from the extension complexity program $p$ as follows. Let $s$ be the shortest program (from a set having the prefix property) for calculating $l(x)$. Then $p'$ is the concatenation $qsp$ where $q$ consists of a few bits to tell the computer to expect two programs and interpret them appropriately. So we have

$$C(x)\leq E(x)+C(l(x))+c_2. \tag{25}$$

From Theorem 2 iii)

$$C(l(x))\leq\log l(x)+\log\log l(x)+\cdots+c_3. \tag{26}$$

Combining (26) and (27) yields Theorem 4.

Let

$$C(x(n)|n)=\min_{U(p,n^*)=x(n)} l(p) \tag{27}$$

be the (conditional) Chaitin complexity of $x(n)$ given $n$, where $n^*$ is the shortest length binary program for $n$ (see Chaitin [3] for definitions of conditional complexities). As before, the domain of $U(\cdot,n^*)$ has the prefix property for each $n$.

The conditional Chaitin complexity of $x$ given its length $l(x)$ and the unconditional Chaitin complexity of $x$ are closely related in the following sense.

*Theorem 5:* There exist constants $c_0$ and $c_1$ such that for all $x \in \{0,1\}^*$,

$$C(x|l(x)) + c_0 \leqslant C(x) \leqslant C(x|l(x)) + \log^* l(x) + c_1. \tag{28}$$

*Proof:* The lower bound follows from Chaitin [3, Theorem 3.1.e]. The upper bound follows from Chaitin [3, Theorems 3.1.d, 3.1.f] where it is shown that

$$C(x) \leqslant C(x, l(x)) + O(1) \leqslant C(x|l(x)) + C(l(x)) + O(1).$$

But from Theorem 2 iii), $C(l(x)) \leqslant \log^* l(x) + O(1)$. Hence the theorem is proved.

*Theorem 6:* There exist constants $c_0$ and $c_1$ such that for all $x \in \{0,1\}^*$,

$$K(x|l(x)) + c_0 \leqslant C(x) \leqslant K(x|l(x)) + \log K(x|l(x)) + \cdots + \log l(x) + \log\log l(x) + \cdots + c_1. \tag{29}$$

*Proof:* The first inequality is a direct consequence of the definitions. To prove the second inequality, we first note that the Chaitin complexity measure is defined with respect to a computer whose programs belong to a set with the prefix property. From Theorem 2 iii), we know that we can transform the domain of a Kolmogorov complexity measure computer into one which has the prefix property by extending the length of the Kolmogorov complexity program from $K(x|l(x))$ to $K(x|l(x)) + \log K(x|l(x)) + \cdots + c_2$. Let us denote this extended program by $p$. From the proof of Theorem 4, we also know that a program $s$ (belonging to a set with the prefix property) which describes the length of $x$ need not be longer than $\log l(x) + \log\log l(x) + \cdots + c_3$. The Chaitin complexity program can be the concatenation $qsp$ where $q$ consists of a few bits to tell the computer to expect two programs and interpret them appropriately. So

$$C(x) \leqslant K(x|l(x)) + \log K(x|l(x)) + \cdots + \log l(x) + \log\log l(x) + \cdots + c.$$

This completes the proof of Theorem 5.

## V. Relation of Chaitin Code Length to Shannon Code Length

Let $\{X_i\}_1^\infty$ be a stationary binary stochastic process with marginals $p(x(n)), x(n) \in \{0,1\}^*$, $n = 1, 2, \cdots$, and Shannon entropy

$$H(X) = \lim_{n\to\infty} H(X_1, X_2, \cdots, X_n)/n. \tag{30}$$

The Shannon entropy $H(X_1, \cdots, X_n)$ is a real number, while the Chaitin complexity $C(X_1, \cdots, X_n|n)$ is a random variable equal to the length of the shortest codeword (program) assigned to $(X_1, \cdots, X_n)$ by $U$. The prefix set of codewords so defined may be thought of as a universal prefix encoding of $n$-sequences for each $n$. Note in particular that the prefix encoding induced by $U$ is completely oblivious to the true underlying statistics $p(x_1, \cdots, x_n)$. We shall show, however, that this universal encoding has an expected word length equal to first order to the optimal Shannon bound $H(X_1, \cdots, X_n)$.

First we remark that Levin [7] has asserted (the proof does not appear) that for any finite alphabet ergodic process (with computable probability distribution) $(1/n)K(X_1, \cdots, X_n|n) \to H(X)$ with probability one. Thus from Theorem 5 it follows that $(1/n)C(X_1, X_2, \cdots, X_n|n) \to H(X)$ with probability one. We shall show that the behavior of $C$ is good for finite $n$, for all $n$.

*Theorem 7:* For every computable probability measure $p: \{0,1\}^* \to [0,1]$ for a stochastic process, there exists a constant $c$ such that for all $n$

$$H(X_1, \cdots, X_n) \leqslant E_p C(X_1, \cdots, X_n|n) \leqslant H(X_1, \cdots, X_n) + c. \tag{31}$$

*Proof:* For each $n$, $C(x(n)|n)$, $x(n) \in \{0,1\}^n$ must satisfy the Kraft inequality. So we have

$$H(X_1, \cdots, X_n) \leqslant E_p C(X_1, \cdots, X_n|n). \tag{32}$$

For the right half of the inequality, we must use a theorem of Chaitin and Willis relating $C$ and a certain universal probability measure $P^*$. We then relate $P^*$ to the true distribution $P$ to achieve the desired proof. We define, for some universal computer $U$,

$$P^*(x(n)|n) = \sum_{U(p,n^*) = x(n)} 2^{-l(p)}. \tag{33}$$

Chaitin has shown [3, Theorem 3.5] (see also Willis [6, Theorem 16]) that there exists a constant $c'$ such that

$$C(x(n)|n) \leqslant \log \frac{1}{P^*(x(n)|n)} + c' \tag{34}$$

for all $n$. In addition, he has shown that for any other prefix domain computer $A$, there exists a constant $c''$ such that

$$P^*(x(n)|n) \geqslant c'' P_A(x(n)|n) \tag{35}$$

for all $n$, where $P_A(\cdot)$ is defined as in (33).

In Lemma 1 below we show that, for the given computable probability mass function $p: \{0,1\}^* \to [0,1]$ for a stochastic process, there exists a prefix domain computer $A$ such that $P_A(x(n)|n) = p(x(n))$ for all $n$. The proof can then be completed as follows

$$E_p C(x(n)|n) = \sum_{x(n)\in\{0,1\}^n} p(x(n)) C(x(n)|n) \tag{36}$$

$$\leqslant \sum_{x(n)\in\{0,1\}^n} p(x(n)) \left( \log \frac{1}{P^*(x(n)|n)} + c' \right), \text{ using (34)}, \tag{37}$$

$$\leqslant \sum_{x(n)\in\{0,1\}^n} p(x(n)) \left( \log \frac{1}{c'' P_A(x(n)|n)} \right) + c', \text{ using (35)}, \tag{38}$$

$$= \sum_{x(n)\in\{0,1\}^n} p(x(n)) \log \frac{1}{p(x(n))} + c''', \text{ using Lemma 1}, \tag{39}$$

$$= H(X_1, \cdots, X_n) + c''', \quad \text{for all } n. \tag{40}$$

Q.E.D.

*Lemma 1:* For any computable probability mass function $p:\{0,1\}^*\to[0,1]$ for a stochastic process, there exists a prefix domain computer $A$ such that $P_A(x(n)|n)=p(x(n))$ for all $n$.

*Remark 1:* Willis [6, Theorem 12] has proved a similar lemma under the constraint that $p(\cdot)$ be "$r$-computable," i.e., that $p(x_1,\cdots,x_n)$ have a finite base-$r$ expansion for every $x_1,x_2,\cdots,x_n$.

*Remark 2:* Here we define a number to be computable if we can calculate its $n$th bit in finite time for all finite $n$. An analogous result can be proved if by a computable number we mean instead of a number which we can approximate arbitrarily closely.

*Proof:* Let $p^{(k)}(x(n))$ denote $p(x(n))$ truncated after $k$ bits. For example, if $p(x(n))=0.001011001\cdots$, then $p^{(5)}(x(n))=0.00101$. Define

$$F^{(k)}(x(n))=\sum_{x'(n)<x(n)} p^{(k)}(x'(n)) \tag{41}$$

where $x'(n)<x(n)$ means $x'(n)$ precedes $x(n)$ in a lexicographic ordering of the $n$-sequences. Note that $p(x(n))$ being computable does not guarantee that $F(x(n))$ is computable.

Let $A$ be a computer that has $n^*$ on its work tape. It also has at its disposal for inspection a random program $p=p_1p_2p_3p_4\cdots\in\{0,1\}^\infty$. We now describe how $A$ operates.

*Step 1:* Calculate $n$.

*Step 2:* Set $m=1$.

*Step 3:* Compute $F^{(m)}(x(n))$, for all $x(n)\in\{0,1\}^n$.

*Step 4:* The error in summing $2^n$ binary terms each in $[0,1]$ and each truncated after $m$ places is bounded above by $2^{n-m}$. Using this crude bound on the difference between $F^{(m)}(x(n))$ and the true distribution function $F(x(n))\triangleq\Sigma_{x'(n)<x(n)}p(x'(n))$, and between $\cdot p^{(m)}=\cdot p_1p_2\cdots p_m$ and $\cdot p$, decide if at this stage it can be guaranteed that

$$\cdot p\in\left(F(x^*(n)),F\left(x^*(n)+\underbrace{00\cdots00}_{n-1}1\right)\right] \tag{42}$$

for some $x^*(n)\in\{0,1\}^n$. Here $x(n)+00\cdots001$ means the sequence obtained by adding $\cdot x(n)$ and $(\frac{1}{2})^n$ and reinterpreting it as a sequence. If (42) can be decided, proceed to step 6.

*Step 5:* Increment $m$ by 1. Go back to Step 3.

*Step 6:* Print out $x^*(n)$ and stop.

It is easily seen that

$$\Pr\left\{\cdot p\in\left(F(x(n)),\quad F\left(x(n)+\underbrace{00\cdots00}_{n-1}1\right)\right]\right\}=p(x(n)) \tag{43}$$

for all $x(n)\in\{0,1\}^n$. Since $\lim_{m\to\infty}\cdot p^{(m)}=\cdot p$ and $\lim_{m\to\infty}F^{(m)}(x(n))=F(x(n))$, $A$ will fail to halt only if $\cdot p=F(x(n))$ for some $x(n)\in\{0,1\}^n$. This event has probability zero. Thus there exists a computer $A$ such that a Bernoulli random program $p$ will induce the stochastic process $\{X_i\}$ as its output. Q.E.D.

## VI. Conclusions

This study can be perceived in three parts. First, the minimal average code length with respect to a known distribution has been shown to be equal to the Shannon entropy $H$ to first order under different coding constraints. Second, the individual complexity measures of Kolmogorov, Chaitin, and others have been shown to be equivalent to one another, also to first order. Finally, the expected code length of the individual algorithmic code has been shown to be equal to first order to the Shannon entropy, thus identifying the statistical and the logical definitions of entropy.

## Acknowledgment

The authors would like to thank Professor John T. Gill for suggesting the method used for lower bounding $L_{1:1}$ in Section III. They also wish to thank both referees for aid in improving the proofs and making the concepts more precise.

## Appendix A: Proof of Theorem 1.

*Theorem 1:*

$$L_{1:1}\geq L_{\mathrm{UD}}-\log\log n-3.$$

*Proof:* From (1)

$$\max\left(L_{\mathrm{UD}}-L_{1:1}\right)<1+\max\left(H(X)-L_{1:1}\right). \tag{A1}$$

We now proceed to find $\max(H(X)-L_{1:1})$. Let $A\triangleq H(X)-L_{1:1}$. Then

$$A=\sum_{i=1}^n p_i\log\frac{1}{p_i}-\sum_{i=1}^n p_i\left\lceil\log\left(\frac{i}{2}+1\right)\right\rceil \tag{A2}$$

$$\leq\sum_{i=1}^n p_i\left(\log\frac{1}{p_i}-\log\left(\frac{i}{2}+1\right)\right), \tag{A3}$$

$$\max A\leq\max\sum_{i=1}^n p_i\left(\log\frac{1}{p_i}-\log\left(\frac{i}{2}+1\right)\right). \tag{A4}$$

Let

$$c_i\triangleq\ln\left(\frac{i}{2}+1\right). \tag{A5}$$

Let

$$J(p_1,\cdots,p_n)\triangleq\sum_{i=1}^n p_i\left(\ln\frac{1}{p_i}-c_i\right)+\lambda\sum_{i=1}^n p_i. \tag{A6}$$

Differentiating $J(p_1,\cdots,p_n)$ with respect to $p_i$, we obtain

$$\frac{\partial J}{\partial p_i}=-c_i+\lambda-1+\ln\frac{1}{p_i}. \tag{A7}$$

Setting $\partial J/\partial p_i=0$, we obtain

$$\ln p_i=\lambda-(c_i+1) \tag{A8}$$

i.e.,

$$p_i=e^{\lambda-(c_i+1)}=ae^{-c_i} \tag{A9}$$

where $a$ is some constant. Now

$$\sum_{i=1}^n p_i=1. \tag{A10}$$

Substituting (A9) in (A10) and using (A5) we get

$$2a\sum_{i=1}^{n}\frac{1}{i+2}=1. \tag{A11}$$

Let

$$H_k \triangleq \sum_{i=1}^{k}\frac{1}{i}. \tag{A12}$$

Then (A11) can be rewritten as

$$2a(H_{n+2}-H_2)=1. \tag{A13}$$

From (A9) and (A5)

$$p_i=\frac{2a}{i+2}. \tag{A14}$$

Therefore

$$\begin{aligned}\max \sum_{i=1}^{n} p_i\left(\ln\frac{1}{p_i}-\ln\left(\frac{i}{2}+1\right)\right)&=\sum_{i=1}^{n}\frac{2a}{(i+2)}\left(\ln\left(\frac{i+2}{2a}\right)-\ln\left(\frac{i+2}{2}\right)\right)\\&=\ln\left(\frac{1}{a}\right)\cdot 2a\sum_{i=1}^{n}\frac{1}{(i+2)}\\&=\ln\left(\frac{1}{a}\right).\end{aligned} \tag{A15}$$

So from (A4)

$$\max A \leqslant \log\left(\frac{1}{a}\right)=1+\log(H_{n+2}-H_2). \tag{A16}$$

Using (A1) we obtain

$$\max(L_{\mathrm{UD}}-L_{1:1})<2+\log(H_{n+2}-H_2). \tag{A17}$$

Knuth [5] has

$$H_n=\ln n+\gamma+\frac{1}{2n}-\frac{1}{12n^2}+\frac{1}{120n^4}-\epsilon, \tag{A18}$$

where $0<\epsilon<1/252n^6$ and $\gamma=0.577\cdots$ is Euler's constant. Therefore,

$$\max(L_{\mathrm{UD}}-L_{1:1})<2+\log(\ln(n+2)) \tag{A19}$$

$$\leqslant 3+\log\log n. \tag{A20}$$

Thus we conclude that

$$L_{1:1}\geqslant L_{\mathrm{UD}}-\log\log n-3. \tag{A21}$$

## Appendix B: Admissible Lengths for Uniquely Decodable Codes.

In this appendix, we prove Theorem 2 which states that the following functions represent possible transformations from 1:1 to UD codes. Recall $l_i=\lceil\log(i/2+1)\rceil$.

i) $f(l_i)=l_i+a\lceil\log l_i\rceil+\log((2^a-1)/(2^a-2))$, where $a>1$; (B1)

ii) $f(l_i)=l_i+2\lfloor\log(l_i+1)\rfloor$; (B2)

iii) $f(l_i)=l_i+\lfloor\log l_i+\log(\log l_i)+\cdots\rfloor+4$. (B3)

*Proof of i):* Define

$$\begin{aligned}S&=\sum_{i=1}^{\infty}2^{-f(l_i)}\\&=\sum_{i=1}^{\infty}2^{-l_i}2^{-a\lceil\log l_i\rceil}2^{-c}.\end{aligned} \tag{B4}$$

But there are $2^k$ 1:1 codewords of length $k$. Therefore

$$S=2^{-c}\sum_{l=1}^{\infty}\frac{1}{2^{a\lceil\log l\rceil}} \tag{B5}$$

$$=2^{-c}\left(\frac{1}{2^0}+\frac{1}{2^a}\cdot 2^0+\frac{1}{2^{2a}}\cdot 2^1+\frac{1}{2^{3a}}\cdot 2^2+\cdots+\frac{1}{2^{ka}}\cdot 2^{k-1}+\cdots\right) \tag{B6}$$

$$=2^{-c}\left(\frac{2^a-1}{2^a-2}\right). \tag{B7}$$

$S$ diverges if $a\leqslant 1$. To make $S\leqslant 1$, it is sufficient (and necessary) to have $a>1$ and $C\geqslant\log[(2^a-1)/(2^a-2)]$. This completes the proof of i).

*Proof of ii):* In this case, define

$$\begin{aligned}S&=\sum_{i=1}^{\infty}2^{-f(l_i)}\\&=\sum_{i=1}^{\infty}2^{-l_i}2^{-2\lfloor\log(l_i+1)\rfloor}\end{aligned} \tag{B8}$$

$$=\sum_{l=1}^{\infty}\frac{1}{2^{2\lfloor\log(l+1)\rfloor}}, \quad \text{using the fact that there are } 2^k \text{ 1:1 codewords of length } k, \tag{B9}$$

$$=\left(\frac{1}{2^2}\right)\cdot 2+\left(\frac{1}{2^{2\cdot 2}}\right)2^2+\left(\frac{1}{2^{2\cdot 3}}\right)2^3+\cdots+\left(\frac{1}{2^{2\cdot k}}\right)2^k+\cdots \tag{B10}$$

$$=\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\cdots+\frac{1}{2^k}+\cdots \tag{B11}$$

$$=\frac{1}{2}\left(\frac{1}{1-\frac{1}{2}}\right)=1. \tag{B12}$$

This proves ii).

*Proof of iii):* Let

$$f(l_i)=l_i+\lfloor\log l_i+\log(\log l_i)+\cdots+\log(\log\cdots(\log l_i)\cdots)\rfloor+c$$

where it is understood that we only consider the first $k$ iterates for which $\log(\log(\cdots(\log l_i)\cdots))$ is positive, for example, if $l_i=2$, $f(l_i)=2+1+c=3+c$, and if $l_i=5$, $f(l_i)=5+\lfloor 2.322+1.215+0.281\rfloor+c=5+\lfloor 3.818\rfloor+c=8+c$. Now

$$\begin{aligned}S&\triangleq\sum_{i=1}^{\infty}2^{-f(l_i)}\\&=2^{-c}\sum_{i=1}^{\infty}2^{-l_i}2^{-\left\lfloor\log l_i+\log(\log l_i)+\cdots+\log\left(\log(\cdots(\log l_i)\cdots)\right)\right\rfloor}\end{aligned} \tag{B13}$$

$$=2^{-c}\sum_{l=1}^{\infty}2^{-\left\lfloor\log l+\log(\log l)+\cdots+\log\left(\log(\cdots(\log l)\cdots)\right)\right\rfloor}, \tag{B14}$$

since there are $2^k$ 1:1 codewords of length $k$,

$$\leqslant 2^{-c+1}\sum_{l=1}^{\infty}2^{-\left(\log l+\log(\log l)+\cdots+\log\left(\log(\cdots(\log l)\cdots)\right)\right)} \tag{B15}$$

$$=2^{-c+1}\sum_{l=1}^{\infty}\frac{1}{l\log l\log\log l\cdots} \tag{B16}$$

where the denominator of the last expression includes all the first $j$ iterates for which

$$\underbrace{\log\left(\log\left(\cdots(\log l)\cdots\right)\right)}_{j \text{ times}}$$

is greater than 1.

We now prove a lemma which will be useful in bounding (B16).

*Lemma B.1:* Let

$$g_b(x)=\frac{1}{x\log_b x\log_b(\log_b x)\cdots} \tag{B17}$$

where the denominator is to be interpreted as in (B16). Then

$$I_b \triangleq \int_1^\infty \frac{1}{x\log_b x\log_b(\log_b x)\cdots}\,dx = \begin{cases}\infty, & \text{if } b\geqslant e\\ \text{finite}, & \text{if } b<e.\end{cases} \tag{B18}$$

*Proof:*

$$I_e=\int_1^e \frac{1}{x}\,dx+\int_e^{e^e}\frac{1}{x\log_e x}\,dx + \int_{e^e}^{e^{e^e}}\frac{1}{x\log_e x\log_e(\log_e x)}\,dx + \cdots \tag{B19}$$

$$=\log_e x\Big]_1^e+\log_e(\log_e x)\Big]_e^{e^e} +\log_e(\log_e(\log_e x))\Big]_{e^e}^{e^{e^e}}+\cdots \tag{B20}$$

$$=1+1+1+\cdots$$
$$=\infty.$$

(B20) can be verified by inspection. Thus we have shown that if $b\geqslant e$, then $I_b$ diverges.

Now suppose $b<e$. Let $M\triangleq \log_b e$ where $M>1$.

$$I_b=\int_1^\infty \frac{1}{x(M\log_e x)(M\log_e(M\log_e x))\cdots}\,dx \tag{B21}$$

$$\leqslant \int_1^\infty \frac{1}{x(M\log_e x)(M\log_e(\log_e x))\cdots}\,dx \tag{B22}$$

$$\leqslant \int_1^e\frac{1}{x}\,dx+\int_e^{e^e}\frac{1}{Mx\log_e x} + \int_{e^e}^{e^{e^e}}\frac{1}{M^2x\log_e x\log_e(\log_e x)}\,dx + \cdots$$

$$=1+\frac{1}{M}+\frac{1}{M^2}+\cdots \tag{B23}$$

$$=\frac{1}{\left(1-\frac{1}{M}\right)}=\frac{M}{M-1}<\infty, \qquad \text{since } M>1. \tag{B24}$$

This completes the proof of Lemma B.1.

In particular, from (B24) we have

$$I_2\leqslant\frac{\log_2 e}{\log_2 e-1} <3.26. \tag{B25}$$

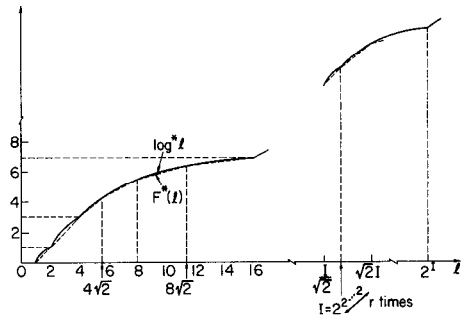



Fig. 1. Sketch of log* $l$ and $F^*(l)$.

But

$$\sum_{l=1}^{\infty}\frac{1}{l\log l\log\log l\cdots}<I_2+1<5. \tag{B26}$$

Using (B16) we obtain $S<5.2^{-c+1}$. If we choose $c=4$, then $S\leqslant 1$. This proves iii).

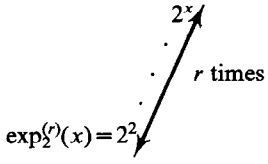
## Appendix C

In this Appendix, we exhibit a piecewise-linear concave function $F^*(l)$ such that

$$F^*(l)\leqslant \log^* l<F^*(l)+2, \qquad l\geqslant 1. \tag{C1}$$

Recall $\log^* l=\log l+\log\log l+\cdots$, stopping at the last positive term (see Fig. 1). The function $F^*(l)$ is also sketched in Fig. 1. For $1\leqslant l\leqslant 4$, $F^*(l)=l-1$ and for $l>4$, $F^*(l)$ is defined as follows. Consider the following sequence of values for $l$: 4, $4\sqrt{2}$, 8, $8\sqrt{2}$, 16, $\cdots$, i.e., a geometric sequence with a ratio of $\sqrt{2}$. Then $F^*(l)$ is obtained by joining adjacent points on the $\log^* l$ curve by straight line segments for the $l$ values mentioned above.

In the following, define




and

$$\log^{(r)}(l)=\underbrace{\log\log\cdots\log}_{r \text{ times}} l,$$

i.e., the $r$-fold composition of the exponential and log functions, respectively.

First we prove the concavity of $F^*(l)$, $l>1$. Let us look at $F^*(l)$ for $l>4$. It is clearly sufficient to prove concavity at points $\exp_2^{(r)}(2)$, $r=3,4,\cdots$ since concavity is automatically satisfied at all other points. Thus we need to show that

$$\frac{f(I)-f\left(\frac{I}{\sqrt{2}}\right)}{I-\frac{I}{\sqrt{2}}}\geqslant\frac{f(\sqrt{2}\,I)-f(I)}{\sqrt{2}\,I-I}, \qquad I=\exp_2^{(r)}(2),\ r=3,4,\cdots, \tag{C2}$$

i.e.,

$$\sqrt{2}\left(f(I)-f\left(\frac{I}{\sqrt{2}}\right)\right)\geqslant f(\sqrt{2}\,I)-f(I) \tag{C3}$$

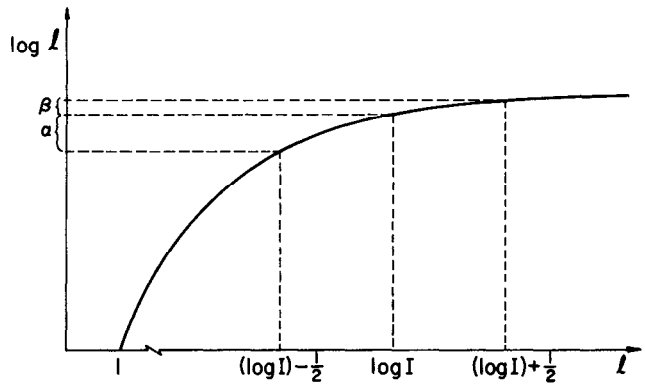



Fig. 2. Graphical interpretation of inequality (C9): $\alpha > \beta$.

where for convenience we have set $\log^* l = f(l)$. By definition,

$$f(I) = \log I + \log\log I + \cdots + \log^{(r)}(I) \tag{C4}$$

$$f\left(\frac{I}{\sqrt{2}}\right) = \left((\log I) - \frac{1}{2}\right) + \log\left((\log I) - \frac{1}{2}\right) + \cdots + \log^{(r-1)}\left((\log I) - \frac{1}{2}\right) \tag{C5}$$

$$f(\sqrt{2}\, I) = \left((\log I) + \frac{1}{2}\right) + \log\left((\log I) + \frac{1}{2}\right) + \cdots + \log^{(r-1)}\left((\log I) + \frac{1}{2}\right) + \log^{(r)}\left((\log I) + \frac{1}{2}\right). \tag{C6}$$

Consider the 1st terms in $f(I)$, $f(I/\sqrt{2})$ and $f(\sqrt{2}\, I)$:

$$\left[f(I) - f\left(\frac{I}{\sqrt{2}}\right)\right]_{\text{1st term}} = \frac{1}{2} \tag{C7}$$

$$\left[f(\sqrt{2}\, I) - f(I)\right]_{\text{1st term}} = \frac{1}{2}. \tag{C8}$$

Considering the 1st terms only, we see that (C3) is satisfied. In fact the difference between the left side and right side of (C3) is $(\sqrt{2} - 1)/2 = 0.207$.

Now consider the 2nd terms in $f(I)$, $f(I/\sqrt{2})$, $f(\sqrt{2}\, I)$. Because the log function is concave, it is clear (see Fig. 2) that

$$\log\log I - \log\left((\log I) - \frac{1}{2}\right) > \log\left((\log I) + \frac{1}{2}\right) - \log\log I. \tag{C9}$$

Considering only the 2nd terms of $f(I)$, $f(I/\sqrt{2})$, $f(\sqrt{2}\, I)$ we see that (C3) is again satisfied. It is clear that by the same argument as above, the 3rd through $r$th terms of $f(I) - f(I/\sqrt{2})$ exceed the corresponding terms of $f(\sqrt{2}\, I) - f(I)$. There is one remaining term in $f(\sqrt{2}\, I)$ which we have to consider, namely $\log^{(r+1)}(\sqrt{2}\, I) \triangleq g(r)$. We now show that $g(r)$ is monotone decreasing in $r$, $r \geq 1$

$$g(r) = \log^{(r+1)}\left(\sqrt{2}\ \exp_2^{(r)}(2)\right) \tag{C10}$$

$$= \log^{(r+2)}\left(2^{\sqrt{2}\ \exp_2^{(r)}(2)}\right) \tag{C11}$$

$$g(r+1) = \log^{(r+2)}\left(\sqrt{2}\ \exp_2^{(r+1)}(2)\right). \tag{C12}$$

So we need to show

$$\left(\exp_2^{(r+1)}(2)\right)^{\sqrt{2}} > \sqrt{2}\ \exp_2^{(r+1)}(2), \tag{C13}$$

i.e.,

$$\left(\exp_2^{(r+1)}(2)\right)^{\sqrt{2}-1} > \sqrt{2}, \tag{C14}$$

which is clearly satisfied for $r \geq 1$.

By inspection $\log\log\log\log 16\sqrt{2} = 0.16$, which is less than $(\sqrt{2} - 1)/2$ so that (C3) is satisfied for $I = \exp_2^{(r)}(2)$, $r \geq 3$. To complete the proof of the concavity of $F^*(l)$, it can easily be verified that concavity of $F^*(l)$ also holds at $l = 4$.

We proceed to show that

$$F^*(l) \leq f(l) \leq F^*(l) + 2, \qquad \text{for } l \geq 1. \tag{C15}$$

Define an auxiliary function $a(l) \triangleq 4 \log l$. Consider the derivative $f'(l)$ of $f(l)$. If $\exp_2^{(r)}(2) \leq l < \exp_2^{(r+1)}(2)$, then

$$f'(l) = \frac{\log e}{l} + \frac{\log e}{\log l} \cdot \frac{\log e}{l} + \cdots + \frac{\log e}{\log^{(r)}(l)} \cdot \frac{\log e}{\log^{(r-1)}(l)} \cdots \frac{\log e}{l}. \tag{C16}$$

We will now show that

$$f'(l) < a'(l), \qquad \text{for } r \geq 2, \tag{C17}$$

i.e.,

$$f'(l) < \frac{4}{l} \cdot \log e, \tag{C18}$$

i.e.,

$$\frac{\log e}{\log l} + \frac{\log e}{\log\log l} \cdot \frac{\log e}{\log l} + \cdots + \frac{\log e}{\log^{(r)}(l)} \cdot \frac{\log e}{\log^{(r-1)}(l)} \cdots \frac{\log e}{\log l} < 3. \tag{C19}$$

It is clear that each term in the left side of (C19) is bounded above by $(\log e)^2/\log l$, and there are $r$ such terms. So it is sufficient to prove that

$$\frac{(\log e)^2}{\log l} \cdot r < 3, \qquad \text{for } r \geq 2. \tag{C20}$$

But $l \geq \exp_2^{(r)}(2)$; hence it is sufficient to prove that

$$\frac{(\log e)^2}{\exp_2^{(r-1)}(2)} \cdot r < 3, \qquad \text{for } r \geq 2, \tag{C21}$$

which is obviously true.

Thus we have shown that for $l \geq 4$, the slope of $f(l)$ is bounded by the slope of $a(l)$. But we know that $a(l)$ increases by 2 when $l$ is multiplied by a factor of $\sqrt{2}$. Therefore $f(l)$, $l \geq 4$, increases by at most 2 every time $l$ is multiplied by $\sqrt{2}$. It is trivial to see that for $1 \leq l \leq 4$ the difference between $f(l)$ and $F^*(l)$ cannot exceed 2. This completes the proof of (C15).

## References


[1] R. Ash, *Information Theory*. New York: Wiley, 1965.
[2] N. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1963.
[3] G. J. Chaitin, "A theory of program size formally identical to information theory," *J. Association for Computing Machinery*, vol. 22, no. 3, pp. 329–340, June 1975.
[4] A. N. Kolmogorov, "Three approaches to the concept of the 'Amount of Information'," *Problemy Peredachi Informatsii*, 1, 1, pp. 3–11, (1965).
[5] D. E. Knuth, *The Art of Computer Programming*, vol. 1, 2nd Ed. Reading, MA: Addison-Wesley, 1973.
[6] D. G. Willis, "Computational complexity and probability constructions," *J. ACM*, vol. 17, no. 2, pp. 241–259, Apr. 1970.
[7] A. K. Zhvonkin, 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," *Russian Math. Surveys* 25, pp. 83–124, 1970.