Kolmogorov—Loveland stochasticity for finite strings

Bruno Durand
LIF, University of Provence

Nikolai Vereshchagin*
Dept. of Mathematical Logic and Theory of Algorithms,
Moscow Lomonossov University,
Leninskie Gory, Moscow 119992

September 9, 2002

Abstract

Asarin [1] showed that any finite sequence with small randomness deficiency has the stability property of the frequency of 1's in their subsequences selected by simple Kolmogorov—Loveland selection rules. Roughly speaking the difference between frequency m/n of zeros and $1/2$ in a subsequence of length n selected from a sequence with randomness deficiency d by a selection rule of complexity k is bounded by $\sqrt{(d+k)/n}$ in absolute value. In this paper we prove a result in the inverse direction: if randomness deficiency of a sequence is large then there is a simple Kolmogorov—Loveland selection rule that selects a sufficiently long subsequence in which frequency of ones is far from $1/2$. Roughly speaking for any sequence of length N there is a selection rule of complexity $O(\log(N/d))$ selecting a subsequence such that $\left|\frac{m}{n}-\frac{1}{2}\right| > \Omega\left(\frac{d}{n\log(N/d)}\right)$.

1 Introduction

One of the first definitions of a random finite sequence of 0s and 1s (=binary string) is due to Kolmogorov [3]. He called a string ε, n-random (he used the term "a table of random numbers" rather then the term "a random string") with respect to a class $\mathcal{R}$ of selection rules if no selection rule in $\mathcal{R}$ selects a subsequence from x of length at least n in which the frequency of 1's deviates from $1/2$ by at least ε. Kolmogorov considered only "admissible" selection rules; the rules addmissible in the sense of [3] were called later "Kolmogorov–Loveland selection rules" (or Kolmogorov–Loveland admissible rules).

*Work done in part while visiting LIF, University of Provence

A Kolmogorov–Loveland selection rule is an algorithm which reads terms of a given finite binary sequence in any order (for example, it may read 7th term first and then 3th term). Before reading each term the algorithm must decide whether this term is selected or not. It is not allowed to read the same term twice. The sequence formed from all the selected terms is called the selected subsequence. The algorithm may be not total and we terminate the selected subsequence once the algorithm is undefined.

An important subfamily of Kolmogorov–Loveland selection rules is formed by Church selection rules (or Church admissible rules). A Kolmogorov–Loveland selection rule is Church admissible if it reads terms of the string in the increasing order (first term, then second term etc.) and is total, that is, it may be applied to every binary string.

Later [4] Kolmogorov defined the notion of Kolmogorov complexity $K(x)$ of a string x and suggested another notion of randomness: a string is random if its complexity is close to its length. More precisely the amount of non-randomness in a string is measured as the difference between its length and its complexity and is called the "randomness deficiency".

In order to distinguish between the two notions of randomness Kolmogorov proposed later to use the term "stochastic" in place of "random" for the first notion and the term "chaotic" for the second one. To compare stochasticity with chaoticity it is natural to reformulate stochasticity using Kolmogorov complexity as follows: first we define Kolmogorov complexity of a selection rule R and then we call a string x Kolmogorov–Loveland ε, n, k-stochastic if no selection rule of complexity at most k selects a subsequence from x of length at least n in which the frequency of 1's deviates from $1/2$ by et least ε. In other words, in the original definition of Kolmogorov, we let the family $\mathcal{R}$ of selection rules be equal to the set of all rules of complexity at most k.

The relations between stochasticity and chaoticity were intensively studied for infinite binary sequences [9, 5, 8, 7, 6]). In contrast, the relations between stochasticity and chaoticity for finite sequences were almost not studied. The first result comparing chaoticity and stochasticity for strings is due to Asarin [1]. He showed that chaoticity implies stochasticity for appropriate choice of parametrs $d(x), \varepsilon, n, k$, Theorem 1 below. No non-trivial implication in the other direction was known. In the present paper we prove the first such implication, Theorem 2 below.

We proceed now to precise definitions.

A Kolmogorov–Loveland selection rule R is specified by two partial computable functions f, g on the set of binary strings. The values of f are positive integers, the values of g are zeros and ones. A sequence selected by R from a sequence $x = x_1, \ldots, x_N$ is defined as follows. Starting with w_0, s_0 equal to the empty words let $k_n = f(w_{n-1})$ (the number of the next observed term), $w_n = w_{n-1}x_{k_n}$ (the sequence of known bits to the algorithm) and $s_n = s_{n-1}x_{k_n}$ if $g(w_{n-1}) = 1$ and $s_n = s_{n-1}$ otherwise (the sequence of selected terms). We terminate these sequences when $f(w_{n-1})$ or $g(w_{n-1})$ are undefined or $k_n \in \{k_1, \ldots, k_{n-1}\}$ or $k_n > N$. Let n be the number of the last step before the termination. The sequence $s = s_n$ of length n is called the

selected subsequence (from the string x by the rule R).

Kolmogorov complexity of strings is defined as follows. A *programming language* is a partial computable function $U(p,x)$ mappimg pairs of binary strings to binary strings. A programming language U is called *optimal* if for any other programming language V there is a constant c such that for every p,x there is p' with

$$|p'| \le |p| + c, \quad U(p',x) = V(p,x).$$

By $|p|$ we denote the length of the string p. By Kolmogorov's theorem [4] optimal programming languages exist. Fix any optimal language and define *Kolmogorov complexity of y conditional to x*, $K(y|x)$, as

$$K(y|x) = \min\{|p| : U(p,x) = y\}.$$

It is easy to verify that for every partial computable function $f(x,y)$ there is a constant c with

$$K(f(x,y)|x) \le K(y|x) + c$$

for all x,y in the domain of f.

The *randomness deficiency* of a string x of length N, $d(x)$, is defined as $N - K(x|N)$.

The notion of complexity is naturally generalized to pairs of strings. Fix any computable one-to-one coresspondence $[x,y]$ between pairs of strings and strings. Then the complexity $K(x,y|z)$ of the pair $\langle x,y\rangle$ conditional to z is defined as $K([x,y]|z)$. Similarly, define $K(x|y,z)$ as $K(x|[y,z])$.

We will use the following inequality relating $K(x,y|z)$, $K(x|z)$, $K(y|x,z)$, which is known as Kolmogorov–Levin theorem [2]. For every positive ε for all sufficiently large N for all string x,y,z of the length at most N it holds

$$K(x,y|z) \le K(x|z) + K(y|x,z) + (1+\varepsilon)\log N, \tag{1}$$

$$K(x|z) + K(y|x,z) \le K(x,y|z) + (3+\varepsilon)\log N. \tag{2}$$

Now we define complexity $K(R|N)$ of a Kolmogorov–Loveland selection rule R. Let the rule R be specified by partial computable functions f,g. Fix any computable one-to-one correspondence $\mathrm{bin}(n)$ between natural numbers and binary strings. Then

$$\begin{aligned} K(R|N) = \min\{|p| :{} & U(p,[0,[\mathrm{bin}(N),w]]) = \mathrm{bin}(f(w)), \\ & U(p,[1,[\mathrm{bin}(N),w]]) = g(w) \text{ for all } w\}. \end{aligned}$$

Informally, $K(R|N)$ is the minimum length of a program of the rule R provided the program is given N.

All logarithms in the paper have the base 2.

Asarin showed that if a selection rule of complexity at most k selects a subsequence of length n having m ones from a sequence whose deficiency is at most d then $|m/n - 1/2|$ is bounded by about $\sqrt{(d+k)/n}$. More specifically the following statement holds:

Theorem 1 ([1]). *For every $\varepsilon > 0$ there is c and $\mu > 0$ such that the following holds for every N, d and every string x of length N with randomness deficiency d. If a Kolmogorov–Loveland selection rule of complexity at most k selects from x a subsequence of length $n > c$ with m 1's and $d + k + 2\log k < \mu N$ then*

$$\left|\frac{m}{n} - \frac{1}{2}\right| < \sqrt{\frac{d + k + 2\log k + (3+\varepsilon)\log N}{2n(1-\varepsilon)\log e}}.$$

Our main result shows that, conversely, if randomness deficiency of a sequence is large then there is a simple selection rule that selects a sufficiently long subsequence in which frequency of ones is far from $1/2$.

Theorem 2. *For some constants c_1, c_2, c_3 for every sufficiently large N and every sequence x of length N and randomness deficiency d such that $d \geq c_1 \log(2N/d)\log N$ there is a Kolmogorov–Loveland selection rule R of complexity at most $c_2 \log(N/d)$ that selects from x a subsequence of length n with m 1's where*

$$\left|\frac{m}{n} - \frac{1}{2}\right| > \frac{d}{32n\log(2N/d)}.$$

Note that the statement of the theorem implies that the selected subsequence is not short: $n \geq \frac{d}{16\log(2N/d)}$, as otherwise $\left|\frac{m}{n} - \frac{1}{2}\right|$ would be greater than $1/2$.

Both theorems give bounds (the first theorem gives the upper bound and the second one gives the lower bound) for the following value

$$f(x,k,n) = \max\{|m/n' - 1/2| : \text{there is a selection rule } R \text{ with } K(R|N) \leq k \text{ that selects from } x \text{ a subsequence of length } n' \geq n \text{ with } m \text{ 1's}\}$$

where $N = |x|$. Theorem 1 gives the upper bound

$$f(x,k,n) < \sqrt{\frac{d + k + 2\log k + (3+\varepsilon)\log N}{2n(1-\varepsilon)\log e}}$$

that holds for certains values of d, N, k and for all strings x of length N and randomness deficiency d and all n. Theorem 2 gives the lower bound

$$f(x,k,n) > \frac{d}{32n\log(2N/d)}$$

that holds for certains values of d, N, k, for all strings x of length N and randomness deficiency d and for *some* n.

Let us analize these bounds. The most interesting is the case when k is small say $k = O(\log N)$ and n is proportional to N: $n \geq cN$ for some constant $c \in (0,1)$. In this case we have more than quadratic gap between the lower and the upper bounds. If d is small say $d = O(\log N)$ then the upper bound is precise for all x and all n: the rule that selects n first terms meets the upper bound, since the randomness deficiency of the prefix x of length n is also $O(\log n)$ and

in every string y with small deficiency the deviation of frequency of 1's from $1/2$ is proportional to $1/\sqrt{|y|}$ [2]. However only large values of d have any interest. Let us analize the upper and the lower bounds when $d \geq c_1 N$ for some constant $c_1 \in (0,1)$. In this case the value of n in the lower bound is at least $c_2 N$ for some positive constant c_2. Theorems 1 and 2 imply therefore the following inequality: For all $c_1 \in (0,1)$ there are $c_2 \in (0,1)$ and $c_3, c_4 > 0$ such that for all $c_5 > 0$

$$c_3 \leq f(x, c_5 \log N, c_2 N) \leq c_4$$

for all sufficiently large N and for all x of length N with $d(x) \geq c_1 N$. An interesting question is: May c_3 be chosen arbitrarily close to c_4 (for some c_1)?

2 The proof of the theorem

We shall use the following theorem that is implicit in [7, proof of Theorem 9.2] (for the seek of completeness we present its proof).

Theorem 3 ([7]). *Given any set $A \subset \{0,1\}^M$ (that is, given the list of elements of A) and any rational K with $|A| < 2^{M-K}$ we can find Church selection rules $R_1 \dots, R_L$ where $L = O(M/K)$ with the following property. For any sequence $y \in A$ there is $j \leq L$ such that R_j selects from y a subsequence of the length n with m ones where $|m/n - 1/2| \geq K/(8n)$.*

Proof. Consider the following game "on credit". The player starting with zero capital observes terms of a sequence y in the increasing order. Before observing a term the player makes a guess of its value and makes a bet; the bet may be any rational number in the interval $[0; 1]$. If the guess is correct the bet is added to the current capital, otherwise it is subtracted. Formally, the definition of a betting strategy is obtained from that of a selection rule by allowing the function g to have any rational values in the segment $[-1; 1]$ and by requiring that $f(w) = |w| + 1$ for all w and that g is total. The capital (=winnings) C_n of the player after n moves is defined as $C_n = C_{n-1} + g(y_1 \dots y_{n-1})$ if $y_n = 1$ and $C_n = C_{n-1} - g(y_1 \dots y_{n-1})$ if $y_n = 0$ (where $C_0 = 0$).

The proof is based on the following theorem [7, Theorem 6.1.3]:

Theorem 4 ([7]). *Given any set $A \subset \{0,1\}^M$ and a rational number K with $|A| < 2^{M-K}$ we can find a betting strategy S such that for any sequence $y \in A$ the winnings of S is at least K.*

Let S be the strategy existing by Theorem 4. Let $l = \lceil 2M/K \rceil$. Consider the strategies $S_1 \dots, S_{2l}$ betting as follows. The bets of $S_1 \dots, S_{2l}$ are always equal to 1 so the only difference between them is in the guessed value. If S guesses the value 1 and makes a bet b then $S_1, \dots, S_{l+\lfloor bl \rfloor}$ guess 1 and all the other strategies guess 0. If S guesses the value 0 of a term and makes a bet b then $S_1, \dots, S_{l+\lfloor bl \rfloor}$ guess 0 and and all the other strategies guess 1. The strategies $S_1 \dots, S_{2l}$ are designed so that if S wins its bet b then $l + \lfloor bl \rfloor$ strategies win

1 and the other loose 1 so the arithmetical mean of winnings of $S_1 \ldots, S_{2l}$ is equal to

$$\frac{l + \lfloor bl \rfloor - (l - \lfloor bl \rfloor)}{2l} = \frac{\lfloor bl \rfloor}{l}.$$

The same applies in the case when S looses its bet b: the average loss of $S_1 \ldots, S_{2l}$ is equal to $\lfloor bl \rfloor / l$. Thus the difference between the winnings on S and the average winnings of $S_1 \ldots, S_{2l}$ is bounded by $|b - \lfloor bl \rfloor / l| \leq 1/l$. The difference between the total winnings on S on every y and the average winnings of $S_1 \ldots, S_{2l}$ on y is bounded by $|x|/l = M/l \leq K/2$ and therefore for any sequence $y \in A$ the average total winnings of $S_1, \ldots, S_{2l}$ on y is at least $K/2$. Hence for any sequence $y \in A$ there is S_j that wins at least $K/2$ on y. Convert S_j into two selection rules, R_{j0} and R_{j1}, defined as follows. Both rules read the input sequence y in the increasing order. The rule R_{j0} selects a term if S_j guesses 0 and the rule R_{j1} selects a term if S_j guesses 1. Let n_0, n_1, m_0, m_1 denote the lengths and the number of 1's in the subsequences selected by R_{j0}, R_{j1} respectively. Then the winnings of R_j is equal to $(2m_1 - n_1) - (2m_0 - n_0)$. Therefore either $2m_1 - n_1 \geq K/4$ or $n_0 - 2m_0 \geq K/4$. In the first case we have $m_1/n_1 - 1/2 \geq K/(8n_1)$. The other case is entirely similar. □

We are given that the sequence x belongs to the set A consisting of all strings of complexity at most $N-d$, which has less than 2^{N-d+1} elements. If we applied Theorem 3 with $M = N$ and $K = d - 1$ we would obtain $L = O(N/(d-1))$ selection rules such that at least one of them selects a subsequence from x with $|m/n - 1/2| \geq (d-1)/8n$, which is much better than the lower bound of Theorem 2. The problem however is that the list elements of A has high complexity and therefore the complexity of the obtained rule might be large. To resolve this problem we will use the following idea due to An. Muchnik [7]: If we are given indices of two non-overlapping substrings y, z of x of low complexity then either using z we are able to find a small set (that is, to find the list of its elements) containing y, or using y we are able to find a small set containing z.

To find such substrings we prove the following

Lemma 5. *For any sufficiently large N every string x of length N with $K(x|N) \leq N-d$ may be chopped into 2^i blocks of the same length where $i \leq \log(N/d)+O(1)$ so that the following holds. There are two consecutive blocks y, z such that both $K(y|u, N), K(z|u, N)$ do not exceed $|y| - d/(2\log(2N/d)) + 4\log N$ where u denotes x without y, z.*

Proof. Let $d' = d/(2\log(2N/d))$. Let y be the first half of x and z the second half of x. If $K(y|N), K(z|N) < N/2 - d' + 4\log N$ we are done. Otherwise, assume w.l.o.g. that $K(y|N) \geq N/2 - d' + 4\log N$. By inequality (2) this implies that $K(z|y, N) \leq N/2 - d + d'$. Let $x_1 = z$ and $u_1 = y$ and repeat the argument for x_1 in place of x. Either we will obtain y_1, z_1 of length $N/4$ with $K(y_1|u_1, N), K(z_1|u_1, N) \leq N/4 - d' + 4\log N$ (in this case we are done), or x_2 of length $N/4$ with $K(x_2|u_2, N) < N/4 - d + 2d'$, where u_2 stands for x without x_2. Repeat this several times. After ith repetition we have x_i of length $N/2^i$ such that $K(x_i|u_i, N) \leq N/2^i - d + id'$, where u_i stands for x without x_i. If we

do not obtain y, z satisfying the theorem within $i = \lceil \log(N/d) \rceil$ steps, we get x_i of length $N/2^i$ with $K(x_i|u_i, N) \leq N/2^i - d + id' = N/2^{i+1} - d'$ (the last equality holds by the choice of i and d'). Therefore for both halves y, z of x_i it holds

$$K(y|u_i, N) < K(x_i|u_i, N) + O(1) < |y| - d' + \log N$$

and

$$K(z|u_i, N) < |z| - d' + \log N. \quad \square$$

Now we are in position to define the selection rule R satisfying Theorem 2. Let i, y, z satisfy Lemma 5. Let u stand for x without y, z. Assume that in the enumeration of strings v with $K(v|u, N) \leq |y| - d/(2\log(2N/d)) + 4\log N$ the string y appears earlier than z (the other case is entirely similar). We read u and z and enumerate strings v of length $N/2^i$ with $K(v|u, N) < |y| - d/(2\log(2N/d)) + 4\log N$ until z appears. Let A denote the set of enumerated strings. Then $\log|A| < \frac{N}{2^i} - \frac{d}{2\log(2N/d)} + O(\log N)$. Apply Theorem 3 with $M = N/2^i$ and $K = \frac{d}{2\log(2N/d)} - O(\log N)$. We obtain selection rules $R_1, \ldots, R_L$. Then we apply the selection rule R_j existing by Theorem 3 and select from y a subsequence of length n with m ones where

$$\left|\frac{m}{n} - \frac{1}{2}\right| \geq \frac{K}{8n}.$$

Let us lowerbound K. We have

$$K = \frac{d}{2\log(2N/d)} - O(\log N) \geq \frac{d}{4\log(2N/d)}$$

where the last inequality holds by the lower bound of d in the condition of the theorem. Therefore we have

$$\left|\frac{m}{n} - \frac{1}{2}\right| \geq \frac{K}{8n} \geq \frac{d}{32\log(2N/d)}$$

To specify R we have to specify i in $\log\log O(N/d)$ bits, the ordinal numbers of blocks y, z in $2\log(N/d)$ bits and j in $\log L$ bits. We have $L = O\left(\frac{M}{K}\right) \leq O\left(\frac{N}{K}\right) \leq O\left(\frac{N}{d}\log\frac{2N}{d}\right)$ and $\log L = O(\log(N/d))$, $K(R) = O(\log(N/d))$. $\square$

Remark 1. It is easy to see from the proof that the values of constants c_1, c_2, c_3 in Theorem 2 may be chosen as $12+\varepsilon, 3+\varepsilon, 6+\varepsilon$ respectively, for any positive ε.

The selection rule constructed in the proof of the theorem is not Church admissible. An interesting question is whether an analog of Theorem 2 holds for Church selection rules.

Theorem 2 can be reformulated as a generalization of Theorem 3 to implicitly given sets. We say that a set A is *given implicitly* if a non-halting program enumerating A is given. For example, for any k there is a non-halting program of length $\log k + O(1)$ that enumerates the set A_k of all string of complexity at

most k. Thus the set A_k has a very short implicit description. On the other hand the Kolmogorov complexity of the list of elements of A_k is at least $k-O(1)$. Indeed, given that list we are able to specify a string of complexity more than k (the first string of complexity larger than k, in the lexicographical order).

Theorem 6. *For every positive ε for sufficiently large N the following holds. Given implicitly a set $A \subset \{0,1\}^N$ and given N and a d such that $\log|A| \leq N-d$ and $d \geq (12+\varepsilon)\log(2N/d)\log N$ we can find $O((\frac{N}{d})^{3+\varepsilon})$ selection rules such that for any $x \in A$ at least one of them selects from x a subsequence with*

$$\left|\frac{m}{n} - \frac{1}{2}\right| > \frac{d}{32\log(2N/d)}.$$

Proof. Let p be the given non-halting program enumerating A. For any $x \in A$ it holds $K(x|p,N) \leq N - d + O(1)$. Note that in the proof of Theorem 2 we actually constructed $O((N/d)^{3+\varepsilon})$ selection rules such that for any x with $K(x|N) \leq N-d$ at least one of them satisfies the theorem. Relativizing the proof by p we get the proof of Theorem 6. □

References

[1] E.A. Asarin. Some properties of Kolmogorov δ random sequences. SIAM Theory of Prob. Appl., 32 (1987) 507–508.

[2] M. LI, P. Vitanyi. Introduction to Kolmogorov complexity and its applications. Springer Verlag, 1997.

[3] A.N. Kolmogorov. On tables of random numbers. Sankhyã: The Indian Journ. of Statistics, Series A, v. 25, Part 4 (1963). Reprinted in Theoretical Computer Science 207 (1998) 387–395.

[4] A.N. Kolmogorov. Three approaches to the quantitative definition of information. Problems of Information Transmission 1(1) (1965) 1–7.

[5] P. Martin-Löf. The definition of random sequences. Inform. and Control 9 (1966) 602–619.

[6] W. Merkle. The Kolmogorov-Loveland stochastic sequences are not closed under selecting subsequences. To appear in Proc. of ICALP'02.

[7] An.A. Muchnik, A.L. Semenov, V.A. Uspensky. Mathematical metaphysics of randomness. Theor. Comp. Scie. 207 (1998) 263–317.

[8] C.P. Schnorr. Process complexity and effective random sets. J. Comp. Systems Scie. 7 (1973) 376–388.

[9] Die Wiederspruchsfreiheit des Kollektivbegriffes. Actualitès Scie. Industr. 735 (1938) 79–99.