Statistical Properties of Finite Sequences with High Kolmogorov Complexity

Ming Li*
University of Waterloo

Paul M.B. Vitányi†
CWI and Universiteit van Amsterdam

Abstract

We investigate to what extent finite binary sequences with high Kolmogorov complexity are normal (all blocks of equal length occur equally frequent), and the maximal length of all-zero or all-one runs which occur with certainty.

1 Introduction

Each individual infinite sequence generated by a $(\frac{1}{2}, \frac{1}{2})$ Bernoulli process (flipping a fair coin) has (with probability 1) the property that the relative frequency of zeros in an initial n-length segment goes to $\frac{1}{2}$ for n goes to infinity. A related statement can be made for finite sequences, in the sense that one can say that the majority of all sequences will have about one half zeros.

*Supported by the NSERC operating grant OGP-046506. Address: Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. Email: mli@math.uwaterloo.ca

†Partially supported by the NSERC International Scientific Exchange Award ISE0046203 and by NWO through NFI Project ALADDIN under Contract number NF 62-376. Address: Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

However, whereas the earlier statement is a property about individual infinite random sequences, the classical theory of probability has no machinery to define or deal with individual finite random sequences.

In [7], Kolmogorov established a notion of complexity (self-information) of finite objects which is essentially finitary and combinatorial. Says Kolmogorov [8]: "Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory must have a finite combinatorial character." It is the aim of this paper to derive randomness related statistical properties of individual (high) complexity finite binary sequences by combinatorial arguments. (A previous version of this paper appeared as part of M. Li and P. Vitányi, Combinatorics and Kolmogorov complexity, *6th IEEE Structure in Complexity Theory Conference*, 1991, pp. 154-163. There we also demonstrated the utility of a Kolmogorov complexity method in combinatorial theory by proving several combinatorial lower bounds (like the 'coin-weighing' problem), [11].)

1.1 Normal Sequences

E. Borel (1909) has called an infinite sequence of zeros and ones 'normal' in the scale of two if, for each k, the frequency of occurrences of each block y of length k in the initial segment of length n goes to limit 2^{-k} for n grows unbounded, [6]. It is known that normality is not sufficient for randomness, since Champernowne's sequence

$$123456789101112\ldots$$

is normal in the scale of ten. On the other hand, it is universally agreed that a random infinite sequence must be normal. (If not, then some blocks occur more frequent than others, which can be used to obtain better than fair odds for prediction.)

P. Martin-Löf (1965), [12], succeeded in characterizing the set of individual infinite random sequences as precisely those sequences which pass all effective tests for randomness: tests of all known and as yet unknown effectively verifyable properties of randomness alike. The criterium for randomness is that an infinite sequence must survive the 'universal Martin-Löf test'. The set of these sequences has measure 1 with respect to the uniform distribution. Martin-Löf random sequences are characterized by the fact that all initial prefixes have as much maximal Kolmogorov complexity as possible. Thus, each individual infinite sequence with this maximal Kolmogorov complexity of all initial segments has by definition all effective properties which hold on the average for sequences produced by a $(\frac{1}{2}, \frac{1}{2})$ Bernoulli process. For example, each Martin-Löf random infinite sequence is normal, it satisfies the so-called Law of the Iterated Logarithm, the number of 1's minus the number of 0's in an initial n-length segment is positive for infinitely many n and negative for another infinitely many n, and so on. This means that the statistical properties of patterns of 0's and 1's in high Kolmogorov complexity infinite sequences are well-known. With respect to normality the known situation, [12], is as follows. Consider infinite sequences over a fixed finite alphabet (equivalently, real numbers in the unit interval $[0, 1)$ represented in a fixed finite base) with respect to the uniform distribution.

Proposition 1 *The set of infinite sequences which are Martin-Löf random has uniform measure 1. Each such sequence is normal in the sense of Borel.*

In the infinite case the distinction between random sequences and nonrandom sequences can be sharply drawn, while for finite sequences randomness is necessarily a matter of degree. Namely, in the infinite case one considers limiting values of quantitive properties which hold for each individual sequence of a set of probability 1. In the finite case randomness is a matter of degree, because it would be clearly unreasonable to say that a sequence x of length n is random, and to say that a sequence y obtained by flipping one bit of x is nonrandom. What we can do is to express the degree of randomness of a finite sequence in the form of its Kolmogorov complexity, and then analyse the statistical properties of the sequence—for example, the number of 0's and 1's in it. As an example of normality properties of infinite and finite sequences expressed in terms of Kolmogorov complexity we cite the following, [12].

Example 1 If $\omega = \omega_1\omega_2\ldots$ in $\{0,1\}^\infty$ satisfies $C(\omega_{1:n}|n) \geq n - 2\log n$ for all n (this includes all Martin-Löf random sequences) then $\lim_{n\to\infty} \sum_{i=1}^{n} \omega_i/n = 1/2$. If $x = x_1x_2\ldots x_n$ in $\{0,1\}^n$ satisfies $C(x) \geq n \pm O(1)$, then $\sum_{i=1}^{n} x_i = n/2 \pm O(\sqrt{n})$

1.2 Normality of Finite Sequences

Previously, in the finite case one considers the *expected* value of quantities over a set of all sequences of a given length. We would like to obtain statements that *individual* random finite sequences have such-and-such quantitative properties in terms of their lengths. But as the result of a sequence of n fair coin flips, *any* sequence of length n can turn up. This raises the question which subset of finite sequences can be regarded as genuinely random. In [12] the viewpoint is taken that finite sequences which satisfy all *effective* tests for randomness (known and unknown alike), are as random as we will ever be able to verify. This form of randomness of individual sequences turns out to be equivalent to such sequences having maximal Kolmogorov complexity.

Since almost all finite sequences have about maximal Kolmogorov complexity, each individual maximal complexity sequence must possess approximately the expected (average) statistical properties of the overall set. For example, the existing body of knowledge tells us that each high complexity finite binary sequence is 'normal' in the sense that each binary block of length k occurs about equally frequent for k relatively small. In particular, this holds for $k = 1$. In this paper we quantify exactly the 'about' and the 'relatively small' in this statement. We quantify the extent of 'Borel normality' in relation with the Kolmogorov complexity of a finite sequence. (In the sequel we use 'complexity' in the sense of 'Kolmogorov complexity'.)

To distinguish individual random sequences obtained by flipping a physical coin from random sequences written down by human subjects, psychological tests [the correct reference is unknown to the authors] have shown that a consistent high classification score is reached by using the criterion that a real random sequence of length, say 40, on the average contains a run of zeros or ones of length 6. In contrast, human subjects feel that short random sequences should not contain such long uniform runs.

We determine the maximal length of runs of zeros or ones which are *with certainty* contained in each high complexity finite sequence. We prove that each high complexity sequence must contain a quite long run of zeros.

The properties must be related to length of the sequence. In a sequence of length 1, or odd length, the number of zeros and ones cannot be equal. To apply such properties in mathematical arguments, it is often of importance that the precise extent to which such properties hold is known.

2 Kolmogorov Complexity

To make this paper self-contained we briefly review notions and properties needed in the sequel. We identify the natural numbers $\mathcal{N}$ and the finite binary sequences as

$$(0, \epsilon), (1, 0), (2, 1), (3, 00), (4, 01), \ldots,$$

where ϵ is the empty sequence. The *length* $l(x)$ of a natural number x is the number of bits in the corresponding binary sequence. For instance, $l(\epsilon) = 0$. If A is a set, then $d(A)$ denotes the *cardinality* of A. Let $\langle \cdot \rangle : \mathcal{N} \times \mathcal{N} \to \mathcal{N}$ denote a standard computable bijective 'pairing' function. Throughout this paper, we will assume that $\langle x, y \rangle = 1^{l(x)} 0xy$.

Define $\langle x, y, z \rangle$ by $\langle x, \langle y, z \rangle \rangle$.

We need some notions from the theory of algorithms, see [13]. Let $T_1, T_2, \ldots$ be a standard enumeration of Turing machines. each of which has a binary one-way input tape, a two-way work tape, and an output tape. At the start the input tape contains a binary input delimited by endmarkers. A Turing machine T computes the function $T : \mathcal{N} \to \mathcal{N}$, defined as $T(p) = x$, with p is the input on the tape when the machine starts its computation, and x is the output on the tape when the machine halts, otherwise $T(x)$ is undefined. The input is sometimes called a *description* or *program*, and the output is called the *described object*. The description of an object x can be facilitated by an object y. The complexity of $x \in \mathcal{N}$, given y, with respect to T is defined as

$$C_T(x|y) = \min\{l(p) : T(\langle p, y \rangle) = x\}.$$

Let U be a universal Turing machine such that $U(\langle n, p \rangle) = T_n(x)$. The invariance theorem (see for example [10]) states that for each T there is a positive constant c_T such that for all $x, y \in \mathcal{N}$:

$$C_U(x|y) \leq C_T(x|y) + c_T. \tag{1}$$

Hence, for each pair of such universal machines U, U', there is a constant $c_{U,U'}$ such that for all x, y:

$$|C_U(x|y) - C_{U'}(x|y)| \leq c_{U,U'}.$$

Fixing a standard reference U, we drop the subscript and define the *Kolmogorov complexity* of x, given y, as $C_U(x|y) = C(x|y)$. This is the *minimal* number of bits in a description from which x can be effectively reconstructed, given y. The *unconditional* complexity is defined as $C(x) = C(x|\epsilon)$.

We will also make use of the *prefix* complexity $K(x)$, which denotes the shortest *self-delimiting* description. To this end, we consider so called *prefix* Turing machines, where the input tape initially contains only an infinite sequence of 0's and 1's. Thus, the input is not delimited by special endmarker symbols. Correspondingly, we define $T(p) = x$ if T started on a program beginning with p halts with output x when it has read all of p but not the next input symbol. Since the input tape is one-way, the set of programs $A = \{p : T(p) < \infty\}$ is a prefix-code: no program in A is a prefix of another program in A. A program in A is called *self-delimiting*, since T can determine where it ends without reading the next symbol of input. We define $K(x)$ and $K(x|y)$ *precisely* as $C(x)$ and $C(x|y)$ with the enumeration of Turing machines replaced by a standard enumeration of prefix Turing machines, with the universal reference machine replaced by a prefix universal reference machine, and with C replaced by K.

A survey is [10]. We need the following properties. Throughout this paper, 'log' denotes $\log_2$. For each $x, y \in \mathcal{N}$ we have

$$C(x|y) \leq l(x) + O(1).$$

For each $y \in \mathcal{N}$ there is an $x \in \mathcal{N}$ of length n such that $C(x|y) \geq n$. In particular, we can set $y = \epsilon$. Such x's may be called *random*, since they are without regularities that can be used to compress the description. Intuitively, the shortest effective description of x is x itself. In general, for each n and y, there are at least $2^n - 2^{n-c} + 1$ distinct x's of length n with

$$C(x|y) \geq n - c.$$

In some cases we want to encode x in self-delimiting form x', in order to be able to decompose $x'y$ into x and y. Good upper bounds on the prefix complexity of x are obtained by iterating the simple rule that a self-delimiting (s.d.) description of the length of x followed by x itself is a s.d. description of x. For example, $x' = 1^{l(x)}0x$ and $x'' = 1^{l(l(x))}0l(x)x$ are both s.d. descriptions for x, and this shows that $K(x) \leq 2l(x) + O(1)$ and $K(x) \leq l(x) + 2l(l(x)) + O(1)$.

Similarly, we can encode x in a self-delimiting form of a shortest program x^* ($l(x^*) = C(x)$) in $2C(x)+1$ bits. Iterating this scheme, we can encode x as a selfdelimiting program of $C(x) + 2\log C(x) + 1$ bits, which shows that $K(x) \leq C(x) + 2\log C(x) + 1$.

3 Number of Zeros and Ones

Let x have length n. It is known [12] that if $C(x|n) = n + O(1)$, then the number of zeros it contains is

$$\frac{n}{2} + O(\sqrt{n}).$$

3.1 Fixed Complexity

We analyse what complexity can say about the number of zeros and ones. Choose a large enough benchmark constant c_1 which will remain fixed for the remainder of this paper. The class of *deficiency* functions is the set of functions $\delta : \mathcal{N} \rightarrow \mathcal{N}$ satisfying $K(n, \delta(n) | n - \delta(n)) \leq c_1$ for all n (and hence $C(n | n - \delta(n)) \leq c_1$). (We have choosen c_1 so large that each monotone sublinear recursive function that we are interested in, such as $\log n$, $\sqrt{n}$, $\log\log n$ is such a deficiency function. For the special case $\delta(n) = \log\log n$ approximately the result in the Proposition below was obtained by V.G. Vovk in [14].)

Proposition 2 *There is a constant c, such that for all deficiency functions δ, for each n and $x \in \{0,1\}^n$, if $C(x) > n - \delta(n)$, then*

$$|\#ones(x) - \frac{n}{2}| < \sqrt{(\delta(n) + c)n \ln 2}. \tag{2}$$

Proof. A general estimate of the tail probability of the binomial distribution, with s_n the number of successful outcomes in n experiments with probability of success $0 < p < 1$ and $q = 1 - p$, is given by Chernoff's bounds, [3, 2]:

$$\Pr(|s_n - np| \geq m) \leq 2e^{-m^2/4npq}. \tag{3}$$

Let s_n be the number of 1's in the outcome of n fair coin flips, which means that $p = q = 1/2$. Defining $A = \{x \in \{0,1\}^n : |\#ones(x) - n/2| \geq m\}$, and applying Equation (3):

$$d(A) \leq 2^{n+1} e^{-m^2/n}.$$

Let $m = \sqrt{(\delta(n)+c)n \ln 2}$ where c is a constant to be determined later. We can compress any $x \in A$ in the following way.

1. Let s be a self-delimiting program to retrieve n and $\delta(n)$ from $n-\delta(n)$, of length at most c_1.
2. Given n and $\delta(n)$, we can effectively enumerate A. Let i be the index of x in such an effective enumeration of A. The length of the (not necessarily self-delimiting) description of i satisfies:

$$\begin{aligned} l(i) \leq \log d(A) &= n + 1 + \log e^{-m^2/n} \\ &\leq n + 1 - \delta(n) - c. \end{aligned}$$

The string si is padded to length $n + 1 - \delta(n) - c + c_1$. From si we can reconstruct x by first using $l(si)$ to compute $n - \delta(n)$, then compute n and $\delta(n)$ from s and $n - \delta(n)$, and subsequently enumerate A to obtain the ith element. Let T be the Turing machine embodying the procedure for reconstructing x. Then, by Equation (1),

$$C(x) \leq C_T(x) + c_T \leq n + 1 - \delta(n) - c + c_1 + c_T.$$

Choosing $c = 1 + c_1 + c_T$ we find $C(x) \leq n - \delta(n)$, which contradicts the condition of the theorem. Hence, $|\#ones(x) - \frac{n}{2}| < m$. $\square$

3.2 Fixed Number of Ones

It may be surprising at first glance, but there are no maximally complex sequences with about equal number of zeros and ones. Equal numbers of zeros and ones is a form of regularity, and therefore lack of complexity. That is, for $x \in \{0,1\}^n$, if $|\#ones(x) - n/2| = c$ for some constant c, then the randomness deficiency $\delta(n) = n - C(x)$ is non-constant (order $\log n$). We prove this fact in the following proposition.

Proposition 3 *There is a constant c, such that for all n and all $x \in \{0,1\}^n$, if*

$$|\#ones(x) - \frac{n}{2}| \leq 2^{-\delta(n)-c}\sqrt{n}$$

then $C(x) \leq n - \delta(n)$.

Proof. Let $m = 2^{-\delta(n)-c}\sqrt{n}$, with c a constant to be determined later. Let $A = \{x \in \{0,1\}^n : |\#ones(x) - \frac{n}{2}| \leq m\}$. There is a constant c_2 such that are only

$$d(A) \leq (2m+1)\binom{n}{n/2} \leq c_2 \frac{2^n m}{\sqrt{n}} \tag{4}$$

elements in A (use Sterling's approximation). Thus, for each $x \in A$, we can encode x by its index in an enumeration of A. We can find A from n and $\delta(n)$. We can find n and $\delta(n)$ from $n - \delta(n)$ by a self-delimiting program of size at most c_1. Altogether, this description takes $\log d(A) + c_1 = n - \delta(n) - c + c_1 + \log c_2$ bits. Let this process of reconstructing x be executed by Turing machine T. Choosing $c = c_1 + \log c_2 + c_T$ we obtain by Equation (1):

$$C(x) \leq C_T(x) + c_T \leq n - \delta(n).$$

□

Example 2. As examples, we consider some particular values of $\delta(n)$. Set $\delta(n) = (1/2)\log n - \log\log n$. If $|\#ones(x) - n/2| = O(\log n)$, then $C(x) \leq n - (1/2)\log n + \log\log n + O(1)$. Set $\delta(n) = (1/2)\log n$. If $|\#ones(x) - n/2| = O(1)$, then $C(x) \leq n - (1/2)\log n + O(1)$. That is, if the number of 1's is too close to the number of 0's then the complexity of the string drops significantly below its maximum.

A random string of length n cannot have precisely or almost $n/2$ ones by Proposition 3. Then how many ones should a random string contain? The next proposition shows that for a random x having $j + n/2$ ones, $K(j)$ must be at least about $O(\log n)$.

Proposition 4 *There is a constant c, such that for all n and all $x \in \{0,1\}^n$, if*

$$|\#ones(x) - \frac{n}{2}| = j,$$

then $C(x|n) \leq n - (1/2)\log n + K(j|n) + c$.

Proof. Let $A = \{x \in \{0,1\}^n : |\#ones(x) - \frac{n}{2}| = j\}$. There is a constant c_3 such that there are

$$d(A) \leq \binom{n}{n/2} \leq c_3 \frac{2^n}{\sqrt{n}} \tag{5}$$

elements in A (use Sterling's approximation). In order to enumerate elements in A, we only need to describe j and n. Thus, for any $x \in A$, we can encode x by its index i (in $\log d(A)$ bits) in an enumeration of A. With n given, we can recover x from an encoding of j in $K(j|n)$ bits, followed by i. This description takes of x, given n, takes $\log d(A) + K(j|n) \leq n - (1/2)\log n + \log c_3 + K(j|n)$ bits. Let T be the Turing machine embodying this procedure to recover x given n. Choosing $c = \log c_3 + c_T$, we have:

$$C(x|n) \leq C_T(x|n) + c_T \leq n - (1/2)\log n + K(j|n) + c.$$

□

Example 3. For $j = O(1)$ we have $C(x|n) \leq n - 1/2 \log n + O(1)$ which is slightly stronger than the statement about unconditional $C(x)$ in Example 2. For $j = \sqrt{n}$ and j random ($K(j|n) \geq (1/2)\log n$), we have $C(x|n) \leq n - O(1)$. Only for such j's is it possible that a number x is random.

4 Number of Blocks

An infinite binary sequence is called *normal* if each block of length k occurs with limiting frequency of 2^{-k}. This justifies our intuition, that a random infinite binary sequence contains about as many zeros as ones. But also, blocks 00, 01, 10, and 11 should appear about equally often. In general we expect that each block of length k occurs with about the same frequency. Can we find an analogue for finite binary sequences? We analyse these properties for high complexity finite binary sequences to obtain a quantification of a similar statement in terms of the length of the sequence and its complexity.

4.1 Fixed Complexity

Let $x = x_1 \dots x_n$ be a binary sequence of length n, and y a much smaller string of length l. Let $p = 2^{-l}$ and $\#y(x)$ be the number of (possibly overlapping) distinct occurrences of y in x. For convenience, we assume that x 'wraps around' so that an occurrence of y starting at the end of x and continuing at the start also counts.

Theorem 1 *Let $l = l(y)$, $p = 2^l$. There is a constant c, such that for all n and $x \in \{0,1\}^n$, if $C(x) > n - \delta(n)$, then*

$$|\#y(x) - np| < \sqrt{\alpha np},$$

with $\alpha = [K(y|n) + \log l + \delta(n) + c](1-p)l4\ln 2$.

Proof. We prove by contradiction. Assume that n is divisible by l. (If it is not, then we can put x on a Procrustus bed to make its length divisible by l at the cost of having the above frequency estimate $\#y(x)$ plus or minus an error term of at most $l/2$.) There are l ways of dividing (the ring) x into $N = n/l$ contiguous equal sized blocks, each of length l. For each such division $i \in \{0,1,\ldots,l-1\}$, let $\#y(x,i)$ be the number of (now nonoverlapping) occurrences of block y. We apply the Chernoff bound, Equation (3) again: With $A = \{x \in \{0,1\}^n : |\#y(x,i) - Np| \geq m\}$ this gives $d(A) \leq 2^{n+1}e^{-m^2/4Np(1-p)}$. We choose m, such that for some constant c to be determined later:

$$\frac{m^2 \log e}{4Np(1-p)} = K(\langle y,i\rangle|n) + \delta(n) + c.$$

To describe an element x in A, we now need only to enumerate A and indicate the index of x in such an enumeration. The description contains the following items:

1. *A description used to enumerate A.* Given $n - \delta(n)$, we can retrieve n and $\delta(n)$, using a self-delimiting description of at most c_1 bits, To enumerate A, we also need to know i and y. Therefore, given $n - \delta(n)$, the required number of bits to enumerate A is at most:

$$K(\langle y,i,\delta(n),n\rangle|n-\delta(n)) \leq K(\langle y,i\rangle|n) + c_1.$$

2. *A description of the index of x.* The number of bits to code the index j of x in A is:

$$\log d(A) \quad \leq \quad \log\left(2^{n+1}e^{-m^2/4Np(1-p)}\right)$$

$$
\begin{aligned}
&= n + 1 - \frac{m^2 \log e}{4Np(1-p)} \\
&= n + 1 - K(\langle y, i\rangle | n) - \delta(n) - c.
\end{aligned}
$$

This total description takes at most $n+1-\delta(n)-c+c_1$ bits. Let T be a Turing machine reconstructing x from these items. According to Equation (1), therefore:

$$C(x) \leq C_T(x) + c_T \leq n + 1 - \delta(n) - c + c_1 + c_T.$$

with $c = 1 + c_1 + c_T$ we have $C(x) \leq n - \delta(n)$, which contradicts the assumption of the theorem.

Therefore, $|\#y(x,i) - Np| < m$, which in its turn implies:

$$|\#y(x,i) - Np| < \sqrt{\frac{K(\langle y, i\rangle | n) + \delta(n) + c}{\log e} 4Np(1-p)}.$$

The theorem now follows by noting that $|\#y(x) - np| = \sum_{i=0}^{l-1} |\#y(x,i) - Np|$, and $K(i|l) \leq \log l$. $\square$

4.2 Fixed Number of Blocks

Similar to the analysis of blocks of length 1, the complexity of a string drops below its maximum in case some block y of length l occurs in one of the l block divisions, say i, with frequency exactly pN ($p = 1/2^l$). Then we can point out x by giving n, y, i and its index in a set of cardinality

$$\binom{N}{pN} (2^l - 1)^{N-pN} = O\left(\frac{2^{Nl}}{\sqrt{Np(1-p)}}\right).$$

Therefore,

$$C(x | \langle n, y\rangle) \leq n - \frac{1}{2}\log n + \frac{1}{2}(l + 3\log l) + O(1).$$

5 Length of Runs

It is known from probability theory, that in a randomly generated finite sequence the *expectancy* of the length of the longest run of zeros or ones is pretty high. For each individual finite sequence with high Kolmogorov complexity we are *certain* that it contains each block (say, a run of zeros) up to a certain length.

Theorem 2 *Let x of length n satisfy $C(x) \geq n - \delta(n)$. Then, for sufficiently large n, x contains all blocks y of length*

$$l = \log n - \log\log n - \log(\delta(n) + \log n) - O(1).$$

Proof. We are sure that y occurs at least once in x, if $\sqrt{\alpha np}$ in Theorem 1 is at most np. This is the case if $\alpha \leq np$, that is:

$$\frac{K(y|n) + \log l + \delta(n) + O(1)}{\log e} 4l \leq np.$$

Substitute $K(y|n) \leq l + 2\log l + O(1)$ (since $K(y|n) \leq K(y)$), and $p = 2^{-l}$ with l set at

$$l = \log n - \log(3\delta(n)\log n + 3\log^2 n),$$

(which equals l in the statement of the theorem up to an additive constant). The result is

$$\frac{l + 3\log l + \delta(n) + O(1)}{\log e} 4l \leq 3(\delta(n)\log n + \log^2 n),$$

and it is easy to see that this holds for sufficiently large n. □

Corollary 1 *If $\delta(n) = O(\log n)$ then each block of length $\log n - 2\log\log n - O(1)$ is contained in x.*

Analysing the proof of Theorem 2 we can improve this in case $K(y|n)$ is low.

Corollary 2 *If $\delta(n) = O(\log\log n)$, then for each $\epsilon > 0$ and n large enough, x contains an all-zero run y (for which $K(y|n) = O(\log l)$) of length $l = \log n - (1+\epsilon)\log\log n + O(1)$.*

Remark. Since there are $2^n(1 - O(1/\log n))$ strings x of length n with $C(x) \geq n - \log\log n + O(1)$, the *expected length of the longest run of consecutive zeros* if we flip a fair coin n times, is at least l as in Corollary 2. (This improves a lower bound of $\log n - 2\log\log n$, obtained as example of elementary methods in [2], by a $\log\log n$ additive term.)

We show in what sense Theorem 2 is sharp. Let $x = uvw$, $l(x) = n$ and $C(x) \geq n - \delta(n)$. We can describe x by giving

1. A description of v in $K(v)$ bits.

2. The literal representation of uw.

3. A description of $l(u)$ in $\log n + \log\log n + 2\log\log\log n + O(1)$

Then, since we can find n by $n = l(v) + l(uw)$,

$$\begin{aligned} C(x) &\leq n - l(v) + K(v) + \log n \\ &+ (1 + o(1))\log\log n + O(1). \end{aligned} \tag{6}$$

Substitute $C(x) = n - \delta(n)$ and $K(v) = o(\log\log n)$ (choose v to be very regular) in Equation (6) to obtain:

$$l(v) \leq \delta(n) + \log n + (1 + o(1))\log\log n.$$

This means that, for instance, for each $\epsilon > 0$, no maximally complex string x with $C(x) = n + O(1)$ contains a run of zeros (or the initial binary digits of π) of length $\log n + (1+\epsilon)\log\log n$ for n large enough and regular enough. By Corollary 2, on the other hand, such a string x *must* contain a run of zeros of length $\log n - (1 + \epsilon)\log\log n + O(1)$.

We end this paper by an interesting question raised by one of the referees: if x is known to be Kolmogorov random with respect to some superpolynomial (say $n^{\log n}$) time-bounded Turing machines, what can we say about their statistical properties?

References

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

[2] T. Corman, C. Leiserson, R. Rivest, *Introduction to Algorithms*, McGraw Hill, 1990.

[3] P. Erdös and J. Spencer, *Probabilistic Methods in Combinatorics*, Academic Press, New York, 1974.

[4] P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, **15**(1974), 1477-1480.

[5] R.L. Graham, D.E. Knuth, O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.

[6] D.E. Knuth, *Seminumerical Algorithms*, Addison-Wesley, 1981.

[7] A.N. Kolmogorov, Three approaches to the definition of the concept 'quantity of information', *Problems in Information Transmission*, **1:1**(1965), 1-7.

[8] A.N. Kolmogorov, Combinatorial foundation of information theory and the calculus of probabilities, *Russian Math. Surveys*, **38:4**(1983), 29-40.

[9] L.A. Levin, Laws of Information conservation (non-growth) and aspects of the foundation of probability theory, *Problems in Information Transmission*, **10**(1974), 206-210.

[10] M. Li and P.M.B. Vitányi, Kolmogorov complexity and its applications, pp. 187-254 in: *Handbook of Theoretical Computer Science, Vol. A*, J. van Leeuwen, Ed., Elsevier/MIT Press, 1990.

[11] M. Li and P.M.B. Vitányi, Kolmogorov complexity arguments in combinatorics, *J. Combinatorial Theory, Ser. A*, To appear.

[12] P. Martin-Löf, On the definition of random sequences, *Information and Control*, (1966).

[13] H.J. Rogers, Jr., *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, 1967.

[14] V.G. Vovk, *SIAM Theory Probab. Appl.*, 32(1987), 413-425.