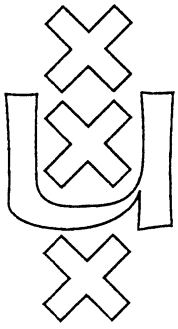
Institute for Language, Logic and Information

COMBINATORIAL PROPERTIES OF FINITE SEQUENCES WITH HIGH KOLMOGOROV COMPLEXITY

Ming Li
Paul M.B. Vitányi

ITLI Prepublication Series
for Computation and Complexity Theory CT-91-09



University of Amsterdam

The ITLI Prepublication Series

1986

86-01 The Institute of Language, Logic and Information
86-02 Peter van Emde Boas — A Semantical Model for Integration and Modularization of Rules
86-03 Johan van Benthem — Categorial Grammar and Lambda Calculus
86-04 Reinhard Muskens — A Relational Formulation of the Theory of Types
86-05 Kenneth A. Bowen, Dick de Jongh — Some Complete Logics for Branched Time, Part I Well-founded Time,
86-06 Johan van Benthem — Logical Syntax Forward looking Operators

1987

87-01 Jeroen Groenendijk, Martin Stokhof — Type shifting Rules and the Semantics of Interrogatives
87-02 Renate Bartsch — Frame Representations and Discourse Representations
87-03 Jan Willem Klop, Roel de Vrijer — Unique Normal Forms for Lambda Calculus with Surjective Pairing
87-04 Johan van Benthem — Polyadic quantifiers
87-05 Víctor Sánchez Valencia — Traditional Logicians and de Morgan's Example
87-06 Eleonore Oversteegen — Temporal Adverbials in the Two Track Theory of Time
87-07 Johan van Benthem — Categorial Grammar and Type Theory
87-08 Renate Bartsch — The Construction of Properties under Perspectives
87-09 Herman Hendriks — Type Change in Semantics: The Scope of Quantification and Coordination

1988

Logic, Semantics and Philosophy of Language:
LP-88-01 Michiel van Lambalgen — Algorithmic Information Theory
LP-88-02 Yde Venema — Expressiveness and Completeness of an Interval Tense Logic
LP-88-03 — Year Report 1987
LP-88-04 Reinhard Muskens — Going partial in Montague Grammar
LP-88-05 Johan van Benthem — Logical Constants across Varying Types
LP-88-06 Johan van Benthem — Semantic Parallels in Natural Language and Computation
LP-88-07 Renate Bartsch — Tenses, Aspects, and their Scopes in Discourse
LP-88-08 Jeroen Groenendijk, Martin Stokhof — Context and Information in Dynamic Semantics
LP-88-09 Theo M.V. Janssen — A mathematical model for the CAT framework of Eurotra
LP-88-10 Anneke Kleppe — A Blissymbolics Translation Program

Mathematical Logic and Foundations:
ML-88-01 Jaap van Oosten — Lifschitz' Realizability
ML-88-02 M.D.G. Swaen — The Arithmetical Fragment of Martin Löf's Type Theories with weak Σ-elimination
ML-88-03 Dick de Jongh, Frank Veltman — Provability Logics for Relative Interpretability
ML-88-04 A.S. Troelstra — On the Early History of Intuitionistic Logic
ML-88-05 A.S. Troelstra — Remarks on Intuitionism and the Philosophy of Mathematics

Computation and Complexity Theory:
CT-88-01 Ming Li, Paul M.B.Vitanyi — Two Decades of Applied Kolmogorov Complexity
CT-88-02 Michiel H.M. Smid — General Lower Bounds for the Partitioning of Range Trees
CT-88-03 Michiel H.M. Smid, Mark H. Overmars, Leen Torenvliet, Peter van Emde Boas — Maintaining Multiple Representations of Dynamic Data Structures
CT-88-04 Dick de Jongh, Lex Hendriks, Gerard R. Renardel de Lavalette — Computations in Fragments of Intuitionistic Propositional Logic
CT-88-05 Peter van Emde Boas — Machine Models and Simulations (revised version)
CT-88-06 Michiel H.M. Smid — A Data Structure for the Union-find Problem having good Single-Operation Complexity
CT-88-07 Johan van Benthem — Time, Logic and Computation
CT-88-08 Michiel H.M. Smid, Mark H. Overmars, Leen Torenvliet, Peter van Emde Boas — Multiple Representations of Dynamic Data Structures
CT-88-09 Theo M.V. Janssen — Towards a Universal Parsing Algorithm for Functional Grammar
CT-88-10 Edith Spaan, Leen Torenvliet, Peter van Emde Boas — Nondeterminism, Fairness and a Fundamental Analogy
CT-88-11 Sieger van Denneheuvel, Peter van Emde Boas — Towards implementing RL

Other prepublications:
X-88-01 Marc Jumelet — On Solovay's Completeness Theorem

1989

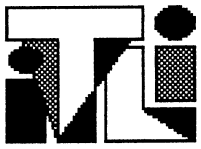
Logic, Semantics and Philosophy of Language:
LP-89-01 Johan van Benthem — The Fine-Structure of Categorial Semantics
LP-89-02 Jeroen Groenendijk, Martin Stokhof — Dynamic Predicate Logic, towards a compositional, non-representational semantics of discourse
LP-89-03 Yde Venema — Two-dimensional Modal Logics for Relation Algebras and Temporal Logic of Intervals
LP-89-04 Johan van Benthem — Language in Action
LP-89-05 Johan van Benthem — Modal Logic as a Theory of Information
LP-89-06 Andreja Prijatelj — Intensional Lambek Calculi: Theory and Application
LP-89-07 Heinrich Wansing — The Adequacy Problem for Sequential Propositional Logic
LP-89-08 Víctor Sánchez Valencia — Peirce's Propositional Logic: From Algebra to Graphs
LP-89-09 Zhisheng Huang — Dependency of Belief in Distributed Systems

Mathematical Logic and Foundations:
ML-89-01 Dick de Jongh, Albert Visser — Explicit Fixed Points for Interpretability Logic
ML-89-02 Roel de Vrijer — Extending the Lambda Calculus with Surjective Pairing is conservative
ML-89-03 Dick de Jongh, Franco Montagna — Rosser Orderings and Free Variables
ML-89-04 Dick de Jongh, Marc Jumelet, Franco Montagna — On the Proof of Solovay's Theorem
ML-89-05 Rineke Verbrugge — Σ-completeness and Bounded Arithmetic
ML-89-06 Michiel van Lambalgen — The Axiomatization of Randomness
ML-89-07 Dirk Roorda — Elementary Inductive Definitions in HA: from Strictly Positive towards Monotone
ML-89-08 Dirk Roorda — Investigations into Classical Linear Logic
ML-89-09 Alessandra Carbone — Provable Fixed points in $I\Delta_0+\Omega_1$

Computation and Complexity Theory:
CT-89-01 Michiel H.M. Smid — Dynamic Deferred Data Structures
CT-89-02 Peter van Emde Boas — Machine Models and Simulations
CT-89-03 Ming Li, Herman Neuféglise, Leen Torenvliet, Peter van Emde Boas — On Space Efficient Simulations
CT-89-04 Harry Buhrman, Leen Torenvliet — A Comparison of Reductions on Nondeterministic Space
CT-89-05 Pieter H. Hartel, Michiel H.M. Smid, Leen Torenvliet, Willem G. Vree — A Parallel Functional Implementation of Range Queries
CT-89-06 H.W. Lenstra, Jr. — Finding Isomorphisms between Finite Fields
CT-89-07 Ming Li, Paul M.B. Vitanyi — A Theory of Learning Simple Concepts under Simple Distributions and Average Case Complexity for the Universal Distribution (Prel. Version)
CT-89-08 Harry Buhrman, Steven Homer, Leen Torenvliet — Honest Reductions, Completeness and Nondeterminstic Complexity Classes
CT-89-09 Harry Buhrman, Edith Spaan, Leen Torenvliet — On Adaptive Resource Bounded Computations
CT-89-10 Sieger van Denneheuvel — The Rule Language RL/1
CT-89-11 Zhisheng Huang, Sieger van Denneheuvel, Peter van Emde Boas — Towards Functional Classification of Recursive Query Processing

Other Prepublications:
X-89-01 Marianne Kalsbeek — An Orey Sentence for Predicative Arithmetic
X-89-02 G. Wagemakers — New Foundations: a Survey of Quine's Set Theory
X-89-03 A.S. Troelstra — Index of the Heyting Nachlass
X-89-04 Jeroen Groenendijk, Martin Stokhof — Dynamic Montague Grammar, a first sketch
X-89-05 Maarten de Rijke — The Modal Theory of Inequality
X-89-06 Peter van Emde Boas — Een Relationele Semantiek voor Conceptueel Modelleren: Het RL-project

1990 SEE INSIDE BACK COVER



Instituut voor Taal, Logica en Informatie
Institute for Language, Logic and Information

Faculteit der Wiskunde en Informatica
(Department of Mathematics and Computer Science)
Plantage Muidergracht 24
1018TV Amsterdam

Faculteit der Wijsbegeerte
(Department of Philosophy)
Nieuwe Doelenstraat 15
1012CP Amsterdam

COMBINATORIAL PROPERTIES OF FINITE SEQUENCES WITH HIGH KOLMOGOROV COMPLEXITY

Ming Li
Computer Science Department
University of Waterloo
Paul M.B. Vitányi
Department of Mathematics and Computer Science
University of Amsterdam
& CWI

ITLI Prepublication Series
for Computation and Complexity Theory
ISSN 0924-8374

Received June 1991

Combinatorial Properties of Finite Sequences with High Kolmogorov Complexity

Ming Li*
University of Waterloo

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

May 29, 1991

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

Infinite sequences generated by a $(\frac{1}{2}, \frac{1}{2})$ Bernoulli process (flipping a fair coin) have 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

*Supported by the NSERC operating grants OGP-0036747 and OGP-046506. Address: Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. Email: mli@math.waterloo.edu

†Partially supported by the NSERC International Scientific Exchange Award ISE0046203. Address: Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

for finite sequences, in the sense that one can say that the majority of all sequences will have about one half zeros. 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 combinatorial properties of high complexity finite binary sequences by combinatorial arguments, without any probabilistic assumptions at all.

It turns out to be quite natural to do combinatorial proofs by Kolmogorov complexity arguments, which are of themselves combinatorial in nature. We have demonstrated the utility of a Kolmogorov complexity method in combinatorial theory by proving several combinatorial lower bounds (like the 'coin-weighing' problem), [11]. Rather than doing combinatorics using Kolmogorov complexity, in this paper we are interested in combinatorial properties of individual finite binary sequences with high Kolmogorov complexity.

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.)

While 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 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 length.

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. In the sequel we use 'complexity' in the sense of 'Kolmogorov complexity'.

We prove 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 $k = 1$. We quantify the 'about' and the 'relatively small' in this statement.

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, 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 such sequence must contain a relatively 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 $|A|$ denotes the *cardinality* of A. Let $< . >: \mathcal{N} \times \mathcal{N} \to \mathcal{N}$ denote a standard computable bijective 'pairing' function. Throughout this paper, we will assume that $< x, y >= 1^{l(x)}0xy$.

Define $< x, y, z >$ by $< x, < y, z >>$.

We need some notions from the theory of algorithms, see [13]. Let $\phi_1, \phi_2, \ldots$ be a standard enumeration of the partial recursive functions. The (Kolmogorov) *complexity* of $x \in \mathcal{N}$, given y, is defined as

$$C(x|y) = \min\{l(< n, z >) : \phi_n(< y, z >) = x\}.$$

This means that $C(x|y)$ is the *minimal* number of bits in a description from which x can be effectively reconstructed, given y. The unconditional complexity is defined as $C(x) = C(x|\epsilon)$.

An alternative definition is as follows. Let

$$C_\psi(x|y) = \min\{l(z) : \psi(< y, z >) = x\} \tag{1}$$

be the conditional complexity of x given y with reference to decoding function ψ. Then $C(x|y) = C_\psi(x|y)$ for a universal partial recursive function ψ that satisfies $\psi(< y, n, z >) = \phi_n(< y, z >)$.

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, which have only 0's and 1's on their input tape, and thus cannot detect the end of the input. Instead we define an input as that part of the input tape which the machine has read when it halts. When $x \neq y$ are two such input, we clearly have that x cannot be a prefix of y, and hence the set of inputs forms what is called a *prefix code*. We define $K(x)$ similarly as above, with reference to a universal prefix machine that first reads 1^n0 from the input tape and then simulates prefix machine n on the rest of the input.

A survey is [10]. We need the following properties. Throughout 'log' denotes the binary logarithm. We often use $O(f(n)) = -O(f(n))$, so that $O(f(n))$ may denote a negative quantity. For each $x, y \in \mathcal{N}$ we have

$$C(x|y) \leq l(x) + O(1). \tag{2}$$

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

$$C(x|y) \geq n - c. \tag{3}$$

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 its shortest program $p(x)$ ($l(p(x)) = C(x)$) 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 $K(x) \leq C(x) + 2\log C(x) + 1$, and so on.

3 Expectation versus Complexity

To derive our results, we often use a common pattern of argument. Following a suggestion of John Tromp, we can formulate it in the form of a general 'Tail Law'.

Consider the sample space $S = \{0,1\}^*$ with uniform probability $\Pr(x) = 2^{-2l(x)}$. Put $S^n = \{0,1\}^n$. Then, $\Pr(x|x \in S^n) = 2^{-n}$. Let $R : S \to \mathcal{Z}$, total recursive, be a function that (in our case) measures the deviation between some function g of x and a reference value $r(l(x))$ for all strings of the same length. We are interested in the relation between the complexity of a string x and this deviation. A natural choice of r would be the average $g(x)$ over S^n. Fix a class D of *deficiency* functions $\delta : \mathcal{N} \to \mathcal{N}$ for which $K(n|n-\delta(n)) = O(1)$. This is satisfied by every monotone sublinear recursive function that we are interested in. The complexity of R can be identified with the complexity of its index in the effective enumeration of recursive functions, which we can assume equals some constant plus (optionally) the complexity of its parameters.

Define the tail probability

$$p(R; n, m) = \Pr\{x \in S^n : |R(x)| \geq m\}.$$

Lemma 1 (Tail Lemma) *Let f be a function from $\mathcal{N} \times \mathcal{N}$ to $\mathcal{N}$ satisfying*

$$-\log p(R; n, f(n,k)) \geq K(R|n) + k + O(1).$$

Then for any $\delta \in D$, we have that all x with $C(x) > n - \delta(n)$ $(n = l(x))$, satisfy

$$|R(x)| < f(n, \delta(n)).$$

Proof. By contradiction. Assume that for some $\delta \in D$, there exists an n such that $A = \{x \in S^n : |R(x)| \geq f(n, \delta(n))\}$ is non-empty. We can describe such an $x \in A$ in the following way:

1. let s be a s.d. program for n given $n - \delta(n)$, of length $l(s) = K(n|n - \delta(n)) = O(1)$.
2. let q be a s.d. program for R given n, of length $l(q) = K(R|n)$.
3. let i be the index of x in an effective enumeration of A, from the x's with the highest $|R(x)|$'s down. From $|A| = 2^n \Pr(A) = 2^n p(R; n, f(n, \delta(n)))$ it follows that the length of the (not necessarily s.d.) description of i satisfies:

$$\begin{aligned} l(i) \leq \log |A| &= n + \log p(R; n, f(n, \delta(n))) \\ &\leq n - K(R|n) - \delta(n) - O(1). \end{aligned}$$

The string sqi has length at most $n - \delta(n) - O(1)$ and can be padded to a string z of length exactly $n - \delta(n) - c$, where c is a constant determined below. From z we can reconstruct x by first using its length plus c to compute n (and $\delta(n)$) from s, then use n to obtain R from q, and finally enumerate A to obtain the ith element. Note that we can enumerate A up to the ith element without using f at all, since we enumerate from the x's with the highest $|R(x)|$ down. So, if recursive function ψ embodies above procedure for reconstructing x, we have, by Equation 1,

$$C(x) \leq C_\psi(x) + c_\psi \leq n - \delta(n) - c + c_\psi.$$

Choosing $c = c_\psi$ finishes the proof. □

Corollary 1 (Tail Lemma Dual) *The exact same argument shows that for sufficiently random x, the deviation $|R(x)|$ is not too small. We thus obtain a Tail Lemma Dual starting from $q(R;n,m) = \Pr\{x \in S^n : |R(x)| \leq m\}$.*

4 Number of Zeros and Ones

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

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

4.1 Fixed Complexity

We analyse what complexity can say about the number of zeros and ones. Let $x = x_1 x_2 \ldots x_n$ and $\delta \in D$ a deficiency function. Suppose,

$$C(x) \geq n - \delta(n).$$

Let $R(x) = \sum x_i - \frac{n}{2}$ be the deviation in the number of ones in x. With $x \in \{0,1\}^n$ uniformly distributed, $\#ones(x) = \sum x_i$ is distributed according to the binomial distribution.

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],

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

The tail probability $p(R;n,m)$ bounded by Equation 4 (with $R(x) = s_n - \frac{n}{2}$ and $p = q = 1/2$) yields:

$$-\log p(R;n,m) \geq \frac{m^2 \log e}{n} - 1$$

Clearly, R is a recursive function with $K(R|n) = O(1)$. Thus, choosing $f(n,k) = \sqrt{(k + O(1))n \ln 2}$, Lemma 1 gives us: all x with $C(x) > n - \delta(n)$ $(n = l(x))$, satisfy

$$|\#ones(x) - \frac{n}{2}| < \sqrt{(\delta(n) + O(1))n \ln 2}. \tag{5}$$

If the complexity of x satisfies that the conditional complexity $C(x|n) \geq n - \delta(n)$, clearly Equation 5 holds a fortiori.

4.2 Fixed Number of Zeros

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. Using the same notation as before, if $R(x) = O(1)$ then the randomness deficiency $\delta(n) = n - C(x)$ is relatively large. For instance,

$$\begin{aligned} q(R;n,m) &= \Pr\{x \in S^n : |R(x)| \leq m\} \\ &\leq (2m+1)2^{-n}\left(\begin{array}{c} n \\ n/2 \end{array}\right) = O(\frac{m}{\sqrt{n}}). \end{aligned}$$

Thus, setting $f(n,k) = 2^{-k-O(1)}\sqrt{n}$, the Tail Law Dual (Corollary 1) gives us: all x with $C(x) > n - \delta(n)$ $(n = l(x))$, satisfy

$$|\#ones(x) - \frac{n}{2}| > 2^{-\delta(n)-O(1)}\sqrt{n}.$$

Perhaps more interestingly, we can define

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

so that $K(R'|n)$ is about $K(j)$. Applying the Tail Law Dual with

$$f(n,k) = 2^{-k-K(j)-O(1)}\sqrt{n},$$

we then find that all x with $C(x) > n - \delta(n)$ satisfy

$$|\#ones(x) - (\frac{n}{2} + j)| > 2^{-\delta(n)-K(j)-O(1)}\sqrt{n}.$$

This means that for a random x having exactly $j + n/2$ ones, $K(j)$ must be at least about $\log n$.

5 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.

5.1 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. Put $R_y(x) = \#y(x) - np$. (So $R_1(x) = \sum x_i - n/2$.) 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 *All x with $C(x) > n - \delta(n)$ satisfy*

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

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

Proof. We prove by contradiction. Assume that n is divisible by l. (If it is not we can put x on a Procrustus bed to make its length divisible by l at the cost of having the above frequency estimate up to a $l/2$ additive term.) 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 $i \in \{0, 1, \dots, l-1\}$, let $R_{y,i}(x)$ be the number of (now nonoverlapping) occurrences of block y minus Np. Notice that $R_{y,i}(x)$ is the deviation from the expectation of a Bernoulli sequence of length N with probability of succes (a block matching y) p, for which we can use the Chernoff bound 4.

$$p(R_{y,i}; n, m) \leq 2e^{-m^2/4Np(1-p)}.$$

Taking the negative logarithm on both sides:

$$-\log p(R_{y,i}; n, m) \geq \frac{m^2 \log e}{4Np(1-p)} - 1. \tag{6}$$

Choose $m = f(n,k)$, such that

$$\frac{f(n,k)^2 \log e}{4Np(1-p)} = K(R_{y,i}|n) + k + O(1). \tag{7}$$

Equations 6, 7 enable us to apply the Tail Lemma 1. Application of the Tail Lemma yields that all x with $C(x) > n - \delta(n)$ satisfy $|R_{y,i}(x)| < f(n, \delta(n))$. Substitution of f according to Equation 7, with $K(R_{y,i}|n) = K(y,i|n)+O(1)$, gives:

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

The theorem now follows by noting that $R_y(x) = \sum_{i=0}^{l-1} R_{y,i}(x)$, and $K(i|l) \leq \log l$ □

With $C(x|n, R_y) \geq n - \delta(n)$, Theorem 1 holds a fortiori.

5.2 Fixed Number of Blocks

Similar to the analysis of blocks of length 1, the complexity 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

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

Therefore,

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

6 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 up to a certain length (like a run of zeros).

Theorem 2 *Let x of length n satisfy $C(x) \geq n - \delta(n)$. Then 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 2 *If $\delta(n) = O(\log n)$ then each block of length $\log n - 2\log\log n - O(1)$ is contained in x.*

Corollary 3 *Analysing the proof of Theorem 2 we can improve this in case $K(y|n)$ is low. 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 $l = \log n - (1+\epsilon)\log\log n + O(1)$.*

Corollary 4 (improving [2]) *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 3. This improves the lower bound of* $\log n - 2\log\log n$ *cited 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 $n = l(v) + l(uw)$,

$$\begin{aligned} C(x) &\leq n - l(v) + K(v) + \log n \\ &+ (1 + o(1))\log\log n + O(1). \end{aligned} \tag{8}$$

Substitute $C(x) = n - \delta(n)$ and $K(v) = o(\log\log n)$ (choose v to be very regular) in Equation 8 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 3, on the other hand, such a string x *must* contain a run of zeros of length $\log n - (1 + \epsilon)\log\log n + O(1)$.

References

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

[2] Corman, C. Leiserson, R. Rivest, *Introduction to Algorithms*, 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, Manuscript, 1990, submitted.

[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] C.E. Shannon, A mathematical theory of communication, *Bell System Tech. J.*, **27**(1948), 379-423, 623-656.

The ITLI Prepublication Series

1990 *Logic, Semantics and Philosophy of Language*

LP-90-01 Jaap van der Does — A Generalized Quantifier Logic for Naked Infinitives
LP-90-02 Jeroen Groenendijk, Martin Stokhof — Dynamic Montague Grammar
LP-90-03 Renate Bartsch — Concept Formation and Concept Composition
LP-90-04 Aarne Ranta — Intuitionistic Categorial Grammar
LP-90-05 Patrick Blackburn — Nominal Tense Logic
LP-90-06 Gennaro Chierchia — The Variablity of Impersonal Subjects
LP-90-07 Gennaro Chierchia — Anaphora and Dynamic Logic
LP-90-08 Herman Hendriks — Flexible Montague Grammar
LP-90-09 Paul Dekker — The Scope of Negation in Discourse, towards a flexible dynamic Montague grammar
LP-90-10 Theo M.V. Janssen — Models for Discourse Markers
LP-90-11 Johan van Benthem — General Dynamics
LP-90-12 Serge Lapierre — A Functional Partial Semantics for Intensional Logic
LP-90-13 Zhisheng Huang — Logics for Belief Dependence
LP-90-14 Jeroen Groenendijk, Martin Stokhof — Two Theories of Dynamic Semantics
LP-90-15 Maarten de Rijke — The Modal Logic of Inequality
LP-90-16 Zhisheng Huang, Karen Kwast — Awareness, Negation and Logical Omniscience
LP-90-17 Paul Dekker — Existential Disclosure, Implicit Arguments in Dynamic Semantics

Mathematical Logic and Foundations

ML-90-01 Harold Schellinx — Isomorphisms and Non-Isomorphisms of Graph Models
ML-90-02 Jaap van Oosten — A Semantical Proof of De Jongh's Theorem
ML-90-03 Yde Venema — Relational Games
ML-90-04 Maarten de Rijke — Unary Interpretability Logic
ML-90-05 Domenico Zambella — Sequences with Simple Initial Segments
ML-90-06 Jaap van Oosten — Extension of Lifschitz' Realizability to Higher Order Arithmetic, and a Solution to a Problem of F. Richman
ML-90-07 Maarten de Rijke — A Note on the Interpretability Logic of Finitely Axiomatized Theories
ML-90-08 Harold Schellinx — Some Syntactical Observations on Linear Logic
ML-90-09 Dick de Jongh, Duccio Pianigiani — Solution of a Problem of David Guaspari
ML-90-10 Michiel van Lambalgen — Randomness in Set Theory
ML-90-11 Paul C. Gilmore — The Consistency of an Extended NaDSet

Computation and Complexity Theory

CT-90-01 John Tromp, Peter van Emde Boas — Associative Storage Modification Machines
CT-90-02 Sieger van Denneheuvel, Gerard R. Renardel de Lavalette — A Normal Form for PCSJ Expressions
CT-90-03 Ricard Gavaldà, Leen Torenvliet, Osamu Watanabe, José L. Balcázar — Generalized Kolmogorov Complexity in Relativized Separations
CT-90-04 Harry Buhrman, Edith Spaan, Leen Torenvliet — Bounded Reductions
CT-90-05 Sieger van Denneheuvel, Karen Kwast — Efficient Normalization of Database and Constraint Expressions
CT-90-06 Michiel Smid, Peter van Emde Boas — Dynamic Data Structures on Multiple Storage Media, a Tutorial
CT-90-07 Kees Doets — Greatest Fixed Points of Logic Programs
CT-90-08 Fred de Geus, Ernest Rotterdam, Sieger van Denneheuvel, Peter van Emde Boas — Physiological Modelling using RL
CT-90-09 Roel de Vrijer — Unique Normal Forms for Combinatory Logic with Parallel Conditional, a case study in conditional rewriting

Other Prepublications

X-90-01 A.S. Troelstra — Remarks on Intuitionism and the Philosophy of Mathematics, Revised Version
X-90-02 Maarten de Rijke — Some Chapters on Interpretability Logic
X-90-03 L.D. Beklemishev — On the Complexity of Arithmetical Interpretations of Modal Formulae
X-90-04 — Annual Report 1989
X-90-05 Valentin Shehtman — Derived Sets in Euclidean Spaces and Modal Logic
X-90-06 Valentin Goranko, Solomon Passy — Using the Universal Modality: Gains and Questions
X-90-07 V.Yu. Shavrukov — The Lindenbaum Fixed Point Algebra is Undecidable
X-90-08 L.D. Beklemishev — Provability Logics for Natural Turing Progressions of Arithmetical Theories
X-90-09 V.Yu. Shavrukov — On Rosser's Provability Predicate
X-90-10 Sieger van Denneheuvel, Peter van Emde Boas — An Overview of the Rule Language RL/1
X-90-11 Alessandra Carbone — Provable Fixed points in $I\Delta_0+\Omega_1$, revised version
X-90-12 Maarten de Rijke — Bi-Unary Interpretability Logic
X-90-13 K.N. Ignatiev — Dzhaparidze's Polymodal Logic: Arithmetical Completeness, Fixed Point Property, Craig's Property
X-90-14 L.A. Chagrova — Undecidable Problems in Correspondence Theory
X-90-15 A.S. Troelstra — Lectures on Linear Logic

1991 *Logic, Semantics and Philosophy of Langauge*

LP-91-01 Wiebe van der Hoek, Maarten de Rijke — Generalized Quantifiers and Modal Logic
LP-91-02 Frank Veltman — Defaults in Update Semantics

Mathematical Logic and Foundations

ML-91-01 Yde Venema — Cylindric Modal Logic
ML-91-02 Alessandro Berarducci, Rineke Verbrugge — On the Metamathematics of Weak Theories
ML-91-03 Domenico Zambella — On the Proofs of Arithmetical Completeness for Interpretability Logic
ML-91-04 Raymond Hoofman, Harold Schellinx — Collapsing Graph Models by Preorders
ML-91-05 A.S. Troelstra — History of Constructivism in the Twentieth Century
ML-91-06 Inge Bethke — Finite Type Structures within Combinatory Algebras

Computation and Complexity Theory

CT-91-01 Ming Li, Paul M.B. Vitányi — Kolmogorov Complexity Arguments in Combinatorics
CT-91-02 Ming Li, John Tromp, Paul M.B. Vitányi — How to Share Concurrent Wait-Free Variables
CT-91-03 Ming Li, Paul M.B. Vitányi — Average Case Complexity under the Universal Distribution Equals Worst Case Complexity
CT-91-04 Sieger van Denneheuvel, Karen Kwast — Weak Equivalence
CT-91-05 Sieger van Denneheuvel, Karen Kwast — Weak Equivalence for Constraint Sets
CT-91-06 Edith Spaan — Census Techniques on Relativized Space Classes
CT-91-07 Karen L. Kwast — The Incomplete Database
CT-91-08 Kees Doets — Levationis Laus
CT-91-09 Ming Li, Paul M.B. Vitányi — Combinatorial Properties of Finite Sequences with high Kolmogorov Complexity

Other Prepublications

X-91-01 Alexander Chagrov, Michael Zakharyaschev — The Disjunction Property of Intermediate Propositional Logics
X-91-02 Alexander Chagrov, Michael Zakharyaschev — On the Undecidability of the Disjunction Property of Intermediate Propositional Logics
X-91-03 V. Yu. Shavrukov — Subalgebras of Diagonizable Algebras of Theories containing Arithmetic
X-91-04 K.N. Ignatiev — Partial Conservativity and Modal Logics
X-91-05 Johan van Benthem — Temporal Logic
X-91-06 — Annual Report 1990
X-91-07 A.S. Troelstra — Lectures on Linear Logic, Errata and Supplement
X-91-08 Giorgie Dzhaparidze — Logic of Tolerance
X-91-09 L.D. Beklemishev — On Bimodal Provability Logics for Π_1-axiomatized Extensions of Arithmetical Theories
X-91-10 Michiel van Lambalgen — Independence, Randomness and the Axiom of Choice
X-91-11 Michael Zakharyaschev — Canonical Formulas for K4. Part I: Basic Results
X-91-12 Herman Hendriks — Flexibele Categoriale Syntaxis en Semantiek: de proefschriften van Frans Zwarts en Michael Moortgat