

A Relation between Complexity and Entropy

James F. Lynch¹

Clarkson University
Department of Mathematics and Computer Science
Potsdam, New York 13699-5815

Abstract

We derive two asymptotic formulas relating the Kolmogorov complexity of strings over a finite alphabet to the entropy of a discrete Markov information source that generates the strings.

Key words: Kolmogorov complexity, information theoretic entropy.

In [2], Beyer, Stein, and Ulam discussed several notions of complexity of integers and made a conjecture relating Kolmogorov complexity and information theoretic entropy. We will state this conjecture, after summarizing the basic notions that it refers to. More comprehensive introductions are [3] and [6] for Kolmogorov complexity, and [1] for information theory.

Let A be an algorithm, i.e. a Turing machine, that transforms binary strings into binary strings. (The restriction to binary strings is for simplicity; all the definitions and results given here easily extend to strings over any finite alphabet.) The *complexity* of a string x relative to A , $K_A(x)$, is the length of the shortest string w such that $A(w) = x$, or if no such string exists, it is ∞ . Similarly, if A is an algorithm that transforms pairs of binary strings into binary strings, the conditional complexity of x with respect to y , $K_A(x|y)$, is the length of the shortest string w such that $A(w, y) = x$, or if no such string exists, it is ∞ . As in [2], we will assume A is of this form, and we will consider conditional complexities $K_A(x|n)$, where $n = |x|$, the length of x .

Let S be a discrete 0-memory binary information source with probability(0) = p and probability(1) = $1 - p$. That is, S generates a sequence of Bernoulli trials whose outcomes are 0 or 1, and for a binary string x of length n with m 0's, the probability of x , $\text{pr}(x)$, is $p^m(1 - p)^{n-m}$. The *entropy* H of S is $-p \log p - (1 - p) \log(1 - p)$. (All our logarithms are base 2.)

¹Research supported by NSF Grant CCR-9006303.

Conjecture 1 (Beyer, Stein, Ulam) *For every natural number n , let x_1, \dots, x_{2^n} be the sequence of all binary strings of length n arranged in order of decreasing probability, as given by S . Let $k(n)$ be the least integer such that $\sum_{1 \leq i \leq k(n)} pr(x_i) > r$ for some fixed $r \in (1/2, 1)$. If K_A is normalized so that*

$$\frac{1}{k(n)} \sum_{1 \leq i \leq k(n)} K_A(x_i|n) = 1$$

when $p = 1/2$ then

$$H \sim \frac{1}{k(n)} \sum_{1 \leq i \leq k(n)} K_A(x_i|n).$$

That is, the most likely strings from A have complexity approximately equal to the entropy of S .

Before stating our theorems, which are similar to the Conjecture, there are two observations worth noting. First, the Conjecture is not true for all A . For example, let A be the identity transformation $A(x) = x$. Then after normalization,

$$\frac{1}{k(n)} \sum_{1 \leq i \leq k(n)} K_A(x_i|n) = 1$$

independently of p , but as is well-known, H is a unimodal function of p with a maximum of 1 at $p = 1/2$ and minima of 0 at $p = 0$ and $p = 1$. However, as we will show, the Conjecture is true when A is a *universal* algorithm or Turing machine. The definition of Kolmogorov complexity usually assumes A is universal. That is, for any algorithm B , there is a string u that encodes B 's program for A : for any strings w and y , $B(w, y) = A(uw, y)$ where uw is the concatenation of u and w . The particular universal algorithm A that is used is not important since $|K_A(x|y) - K_{A'}(x|y)|$ is bounded for any other universal algorithm A' and all x and y .

The second point is that the information source S is a 0-memory source. Our theorems apply to the more general Markov source. A *discrete Markov binary information source* is a finite ergodic Markov chain, say with states s_1, \dots, s_m . From each state there is a transition labelled 0 and another labelled 1. If the chain is in state s_i then the 0 transition will be taken with probability p_i , and the 1 transition will be taken with probability $1 - p_i$. All other transitions have probability 0. Again for simplicity, we assume our Markov source S is regular, i.e. aperiodic. Extending our results to periodic chains is straightforward. S generates a binary string by starting in some arbitrary fixed state, say s_1 , and outputting the labels of the transitions it takes. Let $[a_1 \dots a_m]$ be the stationary distribution of S . Then pr will

be the probability distribution of strings of some length n generated by S . Entropy is now defined by

$$H = - \sum_{1 \leq i \leq m} a_i (p_i \log p_i + (1 - p_i) \log(1 - p_i)).$$

One final technical point is that we assume each p_i is a *computable real number*, i.e. there is an algorithm such that given any natural number l , it generates the first l bits in the binary representation of the number.

Theorem 2 *For every natural number n , let*

$$E(K_A) = \sum_{x \in \{0,1\}^n} K_A(x|n) pr(x)$$

be the average complexity of strings of length n . Then

$$Hn \sim E(K_A).$$

Theorem 3 *Let $r \in (0, 1)$ and $k(n)$ be defined as in the Conjecture. Then*

$$Hn \sim \frac{1}{k(n)} \sum_{1 \leq i \leq k(n)} K_A(x_i|n).$$

Thus, the Conjecture holds for all universal algorithms A and all $r \in (0, 1)$, and the normalization factor is just $1/n$.

These theorems follow from some arguments using basic results about the expectation of random variables (see Feller [5]). For any $i = 1, \dots, m$ and binary string x , let $X(x)$ be the random variable that counts the number of occurrences of state s_i when S generates x , and for $1 \leq t \leq |x|$ let $X_t(x)$ be the indicator random variable whose value is 1 if S is in state s_i at time t .

Lemma 4 *For every $c > 0$*

$$\lim_{n \rightarrow \infty} pr(|X - a_i n| > cn^{1/2} \log n) = 0.$$

Proof. Let $E(X)$ be the expectation of X . We first show that

$$E(X) = a_i n + O(\log n). \tag{1}$$

For $1 \leq t \leq n$ let $q_i^{(t)}$ be the probability that S is in state s_i at time t . By linearity of expectation,

$$E(X) = \sum_{1 \leq t \leq n} E(X_t) = \sum_{1 \leq t \leq n} q_i^{(t)}.$$

By Corollary 4.1.5 in Kemeny and Snell [7], there is a constant $\varepsilon < 1$ such that $|q_i^{(t)} - a_i| < \varepsilon^t$ for sufficiently large t . Therefore there is a constant b such that for $t > b \log n$, $|q_i^{(t)} - a_i| < n^{-2}$. Then

$$\begin{aligned} \sum_{1 \leq t \leq n} q_i^{(t)} &= \sum_{1 \leq t \leq b \log n} q_i^{(t)} + \sum_{b \log n < t \leq n} q_i^{(t)} \\ &= O(\log n) + (n - b \log n)(a_i + O(n^{-2})) \\ &= a_i n + O(\log n). \end{aligned}$$

Next, we show that

$$\mathbb{E}(X^2) = a_i^2 n^2 + O(n \log n). \quad (2)$$

Again by linearity,

$$\mathbb{E}(X^2) = \sum_{1 \leq t \leq n} \mathbb{E}(X_t) + 2 \sum_{1 \leq t < u \leq n} \mathbb{E}(X_t X_u).$$

The second sum on the right can be broken into

$$\sum_{1 \leq t \leq b \log n} \mathbb{E}(X_t X_u) + \sum_{u-t \leq b \log n} \mathbb{E}(X_t X_u) + \sum \mathbb{E}(X_t X_u),$$

where the third sum is over all pairs $t < u$ not included in the first two sums.

$$\begin{aligned} &= O(n \log n) + O(n \log n) + (n^2/2 - O(n \log n))(a_i + O(n^{-2}))^2 \\ &= a_i^2 n^2/2 + O(n \log n). \end{aligned}$$

To finish the proof, by Chebyshev's inequality,

$$\begin{aligned} &\text{pr}(|X - a_i n| \geq cn^{1/2} \log n) \\ &\leq \frac{\mathbb{E}(X^2) - \mathbb{E}(X)^2}{c^2 n (\log n)^2} \\ &= \frac{a_i^2 n^2 + O(n \log n) - (a_i^2 n^2 + O(n \log n))}{c^2 n (\log n)^2} \quad \text{by Equations (1) and (2)} \\ &\rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

□

For every natural number n , let L_n be the set of all strings of length n such that when generated by S , each state s_i occurs between $a_i n - n^{1/2} \log n$

and $a_i n + n^{1/2} \log n$ times, and the number of 0 transitions from s_i is between $a_i p_i n - n^{1/2} \log n$ and $a_i p_i n + n^{1/2} \log n$.

Lemma 5 *We have*

$$\lim_{n \rightarrow \infty} \text{pr}(L_n) = 1.$$

Proof. By Lemma 4 we know that for almost all strings generated by S , there are between $a_i n - n^{1/2} \log n/2$ and $a_i n + n^{1/2} \log n/2$ occurrences of s_i for $i = 1, \dots, m$. Fixing i , we use the Lemma again, applying it to the chain T whose states are σ_0 and σ_1 and matrix of transitions is

$$\begin{bmatrix} p_i & 1 - p_i \\ p_i & 1 - p_i \end{bmatrix}. \quad (3)$$

Each time S leaves state s_i , T will perform one transition, going to σ_0 or σ_1 depending on whether S takes the 0 or 1 transition from s_i . Clearly (3) is also the stationary distribution, so after running T m steps, with probability asymptotic to 1, there will have been between $p_i m - m^{1/2} \log m/4$ and $p_i m + m^{1/2} \log m/4$ occurrences of σ_0 . Alternatively, we could use the fact that the transitions from s_i are Bernoulli trials with probability p_i for success, i.e. a 0 transition. Then the same conclusion follows from Chebyshev's inequality or the DeMoivre-Laplace limit theorem (see Feller [5]). Since in the chain S we can assume there were between $a_i n - n^{1/2} \log n/2$ and $a_i n + n^{1/2} \log n/2$ occurrences of s_i , with probability asymptotic to 1, there will be between $a_i p_i n - n^{1/2} \log n$ and $a_i p_i n + n^{1/2} \log n$ 0 transitions out of s_i . \square

Lemma 6 *There is a constant c such that for all sufficiently large n and all $x \in L_n$,*

$$2^{-cn^{1/2} \log n - Hn} \leq \text{pr}(x) \leq 2^{cn^{1/2} \log n - Hn}. \quad (4)$$

Proof. For $x \in L_n$,

$$\text{pr}(x) = \prod_{1 \leq i \leq m} p_i^{a_i p_i n} (1 - p_i)^{a_i (1 - p_i) n} \times \prod_{1 \leq i \leq m} p_i^{c_i} (1 - p_i)^{d_i}$$

where all $|c_i|$ and $|d_i|$ are $O(n^{1/2} \log n)$. Therefore

$$\left| \sum_{1 \leq i \leq m} c_i \log p_i + d_i \log(1 - p_i) \right| \leq cn^{1/2} \log n$$

for some c , and the Lemma follows. \square

Proof of Theorem 2. We will show that

$$Hn - o(n) \leq E(K_A) \leq Hn + o(n).$$

To prove the lower bound, suppose on the contrary that $E(K_A) < an$ infinitely often for some $a < H$. Then by Markov's inequality, the set C_n of strings of complexity $\leq (H + a)n/2$ has probability $\geq (H - a)/(H + a)$ for infinitely many n . Since $\text{pr}(L_n) \sim 1$, $\text{pr}(C_n \cap L_n) \geq (H - a)/(2(H + a))$ infinitely often. Therefore by Lemma 6, $|C_n| = \Omega(2^{-cn^{1/2} \log n + Hn})$ infinitely often. But the number of strings of complexity $\leq (H + a)n/2$ is bounded by $2^{(H+a)n/2} = o(2^{-cn^{1/2} \log n + Hn})$, contradiction.

We will use the following encoding to prove the upper bound. Let $x \in \{0, 1\}^n$. The first bit of the encoding indicates whether $x \in L_n$ or not. If not, then the rest of the encoding is simply x , giving an encoding of length $n + 1$. If $x \in L_n$, say x is the u th string in L_n , where we assume some fixed effective ordering on L_n . One possible ordering is obtained by simulating S for n steps in all 2^n different ways, and enumerating only those generated strings that are in L_n . (This is where we use the assumption that the p_i 's are computable reals.) The rest of the encoding of x is the binary representation of u . By Lemma 6, $\log |L_n| \leq Hn + o(n)$, so $Hn + o(n)$ bits suffice for the encoding of x . By Lemma 5, $\text{pr}(L_n) \sim 1$, so $E(K_A) \leq Hn + o(n)$. \square

Proof of Theorem 3. Let M_n be the set of all $x \in \{0, 1\}^n$ that satisfy Equation (4). Then $L_n \subseteq M_n$. By Lemma 5, $k(n) \in M_n$. Let $h(n)$ be the least h such that $h \in M_n$ and $D_n = \{x_1, \dots, x_{h(n)-1}\}$, $E_n = \{x_{h(n)}, \dots, x_{k(n)}\}$. Clearly $E_n \subseteq M_n$. Since $D_n \cap M_n = \emptyset$, $\text{pr}(D_n) = o(1)$, and since $\text{pr}(D_n \cup E_n) > r$, $\text{pr}(D_n) = o(\text{pr}(E_n))$. Therefore $|D_n| = o(|E_n|)$ *a fortiori* because the probability of every string in D_n is at least as large as the probability of any string in E_n . Then $|E_n| \sim k(n)$, and since all strings in E_n satisfy Equation (4), the lower and upper bounds can be proven using arguments similar to those in the previous proof. \square

The encoding used in the proofs of the upper bounds in the Theorems is simple, but there is no obvious way to decode a string efficiently. Simulating S for n steps in all possible ways in order to determine that the u th string in L_n is x takes exponentially many steps. We now give an encoding that still uses only $Hn + o(n)$ bits for every string in L_n , but can be decoded in polynomial time.

Take $x \in L_n$, and for $i = 1, \dots, m$ let r_i be the number of occurrences of s_i and t_i be the number of occurrences of 0 transitions from s_i when x is generated by S . The $\binom{r_i}{t_i}$ sequences in $\{0, 1\}^{r_i}$ with t_i zeros are ordered lexicographically. Say that the sequence of 0 transitions from s_i made by S in generating x is the u_i th in the ordering. Then the entire encoding of x is the concatenation of 1 (indicating $x \in L_n$) followed by binary representations of r_i , t_i , and u_i for $i = 1, \dots, m$. To permit easy decoding, we use exactly $\lceil \log n \rceil$ bits to represent each r_i and t_i . Clearly this is sufficient. We encode each u_i with $-a_i(p_i \log p_i + (1 - p_i) \log(1 -$

$p_i))n + n^{1/2}(\log n)^2$ bits. This is sufficient because

$$\begin{aligned} \log u_i &\leq \log \binom{r_i}{t_i} \\ &\leq -a_i(p_i \log p_i + (1 - p_i) \log(1 - p_i))n + n^{1/2}(\log n)^2 \end{aligned}$$

since $x \in L_n$. Therefore the total length of the encoding is still $Hn + o(n)$.

Decoding can be done in polynomial time. The only difficulty is determining which string with t_i zeros in $\{0, 1\}^{r_i}$ is number u_i in the lexicographic ordering. Let $y = y_1 \dots y_{r_i}$ be this string and b be the least index such that $y_b = 1$. Then $b \leq t_i + 1$ and b is the largest integer such that

$$u_i \leq \sum_{j=r_i-t_i-1}^{r_i-b} \binom{j}{r_i-t_i-1}.$$

Since $r_i \leq n$, each term $\binom{j}{r_i-t_i-1}$ in the sum can be computed in polynomial time (using binary notation), and thus b can be found in polynomial time. Having found b , the process is iterated to find the next 1 bit of y , and repeated until all of y is determined.

An open problem raised by the Conjecture and Theorems is to broaden the class of algorithms for which the Theorems hold. The report [2] contains several very simple number theoretic encodings of integers that seem to satisfy the asymptotic formulas in the Theorems.

Bibliography

- [1] N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York (1963).
- [2] W. A. Beyer, M. L. Stein, and S. M. Ulam, The notion of complexity, Los Alamos Report LA-4822, U. S. Dept. of Commerce, Springfield, VA (1971).
- [3] G. J. Chaitin, On the Length of Programs for Computing Finite Binary Sequences, *J. ACM* **13** (1966), 547-569.
- [4] G. J. Chaitin, A theory of program size formally identical to information theory, *J. ACM* **22** (1975), 329-340.
- [5] W. Feller, *An Introduction to Probability Theory and its Application*, 3rd ed., Wiley, New York (1967).
- [6] A. Kolmogorov, Logical basis for information theory and probability theory, *IEEE Trans. Information Theory* **IT-14** (1968), 662-664.

- [7] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Springer-Verlag, New York-Heidelberg (1976).

This electronic publication and its contents are ©copyright 1995 by Ulam Quarterly. Permission is hereby granted to give away the journal and its contents, but no one may “own” it. Any and all financial interest is hereby assigned to the acknowledged authors of the individual texts. This notification must accompany all distribution of Ulam Quarterly.