DISCREPANCY OF COMPLEX SEQUENCES VIA COMPRESSIBILITY

GEORGE DAVIE

Dedicated to a fellow swimmer Ernie Cockayne. Happy 60th birthday!

ABSTRACT. We use the idea of compressibility to examine the discrepancy of set systems coded by complex sequences.

1. INTRODUCTION:

Let ϕ be a universal prefix machine, fixed in what follows. Call an infinite binary sequence α *complex* if there is a constant c such that, for each initial segment $\alpha_{1:n}$ length n of α there is no program shorter than $n - c$ which outputs $\alpha_{1:n}$, see, for example [7]. Let $c(\alpha)$ be the least c for which the above holds. We call $c(\alpha)$ the *compressibility* of α.

In the papers [3,4,5] Fouché, Fouché et al examine properties of complex sequences when viewed as coding for combinatorial configurations.

This leads to interesting properties of such sequences which are not apparent when they are viewed merely as binary sequences.

The paper [6] is a continuation of this theme and studies the discrepancy of complex sequences when seen as coding for set systems. In this note we will use a technique introduced in [1] to lift the effective content of the results in [6].

2. NOTATION

For the sake of clarity, we will use the same notation as [6]. Let $\mathcal{A}$ be a family of subsets of a finite set A and $\sum$ a set of mappings from A to the set $\{-1, 1\}$. The *discrepancy of $\mathcal{A}$ with respect to $\sum$* is

$$\min_{\lambda \in \sum} \max_{X \in \mathcal{A}} |\sum_{x \in X} \lambda(x)|.$$

For $\sum$ the entire set of mappings from A to $\{-1, 1\}$ we will talk of the *discrepancy of A*. We will sometimes talk of elements of $\sum$ as particular *colourings* of the elements of A. Denote the discrepancy by $\sigma(\mathcal{A})$. Since we are going to view binary sequences as codes for the entries of matrices, we let $(i, j) \longmapsto < i, j >$ be a recursive bijection from $\mathcal{N} \times \mathcal{N}$ onto $\mathcal{N}$. For a binary sequence α then, we define the family $\mathcal{A}(\alpha) = (A_i)_{i \geq 1}$ of subsets of $\mathcal{N}$ by $j \in A_i \Leftrightarrow \alpha_{ij} = 1$ where we now write ij for $< i, j >$. For $n \geq 1$, let $\mathcal{A}_n(\alpha)$ be the family of sets $A_i \cap [n], i = 1, ..., n$. That is, we can see $\mathcal{A}_n(\alpha)$ as the set system given by the $n \times n$ matrix in the top lefthand corner of the infinite matrix generated by α. For brevity, when we are dealing directly with the matrix, we will often talk of the *discrepancy of $n \times n$ submatrices* instead of the *discrepancy of the set system given by the submatrix*. For a binary string x

we will denote by $|x|$ the length of x. We will often identify algorithms with their binary representation and will write $|\phi|$ for the length of a binary representation of an algorithm ϕ. Also, $\log x$ will denote the base 2 logarithm of x.

The first results in [6] are for the case where the matrix is infinite and is generated by a complex sequence α. For each of the top left hand $n \times n$ submatrices, the discrepancy of the set system represented by this submatrix is considered, this leads to Theorem 1 of [6]. Secondly the case where the matrix is recursive and the colouring complex, is examined. This leads to Theorem 3 in [6].

In this note we use the compressibility of a complex sequence α to prove more effective versions of these results.

3. Complex matrices.

Theorem 1 of [6] states:

Theorem 1. *(Fouché) 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 $\mathcal{A}_n(\alpha)$ satisfies $\sigma(\mathcal{A}_n(\alpha)) \geq \tau\sqrt{n}$.*

The idea in using compressibility is to show that the compressibility of a sequence must increase the longer the discrepancy stays low. Put differently, we show that for given c, if the discrepancy $\mathcal{A}_n(\alpha)$ stays low for too large n, then the compressibility of α must be larger than c. We prove the following:

Theorem 2. *There exists a universal constant $\tau > 0$ such that, for each complex string α with $c(\alpha) = c$ we can find a natural number n_c such that, for all $n \geq n_c$, the discrepancy of $\mathcal{A}_n(\alpha)$ satisfies $\sigma(\mathcal{A}_n(\alpha)) \geq \tau\sqrt{n}$.*

In other words, the stage after which the discrepancy must be at least $\tau\sqrt{n}$ is recursive in $c(\alpha)$. We use the following lemma from [6].

Lemma 1. *(Fouché) Let $(W_{i,j} : 1 \leq i,j \leq n)$ be n^2 random variables such that every $W_{i,j}$ assumes each of the values 0 or 1 with probability 1/2. For each $v = (v_1, v_2, ..., v_n) \in \{-1, 1\}^n$, set $L_i(v) = \sum_{j=1}^{n} v_j W_{i,j}$. Then, we can find numbers $\tau, \varepsilon, c_0 < 1/2 - \varepsilon$ and n such that*

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

Proof. Directly from Lemma 1 in [6]. The effectivity follows from the following error terms for convergence to the normal distribution, see [2].
Let $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ be the normal density function, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy$ the normal distribution function and let $a_k = b(v+k; 2v, \frac{1}{2})$ where b is the binomial distribution. Following Feller, [2], we see that for $h = 2/\sqrt{n}$ we have (1) $a_k \sim h\varphi(kh)$ with error smaller than $\frac{k^3}{n^2}$. Hence we can, given $\varepsilon > 0$ choose n large enough such that $(1-\varepsilon)[\Phi(z_2) - \Phi(z_1)] < \sum_{\frac{1}{2}z_1\sqrt{n} < k < \frac{1}{2}z_2\sqrt{n}} a_k < (1+\varepsilon)[\Phi(z_2) - \Phi(z_1)]$. ∎

Proof of theorem 2: Consider a complex string α with $c(\alpha) = c$. Let $\phi : N \times N \mapsto N$ be our recursive bijection. Now consider the sequence of $n \times n$ submatrices in the top lefthand corner of the generated infinite matrix. It follows by the lemma that for n as in the lemma the probability that the submatrix has discrepancy lower

than $\tau\sqrt{n}$ is less than $(1/2 - 2\varepsilon)^n$. We can therefore specify an A^n with discrepancy lower than $\tau\sqrt{n}$ by giving the lexicographic position of A^n amongst the $n \times n$ matrices with discrepancy less than $\tau\sqrt{n}$.

We show that for given c, we can find an n large enough that this description will lead to an algorithm of length less than $n' - c$ for an initial segment $\alpha_{1:n'}$ of α. Let $\alpha_{1:n'}$ be the shortest initial segment of α which gives each of the values of A^n. We will show how to choose n in order to obtain a program for $\alpha_{1:n'}$ of length less than $n' - c$.

To be specific, to specify $\alpha_{1:n'}$ we will need τ, ε and the position of A^n amongst those matrices B^n with this low discrepancy.

Now for given c take n large enough such that

$$(1/2 - 2\varepsilon)^n < 2\exp(-c - 2\log\tau - 2\log\varepsilon - 2\log c - |\phi| - 1)$$

We claim that there is a universal constant k (which we can find) such that any complex α which has low discrepancy up to here must have $c(\alpha_{1:n'}) > c - k$.

This follows from the fact that we can write a program φ outputting any of the strings x length n' which code for low discrepancy up to A^n under ϕ. Our program takes as input a concatenation of the following:

1) A self delimiting code for c (length less than $2\log c$),
2) self delimiting code for ϕ and the constants τ, ε and
3) the lexicographic position s (of length at most $n' - c - 2\log\tau - 2\log\varepsilon - 2\log c - |\phi|$) of $\alpha_{1:n'}$ in the set of strings length n' which, under ϕ, code for matrices with A^n having this low discrepancy.

Our program reads c and the codes for τ, ε and finds the first n such that $(1/2 - 2\varepsilon)^n < 2\exp(-c - 2\log\tau - 2\log\varepsilon - 2\log c - |\phi| - 1)$. It now finds n' and reads the position length $(n' - c - 2\log\tau - 2\log\varepsilon - 2\log c - |\phi| - 1)$ and outputs $\alpha_{1:n'}$. Note that φ has total input length less than

$$(2\log c + 2\log\tau + 2\log\varepsilon + |\phi|) + (n' - c - 2\log\tau - 2\log\varepsilon - 2\log c - |\phi| - 1)$$
$$= n' - c - 1.$$

Clearly if we take $k = |\varphi|$ our claim holds. ∎

1. Complex partitions

The case where our countable matrix is recursive and our partition (column vector) complex is now considered.

Notation: For A an $\omega \times \omega$ matrix over $\{0,1\}$ let A^n be the $n \times n$ submatrix in the upper left corner of A. For X a countable column vector over $\{-1,1\}$ write $X(n)$ for the first n entries of X. For an $n \times n$ matrix B and column vector X as above we write $||BX||$ for $\sup\{|B_i X| : i = 1, \ldots, n\}$, where B_i denotes the ith row of B. The following theorem appears in [6]:

Theorem 3. *(Fouché) 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 $n \geq n_\alpha$,*

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

We will prove the following:

Theorem 4. *Let A be a recursive countable matrix over $\{0,1\}$. There is a universal constant $C > 0$ such that, for every complex α, we can effectively find in $c(\alpha) = c$ an n_c, such that, for all $n \geq n_c$*

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

We use the following

Lemma 2. *(Fouché) Let M be an $n \times n$ matrix and let X be a complex column vector over $\{-1, 1\}$ with each entry taking each of -1 and 1 with probability $1/2$. Then*

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

Proof of theorem 3: Choose $D > 4$. Let φ be an algorithm for A. Given n, consider those column vectors for which $[\exists_{k \geq n} |M^k \bar{X}(k)| > \sqrt{Dk \log k}]$.
We claim that we can find a universal constant r such that any column vector X for which $\exists_{k \geq n^*} |A^k \bar{X}(k)| > \sqrt{Dk \log k}$, where n^* is such that

$$2 \sum_{k \geq n^*} \frac{k^2}{k^{D/2}} < 2 \exp(-c - 2 \log c - |\phi|)$$

must have an initial segment $X_{1:l}$ compressibility by more than $c - r$. This follows from the fact that few column vectors will satisfy the condition.

Indeed, consider any such column vector Y. The following program φ outputs an initial segment of Y. φ operates on a concatenation of inputs of the form:

1) Self delimiting codes for c (shorter than $2 \log c$), ϕ and
2) l (to be specified) (shorter than $2 \log l$)
3) The position s of the initial segment of Y amongst the column vectors for which

$$\exists_{k \geq n^*} |M^k \bar{X}(k)| > \sqrt{Dk \log k}.$$

The algorithm works as follows: φ reads c and then finds the first n^* such that

$$2 \sum_{k \geq n^*} \frac{k^2}{k^{D/2}} < 2 \exp(-c - 2 \log c - |\phi|)$$

Now φ reads l where l is that number such that X appears between the first time that

$$2 \sum_{k \geq n^*} \frac{k^2}{k^{D/2}} < 2 \exp(-c - 2 \log c - |\phi| - l)$$

and the first time that

$$2 \sum_{k \geq n^*} \frac{k^2}{k^{D/2}} < 2 \exp(-c - 2 \log c - |\phi| - l - 1)$$

φ now enumerates all column vectors appearing between these two times and extends them all to the maximum vector length listed between these two times.
φ then reads the position s of Y in this set of measure less than $2 \exp(-c - 2 \log c - |\phi| - l)$ and outputs Y.

Note that the length of the input for φ is bounded by:

$$2\log c + |\phi| + 2\log l + (n - c - 2\log c - |\phi| - l) = n - c - l + 2\log l < n - c.$$

Clearly if we take $r = |\varphi|$ the claim holds. ∎

So even though we clearly can, given a complex vector α and fixed n, tailor make the initial segment A^n of A to make the discrepancy of the set system given by A^n very high, the complexity of α and the recursiveness of A rules out the discrepancy being high for infinitely many n.

References

[1] Davie, G.: Recursive events in random sequences, (submitted)

[2] Feller, W.: *An introduction to probability theory and its applications* (Wiley, 1968) 3rd ed.

[3] Fouché, W.L.: Descriptive Complexity and Reflective Properties of Combinatorial Configurations, *Journal of the London Mathematical Society* 54 (1996) 199-208

[4] Fouché, W.L.: Identifying randomness given by high descriptive complexity, *Acta Applicandae Mathematica* 34 (1994) 313-328

[5] Fouché, W.L.and Potgieter, P.H.: Kolmogorov Complexity and Symmetric Relational Structures *The Journal of Symbolic Logic* to appear

[6] Fouché, W.L.: Discrepancies of hypergraphs of high Kolmogorov complexity, (this journal, this edition)

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

Department of Mathematics, Applied Mathematics and Astronomy, Unisa, PO Box 293 Pretoria, South-Africa

E-mail address: davieg@alpha.unisa.ac.za