Discrepancies of hypergraphs of high Kolmogorov complexity

Willem L. Fouché
Department of Quantitative Management, University of South Africa, PO Box 392, 0003 Pretoria, South Africa
E-mail:fouchwl@unisa.ac.za

Dedicated to Ernie Cockayne on the occasion of his 60th birthday

Abstract

We study the discrepancies of set systems whose incidence matrices are encoded by binary strings which are complex in the sense of Kolmogorov-Chaitin. We show that these systems display an optimal degree of irregularity of distribution.

1 Introduction

Random sequences were introduced by Richard von Mises [20] under the name of "collectives", as a foundation of probability theory. The crucial features characterising collectives are, on the one hand, the existence of limiting frequencies within the sequence of outcomes (uniformity of distribution of the values of the sequence) and, on the other hand, invariance of the limiting frequencies for subsequences chosen by means of any gambling strategy against the random sequence. (For the history and a further development of these ideas, see [18]). In the 1960s there emerged two approaches to random sequences by Solomonoff, Kolmogorov and Chaitin [10, 3] and Martin-Löf [11], respectively, which are based on two rather distinct intuitions. The definition of Martin-Löf, in analogy with van Mises' conception, is based on the idea that a random sequence should satisfy all properties which hold with probability one. Strictly speaking, this is, of course, quite impossible but Martin-Löf delineated a countable class of such properties which should be satisfied by a random sequence. The definition of Kolmogorov et al, on the other hand, is based on the intuition

JCMCC **33** (2000), pp. 199-207

that a random sequence should be its own shortest description. The initial definitions worked for finite sequences only but were refined in the 1970s by many people. (See [19, 4, 18, 16] for a thorough discussion.) Thereafter, it was found that the two approaches to randomness lead to the same class of binary strings, the so-called Kolmogorov-Chaitin strings.

Irregularities of distribution are studied in such diverse areas as combinatorics, number theory and the theory of disordered media. In the combinatorial context, one considers partitions of the elements of hypergraphs and measures the irregularity of distribution of the hyperedges with respect to the partitions. For example, we consider the following classical result of Roth [13]: Let H_n be the hypergraph on $[n] := \{1, \cdots, n\}$ having as hyperedges all the arithmetical progressions contained in $[n]$. Roth's theorem states that for each partition $\chi : [n] \to \{-1, 1\}$, one can always find some hyperedge E of H_n such that

$$|\sum_{x \in E} \chi(x)| \gg n^{1/4}.$$

In other words, there is a limit, of order $n^{1/4}$, to the degree in which the elements of all the hyperedges of H_n can be uniformly partitioned into two classes. On the other hand, Van der Waerden [17] has shown that for given k one can always find some n such that for each partition $\chi : [n] \to \{-1, 1\}$, there is some hyperedge E of size k which is monochromatic with respect χ; that is, one can always find in one of the blocks of a partition an arithmetic progression of length k. In both cases we have an irregularity of distribution (a lack of statistical uniformity in at least one hyperedge of the hypergraph), with respect to partitions, on the one hand, which is manifested as an unavoidable regularity, or a prescribed organisation, on the other hand. (For background on these topics see [9, 2, 12].)

Let us agree to call an infinite binary string $\alpha = \alpha_1 \alpha_2 \ldots$ *complex* when it is random in the sense of Kolmogorov, Martin-Löf, Chaitin et al. In the papers [5, 8], it was shown how complex partitions give rise to irregularities of distribution, in the Ramsey-theoretic sense, of countable combinatorial configurations. In this paper we show that complex strings, when viewed as codes for hypergraphs, themselves display a certain optimal degree of irregularity of distribution.

Let $\mathcal{A}$ be a family of subsets of some finite set A and let Σ be a family of mappings $\chi : A \to \{-1, 1\}$. The *discrepancy* of $\mathcal{A}$ with respect to Σ is given by

$$\min_{\chi \in \Sigma} \max_{X \in \mathcal{A}} |\Sigma_{x \in X} \chi(x)|.$$

If Σ is the set consisting of all the mappings $\chi : A \to \{-1, 1\}$, we simply speak of the discrepancy of $\mathcal{A}$ and denote it by $\delta(\mathcal{A})$. Roth's theorem states exactly that $\delta(H_n) \gg n^{1/4}$.

Let $(i,j) \mapsto < i,j >$ be a recursive bijection from $\mathbf{N} \times \mathbf{N}$ onto $\mathbf{N}$. For a complex string α, we define the family $\mathcal{A}(\alpha) = (A_i)_{i \geq 1}$ of subsets of $\mathbf{N}$ by:

$$j \in A_i \Leftrightarrow \alpha_{ij} = 1.$$

(Here we wrote ij instead of $< i,j >$.) For $n \geq 1$, let $\mathcal{A}_n(\alpha)$ be the family of sets $A_i \cap [n]$, $i = 1, \cdots, n$. In other words, the hypergraph $\mathcal{A}_n(\alpha)$ has $[\alpha_{ij} : 1 \leq i,j \leq n]$ as incidence matrix. We shall prove

Theorem 1 *There exists a universal constant $\tau > 0$, such that, for each complex string α, there exists a natural number n_α such that, for all $n \geq n_\alpha$, the discrepancy of the set system $\mathcal{A}_n(\alpha)$ satisfies*

$$\delta(\mathcal{A}_n(\alpha)) \geq \tau\sqrt{n}. \tag{1}$$

Consequently, for $n \geq n_\alpha$, for each $\chi : [n] \to \{-1,1\}$, there is some $X \in \mathcal{A}_n(\alpha)$ such that

$$|\sum_{j \in X} \chi(j)| \geq \tau\sqrt{n}.$$

The inequality (1) is essentially the best possible, for Spencer [15] has shown that for any set system $\mathcal{A}$ consisting of n subsets of $[n]$, it is the case that

$$\delta(\mathcal{A}) \leq 6\sqrt{n}.$$

The proof of Theorem 1 appears in Section 2. The proof essentially boils down to a constructivisation of some of the arguments in [15]. In Section 3 we study the discrepancies of recursive hypergraphs with respect to complex partitions.

2 Proof of Theorem 1

If A is a set, we write A^* for the set of words over A. If $s \in A^*$, we write $|s|$ for the length of s. We denote the set $\{0,1\}^{\mathbf{N}}$ of infinite binary sequences by $\mathcal{N}$. We topologise this space (the Baire space) by the product topology. If $\alpha = \alpha_1\alpha_2 \ldots \in \mathcal{N}$, we write $\overline{\alpha}(n)$ for the word $\alpha_1\alpha_2 \ldots \alpha_n$. For $s \in \{0,1\}^*$, we set $[s] = \{\alpha \in \mathcal{N} : \overline{\alpha}(n) = s\}$, where $n = |s|$. We write λ for the unique measure on the Borel subsets of $\mathcal{N}$ with the property that $\lambda([s]) = 2^{-|s|}$ for each of the respective sets $[s]$ (the Lebesgue measure). We shall need the following

Theorem 2 *If $(A_n)_{n \geq 1}$ is a sequence of open subsets of $\mathcal{N}$ such that the relation $[s] \subset A_n$ is recursively enumerable in $s \in \{0,1\}^*$ and $n \in \mathbf{N}$ and if $\sum_n \lambda(A_n) < \infty$, then, if α is complex, $\alpha \in A_n$ for at most finitely many n.*

This theorem can be viewed as a constructive version of the second Borel-Cantelli lemma. A proof of this result appears in [14].

For the proof of Theorem 1 we shall need the following

Lemma 1 *Let $(W_{i,j} : 1 \leq i,j \leq n)$ be n^2 independent random variables such that every $W_{i,j}$ assumes each of the values 0 or 1 with probability $1/2$. For each $\nu = (\nu_1, \cdots, \nu_n) \in \{-1,1\}^n$, set*

$$L_i(\nu) = \sum_{j=1}^{n} \nu_j W_{i,j}.$$

Then, for some real number $c_0 < 1/2$ and some positive rational number $\tau > 0$, we have, for n sufficiently large, that

$$Prob\,[\exists_{\nu\in\{-1,1\}^n}\, \forall_{i\leq n}(|L_i(\nu)| < \tau\sqrt{n}\,)\,] \leq (2c_0)^n.$$

We first show how Theorem 1 follows from this lemma:

Let A_n be the subset of $\mathcal{N}$ defined by the following condition:

$$\alpha \in A_n \Leftrightarrow \exists_{\nu\in\{-1,1\}^n}\forall_{i\leq n}\,(|\sum_{j=1}^{n} \alpha_{ij}\nu_j| < \tau\sqrt{n}\,),$$

where τ is the rational number that appears in the formulation of Lemma 1. It is clear that each A_n is open and that the relation $[s] \subset A_n$ is recursively enumerable in s, n. Moreover, it follows from Lemma 1 that

$$\sum_n \lambda(A_n) < \infty.$$

By Theorem 2, we have, for each complex α, that $\alpha \notin A_n$, for all n sufficiently large. This completes the proof of Theorem 1.

PROOF OF LEMMA 1: Since, for fixed ν, the random variables $L_i(\nu)$ are independent and there are 2^n possible values for ν, it suffices to show, for n large, for each $\nu \in \{-1,1\}^n$ and $i \leq n$ that

$$\text{Prob}\,[|L_i(\nu)| < \tau\sqrt{n}\,] \leq c_0,$$

for some $0 \leq c_0 < 1/2$. Fix $i \in [n]$ and $\nu \in \{-1,1\}$. By symmetry, we may assume that

$$|\{j : \nu_j = -1\}| \geq |\{j : \nu_j = 1\}|.$$

Denote the cardinalities of these sets by $k + l$ and k, respectively. Then $L_i(\nu)$ has the same distribution as

$$Z := X_1 + \ldots + X_k - (Y_1 + \ldots + Y_{k+l}),$$

where $X_1, \cdots, X_k, Y_1, \cdots, Y_{k+l}$ are independent random variables each assuming the values 0 or 1 with probability 1/2. Set

$$N = 2(X_1 - \frac{1}{2}) + \ldots + 2(X_k - \frac{1}{2}) + 2(-Y_1 + \frac{1}{2}) + \ldots + 2(-Y_{k+l} + \frac{1}{2}).$$

Then N is a sum of n independent variables each of mean 0 and variance 1. It follows from the central limit theorem that for reals $A \leq B$:

$$\text{Prob}\,[A \leq \frac{N}{\sqrt{n}} \leq B] \longrightarrow \int_A^B \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt,$$

as $n \to \infty$. Choose $\tau > 0$ such that

$$\int_{-2\tau}^{5\tau} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt \leq \frac{1}{4}.$$

Since $N = 2Z + l$, the inequality $|Z| < \tau\sqrt{n}$ is equivalent to:

$$-2\tau + \frac{l}{\sqrt{n}} < \frac{N}{\sqrt{n}} < 2\tau + \frac{l}{\sqrt{n}}.$$

We consider two cases.
CASE 1: $l \leq 3\tau\sqrt{n}$. If $|Z| < \tau\sqrt{n}$, we have:

$$-2\tau < \frac{N}{\sqrt{n}} < 5\tau,$$

Therefore, by our choice of τ:

$$\text{Prob}\,[|Z| < \tau\sqrt{n}] \leq \frac{1}{4} + o(1),$$

as required.
CASE 2: $l > 3\tau\sqrt{n}$. One sees that $N/\sqrt{n} > \tau$ if $|Z| < \tau\sqrt{n}$. But

$$\text{Prob}\,[\frac{N}{\sqrt{n}} > \tau] = \int_\tau^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt + o(1),$$

which is $\leq c_0$ for come absolute $c_0 < 1/2$, for n sufficiently large. This completes the proof of Lemma 1.
REMARK. In [6, 7] the author shows that each complex string can be represented as a "generic" Brownian motion on the unit interval. This opens the possibility of displaying the irregularities of distribution of complex strings in generic Brownian motion. This line of thought will be pursued in a sequel to this paper.

3 Complex partitions

If A is an $\omega \times \omega$-matrix over $\{0,1\}$ (i.e. each entry of A is either 0 or 1), we write A^n for the $n \times n$-submatrix in the northwestern corner of A. If X is a countable column vector over $\{-1,1\}$, we write $\overline{X}(n)$ for the first n entries of X when n is at most the length of X. For an $n \times n$-matrix M over $\{0,1\}$ and a column vector X of length n over $\{-1,1\}$, we write

$$\|MX\| = \sup\{|M_i X| : i = 1, \cdots, n\},$$

where M_i denotes the ith row of M.

Proposition 1 *Suppose A is an $\omega \times \omega$ matrix over $\{0,1\}$. There is a countable column vector X over $\{-1,1\}$ and an absolute constant $C > 0$, such that, for all $n > 1$:*

$$\|A^n \overline{X}(n)\| \leq C\sqrt{n \log n}.$$

PROOF. The proof is based on König's infinity lemma together with the following basic probabilistic result: If $X_1, \cdots, X_n$ are independent random variables assuming each of the values -1 or 1 with a probability $1/2$, then, writing

$$S_n = X_1 + \ldots + X_n,$$

we have, for $\alpha > 0$:

$$\text{Prob}\,[|S_n| > \alpha] \leq 2e^{-\alpha^2/2n}. \tag{2}$$

This is known as Chernoff's inequality. A proof can be found in [1].

The inequality (2) has the following implication:

Lemma 2 *There exists a universal constant $C > 0$, such that, if M is any $n \times n$-matrix over $\{0,1\}$, there is column vector X of length n over $\{-1,1\}$ such that, for all $2 \leq k \leq n$:*

$$\|M^k \overline{X}(k)\| \leq C\sqrt{k \log k}. \tag{3}$$

We first deduce Proposition 1 from Lemma 2: Recall that a subset T of $\{-1,1\}^*$ is a tree iff, whenever $s \in T$ and t is a left factor of s, then $t \in T$. It follows from the König infinity lemma that if a tree T is infinite, then T contains an infinite branch. In terms of the matrix A as in the statement of Proposition 1, we define a tree T as follows: We place all words of length

≤ 1 in T. If $n > 1$ and s is of word of length n, then $s \in T$, iff, when s is viewed as a column vector over $\{-1, 1\}$, we have, for $2 \leq k \leq n$:

$$\|A^k \overline{s}(k)\| \leq C\sqrt{k \log k},$$

where C is given by Lemma 2. It follows from Lemma 2 that T has infinitely many elements. Any infinite branch of T will define an infinite column vector X over $\{-1, 1\}$ which will have the required property.

PROOF OF LEMMA 2: Let X be a random column vector such that its components $X_1, \cdots, X_n$ are independent random variables assuming each of the values -1 or 1 with probability $1/2$. Then, for $2 \leq k \leq n$ and $1 \leq i \leq k$ the linear form $M_i^k \overline{X}(k)$ has the same distribution as $S_{l(i)}$ for some $l(i) \leq k$. It follows from (2) that, for $D > 0$ and $l(i) > 0$:

$$\text{Prob } [|M_i^k \overline{X}(k)| > \sqrt{Dk \log k}\,] \leq 2\exp(-Dk \log k/2l(i)).$$

Since $l(i) \leq k$, the probability is always bounded from above by $k^{-D/2}$. We conclude that

$$\text{Prob}[\,\exists_{2 \leq k \leq n}\, |M^k \overline{X}(k)| > \sqrt{Dk \log k}\,] \leq 2 \sum_{k \geq 2} \frac{k^2}{k^{D/2}},$$

which is < 1 for D sufficiently large. This concludes the proof of Lemma 2.

It is an interesting open problem whether, for given A as in the formulation of Proposition 1, the vector X is recursive in A. The tree T constructed in the proof of the proposition is recursive in A. This, however does not necessarily imply that one can find an infinite branch which is recursive in T. Another probably very difficult problem is whether the proposition holds with an upper bound of the form $\ll \sqrt{n}$ instead of $\ll \sqrt{n \log n}$. For this purpose it will suffice to refine Spencer's theorem in [15] to prove a version of Lemma 2 with (3) replaced by:

$$\|M^k \overline{X}(k)\| \leq C\sqrt{k},$$

for all $2 \leq k \leq n$.

In the following theorem, we view any $\alpha \in \mathcal{N}$ as an infinite column vector over $\{-1, 1\}$ where now $\mathcal{N}$ is the Baire space $\{-1, 1\}^{\mathbf{N}}$. We call an $\omega \times \omega$ matrix $A = (a_{ij})$ over $\{0, 1\}$ recursive if the relation $a_{ij} = 1$ is recursive in i, j.

Theorem 3 *Let A be a recursive countable matrix over $\{0, 1\}$. There is a universal constant $C > 0$ such that, for every complex α, there is some n_α, such that, for all $n \geq n_\alpha$:*

$$\|A^n(\overline{\alpha}(n))\| \leq C\sqrt{n \log n}. \tag{4}$$

PROOF. For $n \geq 1$ define the subset B_n of $\mathcal{N}$ by

$$\beta \in B_n \Longleftrightarrow \exists_{m \geq n} (\|A^m(\overline{\beta}(m))\| > \sqrt{Dm \log m}),$$

where $D > 4$ is a fixed rational number. It is clear that B_n is an open set. Moreover, if $s \in \mathcal{N}$ and $|s| \geq n$ then, writing $m = |s|$:

$$[s] \subset B_n \Longleftrightarrow \|A^m(s)\| > \sqrt{Dm \log m}.$$

Since A is recursive, this is a recursively enumerable relation between n and s. Therefore, there is a total recursive mapping $(n, m) \mapsto s_{nm}$ from $\mathbf{N}^2$ to $\mathcal{N}$ such that for all n:

$$B_n = \cup_m [s_{nm}].$$

It follows from the proof of Lemma 2 that the Lebesgue measure $\lambda(B_n)$ of B_n satisfies:

$$\lambda(B_n) \leq \sum_{m \geq n} \frac{m^2}{m^{D/2}} \ll \frac{n^2}{n^{D/2}}.$$

We conclude that $B := \cap_n B_n$ is a set of constructive measure 0. It is well-known that a set of constructive measure 0 contains no complex strings. (See, for example [4].) We can therefore conclude that B contains no complex strings. This means exactly that if α is complex, then (4) holds for all n sufficiently large.

References

[1] Alon, N. and Spencer, J.H.: *The probabilistic method*, Wiley, New York, 1987.

[2] Beck J. and Chen, W: *Irregularities of distributions*, Cambridge University Press, 1987.

[3] Chaitin, G.J.: On the length of programs for computing binary sequences, *J. Assoc. Comput. Mach.* **13** (1966), 547-569.

[4] Chaitin, G.A.: *Algorithmic information theory*, Cambridge University Press, 1987.

[5] Fouché, W.L.: Descriptive complexity and reflective properties of combinatorial configurations, *J. Lond. Math. Soc.* **54** (1996), 199-208.

[6] Fouché, W.L.: Arithmetic representations of Brownian motion I, *J. Symb. Logic* (to appear).

[7] Fouché, W.L.: Arithmetic representations of Brownian motion II, (submitted).

[8] Fouché, W.L. and Potgieter, P.H.: Kolmogorov complexity and symmetric relational structures, *J. Symb. Logic* **63** (1997), 1083-1094.

[9] Graham, R.L., Rothschild, B.L. and Spencer, J.H: *Ramsey theory*, Wiley, New York, 1990.

[10] Kolmogorov, A.N.: Three approaches to the quantitative definition of randomness, *Probl. Inform. Transmission* **1** (1965), 1-7.

[11] Martin-Löf, P.: The definition of random sequences, *Information and Control* **9** (1966), 602-619.

[12] Nešetřil, J.: Ramsey theory, in R.L. Graham, M. Grötschel and L. Lovász, eds., *Handbook of Combinatorics*, Vol. 2, North Holland, 1995, 1331-1403.

[13] Roth, K.F.: Remarks concerning integer sequences, *Acta Arith.* **9** (1964), 257-260.

[14] Shen, A.Kh.: Connections between different algorithmic definitions of randomness, *Soviet Math. Dokl.* **38** (1989), 316-319.

[15] Spencer, J.H.: Six standard deviations suffice, *Trans. Amer. Math. Soc.* **289** (1986), 679-706.

[16] Uspensky, V.A. and Semenov, A.L.: What are the gains of a theory of algorithms, *Algorithms in modern mathematics and computer science*, Lecture Notes in Computer Science **122** (ed.A.P. Ershov and D.E. Knuth), Springer-Verlag, New York.

[17] van der Waerden, B.L.: Beweis einer Baudetschen Vermutung, *Nieuw Arch. Wisk.* **1** (1927), 212-216.

[18] van Lambalgen, M.: Von Mises' definition of random sequences reconsidered, *J. Symb. Logic* **52** (1987), 725-755.

[19] Vitányi, P. and Li, M.: *An introduction to Kolmogorov complexity and its applications*, Springer-Verlag, New York, 1993.

[20] von Mises R.: Grundlagen der Wahrscheinlichkeitsrechnung, *Math. Zeitschrift* **5** (1919), 52-99.