Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Del azar con dos símbolos al azar con tres símbolos

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Ariel Zylber (azylber@dc.uba.ar)

Directora: Verónica Becher

Buenos Aires, Noviembre 23, 2017

DEL AZAR CON DOS SÍMBOLOS AL AZAR CON TRES SÍMBOLOS

En 1909 Borel definió normalidad como una noción de aleatoriedad de los dígitos de la representación de un número real en cierta base (expansión fraccionaria). Si pensamos la representación de un número en una base dada como una secuencia infinita de símbolos de un alfabeto finito A, se puede dar la definición de normalidad directamente para secuencias de símbolos de A: Una secuencia x es normal para el alfabeto A si cualquier bloque finito de símbolos de A aparece con igual frecuencia asintótica en x que cualquier otro bloque de la misma longitud. Se encontraron muchos ejemplos de secuencias normales siendo Champernowne en 1933 el primero en conseguir dar explícitamente un ejemplo sencillo. También se logró caracterizar cómo seleccionar subsecuencias de una secuencia normal x preservando su normalidad, siempre dejando el alfabeto A fijo. En este trabajo consideramos el problema dual que consiste insertar símbolos en infinitas posiciones de una secuencia dada, de manera tal de preservar la normalidad. Específicamente, dado un símbolo s que no está en el alfabeto original A y dada una secuencia x normal para el alfabeto A, resolvemos el problema de cómo insertar el símbolo s en infinitas posiciones de la secuencia x de modo tal que la secuencia resultante sea normal para el alfabeto extendido $A \cup \{s\}$.

Palabras claves: normalidad, secuencias, azar, inserción, símbolos, alfabeto.

FROM RANDOMNESS IN TWO SYMBOLS TO RANDOMNESS IN THREE SYMBOLS

In 1909 Borel defined normality as a notion of randomness of the digits of the representation of a real number over certain base (fractional expansion). If we think the representation of a number over a base as an infinite sequence of symbols from a finite alphabet A, we can define normality directly for words of symbols of A: A word x is normal to the alphabet A if every finite block of symbols from A appears with the same asymptotic frequency in x as every other block of the same length. Many examples of normal words have been found since its definition, being Champernowne in 1933 the first to show an explicit and simple instance. Moreover, it has been characterized how we can select subsequences of a normal word x preserving its normality, always leaving the alphabet A fixed. In this work we consider the dual problem which consists of inserting symbols in infinite positions of a given word, in such a way that normality is preserved. Specifically, given a symbol s that is not present on the original alphabet A and given a word x that is normal to the alphabet A we solve how to insert the symbol s in infinite positions of the word x such that the resulting word is normal to the expanded alphabet $A \cup \{s\}$.

Keywords: normality, words, randomness, insertion, symbols, alphabet.

CONTENTS

1. INTRODUCTION AND STATEMENT OF RESULTS

In 1909, Borel [3] defined normality as a notion of randomness of the digits of the fractional expansion of a real number over some base. Since then many examples of normal words have been found, Champernowne[5] in 1933 was the first to show an explicit and simple instance,

$$12345678910111213141516171819202122232425262728\ldots$$

the concatenation of all the natural numbers in the natural order is a normal word for the alphabet $A = \{0, 1, \ldots, 9\}$. Moreover, it has been characterized how we can select subsequences of a normal word x preserving its normality, always leaving fixed the alphabet A, see [1, 7, 9].

In this work we consider how normality of words is affected when we add new symbols to the alphabet. Clearly, if a word x is normal to a given alphabet A it is not normal to an alphabet A' that results from adding a new symbol to A, because the word x contains no appearances of this new symbol. A natural question that comes up is if it is possible to insert occurrences of this new symbol along the word x to make it normal in the expanded alphabet. We give a positive answer of this question in Theorem 1.

Fix an alphabet A and a new symbol s. For any given normal word x in A^ω the proof of Theorem 1 gives a way of inserting occurrences of the new symbol s along the word x that depends on the speed of convergence of normality of the word x. The proof is purely combinatorial and it is completely elementary except for the use of the characterization of normality given by Piatetski-Shapiro [8, 4] also known as the Hot Spot Lemma.

The main idea in the proof of Theorem 1 is to use a Champernowne-like word in the expanded alphabet as a reference for insertion of the new symbol s in the given normal word x. We call the discrepancy of a finite word w with respect to the length ℓ to the maximum difference between the expected frequency and the actual frequency in w of any block of ℓ digits. The key ingredient of the proof of Theorem 1 is given by Lemma 9 where we prove that if the discrepancy of a finite word w in the original alphabet with respect to a given length is low enough then inserting occurrences of the new symbol in w according to the pattern of a Champernowne-like word yields an expanded word with also low discrepancy but now with respect to an exponentially shorter length. The proof of this lemma relies on bounding the number of occurrences of a word in the expanded word. In the proof of Theorem 1 we take consecutive segments of the original word x, of increasing length, and expand each of them according to the pattern of digits given by a Champernowne-like word. The difficulty here is in determining the appropriate lengths of these segments. They have to be long enough so that their discrepancy catches up with the discrepancy of the Champernowne-like word. At last, an application of Piatetski-Shapiro's characterization of normality allows us to conclude the normality of the expanded word.

1.1 Primary definitions

We call an alphabet to a finite set A of symbols. Given an alphabet A, we write A^k for the set of all words of length k, A^* for the set of all finite words and A^ω for the set of all infinite words of A. Therefore, $(A^k)^*$ denotes the set of all finite words composed of

the words of length k of A as symbols, or equivalently, the set of all finite words of length multiple of k.

The length of a finite word v is denoted by $|v|$. Given two words u and v with u finite, we denote uv to the word resulting of concatenating u and v. The position of symbols in words are numbered starting from 1. For a word v, we denote $v[i,j]$ as the substring of v from position i to position j. We call $v[i]$ to the symbol corresponding to the i-th position of v. We call substring of a word v to a word of the form $v[i,j]$ for some $i,j \in \mathbb{N}$ such that $1 \leq i \leq j \leq |v|$ and subsequence of v to a word of the form $v[i_1]v[i_2]\dots v[i_k]$ for some $i_1, i_2, \dots, i_k \in \mathbb{N}$ with $i_1 < i_2 < \dots < i_k \leq |v|$.

Given some alphabet A and $u, v \in A^*$, we write

$$\|u\|_v = |\{i \leq |u| - |v| + 1 : u[i, i+|v|-1] = v \text{ and } i \equiv 1 \mod |v|\}|$$

for the number of aligned occurrences of v in u. Thus, if we split the word u in consecutive strings of length $|v|$ and possibly a shorter last string, $\|u\|_v$ is the number of those strings that coincide with v.

With this notation we can state the formulation of normality that is most convenient for to solve our problem. A thorough presentation of normality can be read from the monographs [4, 2].

Definition (Normality to a given alphabet)**.** *Given an alphabet A and some word $u \in A^\omega$, we say that u is simply normal to length ℓ if every $v \in A^\ell$ verifies that*

$$\lim_{n\to\infty} \frac{\|u[1,\ell n]\|_v}{n} = \frac{1}{|A|^\ell}.$$

We say that u is normal if it is simply normal to every length $\ell \in \mathbb{N}$.

Given an alphabet A and a word $v \in A^* \cup A^\omega$, we write $v \upharpoonright n$ to denote $v[1,n]$ which is the word consisting of the first n symbols of v, and we write $v \upharpoonleft n$ to denote the word that results from removing the first n symbols of v.
From now on, we fix a base b and we define $A = \{0, 1, \dots, b-1\}$ and $\widehat{A} = \{0, 1, \dots, b\}$, the alphabets whose symbols are the digits in base b and base $b+1$ respectively.

Definition (reduction operator)**.** *We define the reduction operator $r : \widehat{A}^* \to A^*$ as the operator that removes the symbols b from a word in $\widehat{A}^*$. Precisely, given a word $v \in \widehat{A}^*$,*

$$v = v_1 v_2 \dots v_k$$

where v_i is the i-th symbol of v, then

$$r(v) = v_{r_1} v_{r_2} \dots v_{r_t}$$

where

$$t = |v| - \|v\|_b$$

and

$$r_i = \min(\{j \in \mathbb{N} : |v \upharpoonright j| - \|v \upharpoonright j\|_b = i\}).$$

We define the reduction operator r on infinite words $v \in \widehat{A}^\omega$ in a similar way.

1.2 The main theorem

Theorem 1. *Let $v \in A^\omega$ be a normal word then there exists some normal word $\widehat{v} \in \widehat{A}^\omega$ such that $r(\widehat{v}) = v$.*

Before giving the proof of Theorem 1 we need some intermediate results.

2. TOOLS AND LEMMAS

We define here in a precise way how we expand a word according to the pattern of a Champernowne-like word.

Definition (Champernowne-like words)**.** *For each $n \in \mathbb{N}$, let w_n be the word consisting of the concatenation in lexicographical order of all the words of $\widehat{A}^n$.*

Thus, for $\widehat{A} = \{0, 1\}$, $w_3 = 000001010011100101110111$.

Definition (The wildcard operator)**.** *Let $B = \{b, \star\}$ the alphabet consisting of only the symbols $'b'$ and $'\star'$. We define the wildcard operator $(\star) : \widehat{A}^* \to B^*$ as the operator that given $v \in \widehat{A}^*$ replaces all its symbols different from $'b'$ with a wildcard $'\star'$. Formally, if*

$$v = v_1 v_2 \dots v_k$$

where v_i is the i-th symbol from v, then,

$$(\star)(v) = v_1^\star v_2^\star \dots v_k^\star$$

where

$$v_i^\star = \begin{cases} b, & \text{if } v_i = b \\ \star, & \text{otherwise} \end{cases}$$

We write $v^\star = (\star)(v)$.

It follows easily that if $u, v \in \widehat{A}^*$ then $(uv)^\star = u^\star v^\star$.

Definition (The expansion of order n of a given word)**.** *For each $n \in \mathbb{N}$ we let*

$$\ell_n = \|w_n^\star\|_\star$$
$$\widehat{\ell}_n = |w_n^\star|.$$

For each $i \in \mathbb{N}$ such that $1 \leq i \leq \widehat{\ell}_n$ define

$$m(n, i) = |\{j \leq i : (w_n^\star)_j = \star\}| = \|w_n^\star \restriction i\|_\star.$$

Thus, $m(n, i)$ counts the number of wildcards in w_n up to the i-th symbol.

The expansion $e_n : A^{\ell_n} \to \widehat{A}^{\widehat{\ell}_n}$ is such that, if

$$v = v_1 v_2 \dots v_{\ell_n}$$

then

$$\widehat{v} = \widehat{v}_1 \widehat{v}_2 \dots \widehat{v}_{\widehat{\ell}_n}$$

where

$$\widehat{v}_i = \begin{cases} b, & \text{if } (w_n)_i = b \\ v_{m(n,i)}, & \text{otherwise.} \end{cases}$$

Thus, given a word $v \in A^{\ell_n}$, the expanded word $e_n(v)$ is obtained as follows: take w_n, replace all its symbols different from b by a wildcard symbol, and then replace in each wildcard symbol with the symbols of v in order. Clearly, v is a subsequence of $e_n(v)$ and the only digits that are not part of that subsequence are all b's.

We can extend e_n to $(A^{\ell_n})^*$ by concatenating the expansion of each block of ℓ_n digits. Namely, if $v \in (A^{\ell_n})^*$ such that

$$v = v_1 v_2 \dots v_k$$

where $|v_i| = \ell_n$ for all $1 \leq i \leq k$, then

$$e_n(v) = e_n(v_1)e_n(v_2)\dots e_n(v_k).$$

Clearly, the reduction r is a retraction of e_n for all $n \in \mathbb{N}$, that is,

$$r \circ e_n = id.$$

The next observations follow from the definitions.

Observation 2. *$\widehat{\ell}_n = n(b+1)^n$ and $\ell_n = nb(b+1)^{n-1}$ for all $n \in \mathbb{N}$.*

Proof. Since there are $(b+1)^n$ different words of length n using $b+1$ symbols and each word has length n we get $\widehat{\ell}_n = n(b+1)^n$. Since each symbol appears the same number of times in w_n then $\|w_n\|_b = n(b+1)^{n-1}$. It follows that

$$\ell_n = |w_n| - \|w_n\|_b = n(b+1)^n - n(b+1)^{n-1} = nb(b+1)^{n-1}.$$

□

We denote $\mathbb{1}$ to the indicator function of the diagonal elements of $A^* \times A^*$. Namely, we define $\mathbb{1} : A^* \times A^* \to \mathbb{N}$ as

$$\mathbb{1}(x,y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$

We denote $\mathbb{1}(x,y)$ as $\mathbb{1}(x=y)$.

Observation 3. *Given an alphabet C with $|C| = k$, some $v \in C^n$, some $m \in \mathbb{N}$ such that $m > n$ and some $i \in \mathbb{N}$ such that $0 \leq i \leq m-n$, then*

$$\sum_{u \in A^m} \mathbb{1}(u[i+1, i+n] = v) = k^{m-n}.$$

Observation 4. *Given an alphabet C with $|C| = k$, some $v \in C^n$ and $u \in (C^n)^*$ then*

$$\|u\|_v = \sum_{i=0}^{|u|/n-1} \mathbb{1}(u[in+1, in+n] = v).$$

Observation 5. *Given $v, w \in \widehat{A}^*$ then $v = w$ if and only if $v^\star = w^\star$ and $r(v) = r(w)$.*

Observation 6. *If $v \in B^n$ then $\|w_n^\star\|_v = b^{\|v\|_\star}$.*

Observation 7. *If $v \in A^{\ell_n}$ and $w \in (A^{\ell_n})^*$ then $\|w\|_v = \|e_n(w)\|_{e_n(v)}$.*

Lemma 8. *Given $w \in \widehat{A}^n$ then*

$$\sum_{u \in A^{\ell_n}} \|e_n(u)\|_w = b^{\ell_n}.$$

Proof. By Observation 4 we have

$$\|e_n(u)\|_w = \sum_{i=0}^{|u|/n-1} \mathbb{1}(e_n(u)[in+1, in+n] = w)$$

for all $u \in A^{\ell_n}$. Applying Observation 5 we get $\|e_n(u)\|_w$ is equal to

$$\sum_{i=0}^{|u|/n-1} \mathbb{1}\Big((e_n(u)[in+1, in+n])^\star = w^\star\Big)\mathbb{1}\Big(r(e_n(u)[in+1, in+n]) = r(w)\Big). \tag{2.0.1}$$

Analyzing the definition of $(\star)$ we get that

$$(e_n(u)[in+1, in+n])^\star = e_n(u)^\star[in+1, in+n] = (w_n^\star)[in+1, in+n].$$

By Observation 6 we conclude that

$$\|w_n^\star\|_{w^\star} = b^{\|w^\star\|_\star} = b^{|w|-\|w\|_b}.$$

This means that there are exactly $b^{|w|-\|w\|_b}$ terms of the sum in which

$$\mathbb{1}\Big((e_n(u)[in+1, in+n])^\star = w^\star\Big) = 1.$$

Let

$$I = \{0 \le i < \ell_n/n : \mathbb{1}((w_n^\star)[in+1, in+n] = w^\star) = 1\}.$$

be the set of indexes where the first term of the product of equation 2.0.1 does not vanish. Notice that I does not depend on u.

Analyzing the second term of the product of equation 2.0.1, we observe that

$$\mathbb{1}\Big(r(e_n(u)[in+1, in+n]\Big) = r(w)) = \mathbb{1}\Big(u[m(n, in)+1, m(n, in+n)] = r(w)\Big).$$

applying this we reduce (2.0.1) to

$$\|e_n(u)\|_w = \sum_{i \in I} \mathbb{1}\Big(u[m(n, in)+1, m(n, in+n)] = r(w)\Big). \tag{2.0.2}$$

For $i \in I$ we have that $(e_n(u)[in+1, in+n])^\star = w^\star$, which implies that

$$|u[m(n, in)+1, m(n, in+n)]| = |r(e_n(u)[in+1, in+n])| = |r(w)|.$$

Summing (2.0.2) over all $u \in A^{\ell_n}$ we get

$$\sum_{u \in A^{\ell_n}} \|e_n(u)\|_w = \sum_{u \in A^{\ell_n}} \sum_{i \in I} \mathbb{1}(u[m(n, in)+1, m(n, in+n)] = r(w)).$$

And applying Observation 3 we get

$$\sum_{u \in A^{\ell_n}} \|e_n(u)\|_w = \sum_{i \in I} b^{\ell_n - |r(w)|} = b^{|w|-\|w\|_b} b^{\ell_n - |r(w)|}.$$

And noticing that by definition of r we have that $|r(w)| = |w| - \|w\|_b$ this gives us the desired result

$$\sum_{u \in A^{\ell_n}} \|e_n(u)\|_w = b^{\ell_n}.$$

□

2.1 On discrepancies

Here we introduce a definition of discrepancy for finite words and we relate the discrepancy of a word and the discrepancy of the expanded word. We also consider the concatenation of a sequence of words and we bound the discrepancy of the resulting word in terms of the discrepancies of the individual words. Most of the bounds that we give can be improved but these simple versions will be enough for the proof of Theorem 1.

Given some word $u \in A^*$ and a fixed length $\ell \in \mathbb{N}$, for a word $v \in A^\ell$ the frequency of aligned occurrences of v in u over all aligned substrings of length ℓ in u is

$$\frac{\|u\|_v}{\lfloor |u|/\ell \rfloor}.$$

We can measure how far is this frequency from the case where all words of length ℓ are equiprobable by

$$\left| \frac{\|u\|_v}{\lfloor |u|/\ell \rfloor} - \frac{1}{|A|^\ell} \right|.$$

The discrepancy of a word u in A^* for a length ℓ is the maximum of this distance among all $v \in A^\ell$ and we denote it by $\Delta_{A,\ell}(u)$.

Definition (Discrepancy of a finite word for a given length ℓ)**.**

$$\Delta_{A,\ell}(u) = \max_{v \in A^\ell} \left(\left| \frac{\|u\|_v}{\lfloor |u|/\ell \rfloor} - \frac{1}{|A|^\ell} \right| \right).$$

An easy equivalence is that u is simply normal to length ℓ if and only if

$$\lim_{n \to \infty} \Delta_{A,\ell}(u[1,n]) = 0.$$

and therefore u is normal if and only if this limit is valid for every length $\ell \in \mathbb{N}$.

Let $u \in A^*$, let ℓ be a length and let ε be a real umber between 0 and 1. Then it follows that

$$\Delta_{A,\ell}(u) < \varepsilon$$

is equivalent to have for all $v \in A^\ell$,

$$\lfloor |u|/\ell \rfloor \left(\frac{1}{|A|^\ell} - \varepsilon \right) < \|u\|_v < \lfloor |u|/\ell \rfloor \left(\frac{1}{|A|^\ell} + \varepsilon \right).$$

Lemma 9 (Main Lemma)**.** *For each $n \in \mathbb{N}$ there exists a constant $c_n \in \mathbb{R}$ with $c_n > 0$ such that for every $\varepsilon > 0$ and every word $v \in (A^{\ell_n})^*$ if*

$$\Delta_{A,\ell_n}(v) < \varepsilon \tag{2.1.1}$$

then

$$\Delta_{\widehat{A},n}(e_n(v)) < c_n \varepsilon.$$

Proof. Let $w \in \widehat{A}^n$ be any word, then

$$\|e_n(v)\|_w = \sum_{\widehat{u} \in \widehat{A}^{\widehat{\ell_n}}} \|e_n(v)\|_{\widehat{u}} \|\widehat{u}\|_w.$$

By the definition of e_n, the blocks of length $\widehat{\ell}_n$ of $e_n(v)$ are of the form $e_n(v_i)$ for some $v_i \in A^{\ell_n}$. Then, the only non-zero terms of the sum can be the ones where $\widehat{u}$ is in the image of e_n, and since e_n is injective we can change the sum to iterate over the $e_n(u)$ for $u \in A^{\ell_n}$. It follows that

$$\|e_n(v)\|_w = \sum_{u \in A^{\ell_n}} \|e_n(v)\|_{e_n(u)} \|e_n(u)\|_w.$$

By Observation 7 it reduces to

$$\|e_n(v)\|_w = \sum_{u \in A^{\ell_n}} \|v\|_u \|e_n(u)\|_w.$$

Applying (2.1.1) we get

$$\|e_n(v)\|_w < \sum_{u \in A^{\ell_n}} \frac{|v|}{|u|} \left(\frac{1}{b^{|u|}} + \varepsilon \right) \|e_n(u)\|_w = \frac{|v|}{\ell_n} \left(\frac{1}{b^{\ell_n}} + \varepsilon \right) \left(\sum_{u \in A^{\ell_n}} \|e_n(u)\|_w \right).$$

Using Observation 8 we get

$$\|e_n(v)\|_w < \frac{|v|}{\ell_n} \left(\frac{1}{b^{\ell_n}} + \varepsilon \right) b^{\ell_n} = \frac{|v|}{\ell_n} \left(1 + b^{\ell_n} \varepsilon \right).$$

Multiplying by $\frac{|w|}{|e_n(v)|} = \frac{n}{|e_n(v)|}$ on both sides we obtain

$$\frac{|w|}{|e_n(v)|} \|e_n(v)\|_w < \frac{n|v|}{\ell_n |e_n(v)|} \left(1 + b^{\ell_n} \varepsilon \right). \tag{2.1.2}$$

Since $v \in (A^{\ell_n})^*$ we can write v as

$$v = v_1 v_2 \dots v_t$$

where each v_i satisfies $|v_i| = \ell_n$. Then $|v| = t\ell_n$ and

$$e_n(v) = e_n(v_1) e_n(v_2) \dots e_n(v_t)$$

where $|e_n(v_i)| = \widehat{\ell}_n$ for all $1 \le i \le t$. So, we conclude that $|e_n(v)| = t\widehat{\ell}_n$.

Using this on (2.1.2) we get

$$\frac{|w|}{|e_n(v)|} \|e_n(v)\|_w < \frac{nt\ell_n}{\ell_n t \widehat{\ell}_n} \left(1 + b^{\ell_n} \varepsilon \right)$$

using Observation 2 we can replace the value of $\widehat{\ell}_n$ and get

$$\frac{|w|}{|e_n(v)|} \|e_n(v)\|_w < \frac{n}{n(b+1)^n} \left(1 + b^{\ell_n} \varepsilon \right) = \frac{1}{(b+1)^n} + \frac{b^{\ell_n}}{(b+1)^n} \varepsilon.$$

By a similar argument we get the inequality

$$\frac{|w|}{|e_n(v)|} \|e_n(v)\|_w > \frac{1}{(b+1)^n} - \frac{b^{\ell_n}}{(b+1)^n} \varepsilon.$$

These two inequalities imply that

$$\Delta_{\widehat{A},n}(e_n(v)) < \frac{b^{\ell_n}}{(b+1)^n} \varepsilon.$$

The desired result follows taking

$$c_n = \frac{b^{\ell_n}}{(b+1)^n}.$$

□

2.2 Some other useful results

Proposition 10. *Given a finite alphabet C with $|C| = k$ and some $m, n \in \mathbb{N}$. We have that for each word $v \in (C^{mn})^*$ and $\varepsilon \in \mathbb{R}$ with $\varepsilon > 0$ such that*

$$\Delta_{C,mn}(v) < \varepsilon \tag{2.2.1}$$

then

$$\Delta_{C,n}(v) < k^{(m-1)n}\varepsilon.$$

Proof. Let $w \in C^n$. We have that

$$\|v\|_w = \sum_{u \in C^{mn}} \|v\|_u \|u\|_w.$$

Using 2.2.1 we get

$$\|v\|_w < \sum_{u \in C^{mn}} \frac{|v|}{|u|} \left(\frac{1}{k^{mn}} + \varepsilon \right) \|u\|_w.$$

Using Observation 4 we get

$$\|v\|_w < \frac{|v|}{mn} \left(\frac{1}{k^{mn}} + \varepsilon \right) \sum_{u \in C^{mn}} \sum_{i=0}^{m} \mathbb{1}(u[in+1, in+n] = v).$$

Using Observation 3 we get

$$\|v\|_w < \frac{|v|}{mn} \left(\frac{1}{k^{mn}} + \varepsilon \right) \sum_{i=0}^{m} k^{mn-n} = \frac{|v|}{n} \left(\frac{1}{k^n} + k^{(m-1)n}\varepsilon \right).$$

□

Proposition 11. *Given a finite alphabet C, some $n \in \mathbb{N}$ and $u, v \in (C^n)^*$, if*

$$\Delta_{C,n}(u) < \varepsilon \tag{2.2.2}$$

and

$$\Delta_{C,n}(uv) < \varepsilon \tag{2.2.3}$$

then

$$\Delta_{C,n}(v) < \frac{|uv| + |u|}{|v|}\varepsilon.$$

Proof. Let $w \in C^n$. Then,

$$\|v\|_w = \|uv\|_w - \|u\|_w.$$

Using (2.2.2) and (2.2.3) we get

$$\|v\|_w < \frac{|uv|}{|w|} \left(\frac{1}{k^{|w|}} + \varepsilon \right) - \frac{|u|}{|w|} \left(\frac{1}{k^{|w|}} - \varepsilon \right)$$

which using $|uv| = |u| + |v|$ is equivalent to

$$\|v\|_w < \frac{|v|}{|w|} \frac{1}{k^{|w|}} + \frac{|uv| + |u|}{|w|}\varepsilon$$

which is equivalent to

$$\|v\|_w < \frac{|v|}{|w|}\left(\frac{1}{k^{|w|}} + \frac{|uv|+|u|}{|v|}\varepsilon\right).$$

In a similar way we can conclude

$$\|v\|_w > \frac{|v|}{|w|}\left(\frac{1}{k^{|w|}} - \frac{|uv|+|u|}{|v|}\varepsilon\right).$$

Since both inequalities are valid for all $w \in C^n$ we conclude the result. □

Proposition 12. *Given a finite alphabet C, some $n \in \mathbb{N}$ and $u, v \in (C^n)^*$, if*

$$\Delta_{C,n}(u) < \varepsilon \tag{2.2.4}$$

and

$$\Delta_{C,n}(v) < \frac{|uv|+|u|}{|v|}\varepsilon \tag{2.2.5}$$

then

$$\Delta_{C,n}(uv) < 3\varepsilon.$$

Proof. Let $w \in C^n$. Then,

$$\|uv\|_w = \|u\|_w + \|v\|_w.$$

Using (2.2.4) and (2.2.5) we get

$$\|uv\|_w < \frac{|u|}{|w|}\left(\frac{1}{k^{|w|}} + \varepsilon\right) + \frac{|v|}{|w|}\left(\frac{1}{k^{|w|}} + \frac{|uv|+|u|}{|v|}\varepsilon\right)$$

which using $|uv| = |u| + |v|$ is equivalent to

$$\|uv\|_w < \frac{|u|+|v|}{|w|}\frac{1}{k^{|w|}} + \frac{3|u|+|v|}{|w|}\varepsilon,$$

and since $3|u| + |v| < 3(|u| + |v|)$ we get

$$\|uv\|_w < \frac{|u|+|v|}{|w|}\left(\frac{1}{k^{|w|}} + 3\varepsilon\right).$$

In a similar way we can conclude

$$\|uv\|_w > \frac{|u|+|v|}{|w|}\left(\frac{1}{k^{|w|}} - 3\varepsilon\right).$$

Since both inequalities are valid for all $w \in C^n$ we conclude the result. □

Our analysis so far focuses in aligned occurrences of a given word in an expanded word. For a technical reason the proof of Theorem 1 needs to consider the number of occurrences of any given word in the constructed expanded word. We define the number of non-aligned occurrences of a word v in a word u as

$$|u|_v = |\{i \leq |u| - |v| + 1 : u[i, i + |v| - 1] = v\}|$$

Notice that for every symbol $b \in A$ and for every word $u \in A^*$,

$$|u|_b = \|u\|_b.$$

The following proposition gives the needed result.

Proposition 13. *Given a finite alphabet C, some $n, m \in \mathbb{N}$ with $m < n$ some $u \in (C^n)^*$ and $v \in C^m$, if*

$$\Delta_{C,n}(u) < \varepsilon \tag{2.2.6}$$

then

$$|u|_v < |u| \left(\frac{m-1}{n} + \frac{1}{|C|^m} + |C|^n \varepsilon \right) - (m-1).$$

Proof. For every pair of consecutive blocks of length n in u there are exactly $m-1$ substrings of length m that are not fully contained in one of these blocks. Since there are $|u|/n$ blocks of length n in u, there are $(|u|/n-1)(m-1)$ substrings of length m not fully contained in one of the blocks. This gives us the following bound on the number of occurrences of v in u:

$$\begin{aligned} |u|_v &\leq (|u|/n-1)(m-1) + \sum_{i=0}^{|u|/n-1} |u[in+1, in+n]|_v \\ &= (|u|/n-1)(m-1) + \sum_{w \in C^n} \|u\|_w |w|_v. \end{aligned}$$

Using (2.2.6) we get,

$$|u|_v < (|u|/n-1)(m-1) + \sum_{w \in C^n} \frac{|u|}{|v|} \left(\frac{1}{|C|^n} + \varepsilon \right) |w|_v.$$

Using that $|w|_v = \sum_{i=1}^{|w|-|v|} \mathbb{1}(w[i, i+|v|] = v)$ we get,

$$|u|_v < (|u|/n-1)(m-1) + \frac{|u|}{|v|} \left(\frac{1}{|C|^n} + \varepsilon \right) \sum_{w \in C^n} \sum_{i=1}^{n-m} \mathbb{1}(w[i, i+|v|] = v).$$

Using Observation 3 we get,

$$|u|_v < (|u|/n-1)(m-1) + \frac{|u|}{|v|} \left(\frac{1}{|C|^n} + \varepsilon \right) \sum_{i=1}^{n-m} |C|^{n-m}.$$

Which is equivalent to

$$|u|_v < (|u|/n-1)(m-1) + \frac{|u|}{n} \left(\frac{1}{|C|^m} + |C|^{n-m} \varepsilon \right) (n-m).$$

And since $m < n$ we get,

$$|u|_v < |u| \left(\frac{m-1}{n} + \frac{1}{|C|^m} + |C|^n \varepsilon \right) - (m-1),$$

as desired. □

The first paragraph in the proof above yields the following result.

Observation 14. *Given a finite alphabet C, some $u, v, w \in C^*$ then*

$$|uv|_w \leq |u|_w + |v|_w + |w| - 1.$$

Finally we introduce a characterization of normality that is seemingly easier than the actual definition, because instead of asking for the limit it asks for the limsup.

Lemma 15 (Hot Spot Lemma, Piatetski-Shapiro 1951)**.** *Let $x \in A^\omega$. Then, x is normal if and only if there is positive constant C such that for all lengths ℓ and for every word $u \in A^\ell$,*

$$\limsup_{n\to\infty} \frac{|x[1,n]|_u}{n} < \frac{C}{|A|^\ell}.$$

3. PROOF OF THEOREM 1

We construct inductively a sequence of nonempty finite substrings $\{v_i\}_{i\in\mathbb{N}}$ of v that verifies that $v_1v_2\dots v_k$ is a prefix of v for all k in $\mathbb{N}$. Suppose that we have already defined $v_1, v_2, \dots, v_{n-1}$ and we want to define v_n. Let $L_{n-1} = |v_1v_2\dots v_{n-1}|$ be the total length of all substrings already defined. Since v is normal, then $v \upharpoonright L_{n-1}$ is also normal and consequently given

$$\varepsilon_n = \frac{1}{(b+1)^{2^n}n}\frac{1}{3\max(b^n c_{2^n}, (b+1)^n c_{2^{n+1}})}$$

where c_{2^n} and $c_{2^{n+1}}$ are the constants from Lemma 9, there exists a k_n such that for all $k > k_n$ in $\mathbb{N}$ we have

$$\Delta_{A,\ell_{2^{n+1}}}(v[L_{n-1}+1, L_{n-1}+k]) < \varepsilon_n$$

Take t_n such that $t_n\ell_{2^n} > \max(k_n, \ell_{2^{n+1}})$ and define v_n as

$$v_n = v[L_{n-1}+1, L_{n-1}+t_n\ell_{2^n}]$$

It is clear that $v_1v_2\dots v_n = v[1, L_{n-1}+t_n\ell_{2^n}]$ and thus is a prefix of v.

Given $\{v_i\}_{i\in\mathbb{N}}$ defined as above, we define the expansion $\widehat{v}$ as

$$\widehat{v} = e_{2^1}(v_1)e_{2^2}(v_2)\dots e_{2^i}(v_i)\dots$$

Since each v_i has length $t_i\ell_{2^i}$ which is multiple of ℓ_{2^i}, the expansion is well defined. It follows easily that

$$r(\widehat{v}) = r(e_{2^1}(v_1))r(e_{2^2}(v_2))\dots r(e_{2^i}(v_i))\dots = v.$$

We claim that $\widehat{v}$ is normal in base $b+1$. We can write each v_n as

$$v_n = v_{n,1}v_{n,2}\dots v_{n,t_n}$$

where each $v_{n,i}$ satisfies $|v_{n,i}| = \ell_{2^n}$. Fix $n \in \mathbb{N}$ and $j \in \mathbb{N}_0$ with $0 \le j \le t_{n+1}$, and define

$$v'_{n+1} = v_{n+1,1}v_{n+1,2}\dots v_{n+1,j}$$

as the prefix of v_{n+1} that consists of the first j blocks of length $\ell_{2^{n+1}}$. By definition of v_n, we have that

$$\Delta_{A,\ell_{2^{n+1}}}(v_n) < \varepsilon_n \tag{3.0.1}$$

and since $v_nv'_{n+1}$ is a prefix of $v \upharpoonright L_{n-1}$ of length greater than k_n we have

$$\Delta_{A,\ell_{2^{n+1}}}(v_nv'_{n+1}) < \varepsilon_n.$$

Using Proposition 11 we have that

$$\Delta_{A,\ell_{2^{n+1}}}(v'_{n+1}) < \frac{|v_nv'_{n+1}| + |v_n|}{|v'_{n+1}|}\varepsilon_n. \tag{3.0.2}$$

Now, by (3.0.1) and Proposition 10 we have that

$$\Delta_{A,\ell_{2^n}}(v_n) < b^{\ell_{2^n}}\varepsilon_n$$

and applying Lemma 9 we get

$$\Delta_{\widehat{A},2^n}(e_{2^n}(v_n)) < b^{\ell_{2^n}}c_{2^n}\varepsilon_n \tag{3.0.3}$$

Similarly, applying Lemma 9 to (3.0.2) we get

$$\Delta_{\widehat{A},2^{n+1}}(e_{2^{n+1}}(v'_{n+1})) < \frac{|v_nv'_{n+1}|+|v_n|}{|v'_{n+1}|}c_{2^{n+1}}\varepsilon_n$$

and by Proposition 10 we conclude

$$\Delta_{\widehat{A},2^n}(e_{2^{n+1}}(v'_{n+1})) < \frac{|v_nv'_{n+1}|+|v_n|}{|v'_{n+1}|}(b+1)^{2^n}c_{2^{n+1}}\varepsilon_n. \tag{3.0.4}$$

Using Proposition 12 with (3.0.3) and (3.0.4) we get that

$$\Delta_{\widehat{A},2^n}(e_{2^n}(v_n)e_{2^{n+1}}(v'_{n+1})) < 3\max(b^{\ell_{2^n}}c_{2^n},(b+1)^{2^n}c_{2^{n+1}})\varepsilon_n < \frac{1}{(b+1)^{2^n}n}. \tag{3.0.5}$$

Notice that the bound does not depend on j. If $j=0$ we get the special case

$$\Delta_{\widehat{A},2^n}(e_{2^n}(v_n)) < \frac{1}{(b+1)^{2^n}n}. \tag{3.0.6}$$

Now, we fix $u \in \widehat{A}^m$ for some $m \in \mathbb{N}$. For $n \in \mathbb{N}$ and $j \in \mathbb{N}_0$ with $0 \le j \le t_n$, we define

$$L_{n,j} = |v_1v_2\dots v_nv_{n+1,1}v_{n+1,2}\dots v_{n+1,j}|.$$

Notice that $L_{n,t_n} = L_{n+1,0}$. We define $L_{0,0}=0$. Given some $M \in \mathbb{N}$ with $M > L_{1,0}$, there exists some $n,j \in \mathbb{N}$ with $n>1$ such that

$$L_{n,j-1} \le M \le L_{n,j}. \tag{3.0.7}$$

By Observation 14 we get

$$|\widehat{v}[1,M]|_u \le |\widehat{v}[1,L_{n,j}]|_u \le$$
$$\left(\sum_{i=1}^{n-1}|e_{2^i}(v_i)|_u\right) + |e_{2^n}(v_n)e_{2^{n+1}}(v_{n+1,1})\dots e_{2^{n+1}}(v_{n+1,j})|_u + (n-1)(|u|-1). \tag{3.0.8}$$

Given that we have (3.0.6) for each term of the sum, we can apply Proposition 13 and we get the bound

$$\sum_{i=1}^{n-1}|e_{2^i}(v_i)|_u \le \sum_{i=1}^{n-1}|e_{2^i}(v_i)|\left(\frac{|u|-1}{2^i}+\frac{1}{(b+1)^{|u|}}+\frac{(b+1)^{2^i}}{(b+1)^{2^i}i}\right) - (n-1)(|u|-1).$$

Noticing that

$$\frac{|u|-1}{2^i}+\frac{1}{i} \to 0 \text{ as } i \to \infty$$

there exists some i_0 such that for all $i > i_0$ we have

$$\frac{|u|-1}{2^i} + \frac{1}{i} \leq \frac{1}{(b+1)^{|u|}}. \tag{3.0.9}$$

If M is sufficiently large, we will have $n > i_0$ and then we can split the sum and get

$$\begin{aligned}\sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u \leq & \sum_{i=1}^{i_0} |e_{2^i}(v_i)| \left(\frac{|u|-1}{2^i} + \frac{1}{(b+1)^{|u|}} + \frac{1}{i} \right) + \\ & \sum_{i=i_0+1}^{n-1} |e_{2^i}(v_i)| \left(\frac{|u|-1}{2^i} + \frac{1}{(b+1)^{|u|}} + \frac{1}{i} \right) - \\ & (n-1)(|u|-1).\end{aligned} \tag{3.0.10}$$

Calling

$$\delta = \sum_{i=1}^{i_0} |e_{2^i}(v_i)| \left(\frac{|u|-1}{2^i} + \frac{1}{(b+1)^{|u|}} + \frac{1}{i} \right)$$

(notice that δ does not depend on M) and using (3.0.9) we get

$$\sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u \leq \delta + \frac{2}{(b+1)^{|u|}} \left(\sum_{i=i_0+1}^{n-1} |e_{2^i}(v_i)| \right) - (n-1)(|u|-1).$$

Using that $|e_{2^i}(v_i)| = L_{i,0} - L_{i-1,0}$ we can reduce this to

$$\sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u \leq \delta + (L_{n-1,0} - L_{i_0,0}) \frac{2}{(b+1)^{|u|}} - (n-1)(|u|-1). \tag{3.0.11}$$

Having (3.0.5) and using Proposition 13 with the second term of (3.0.8) we get

$$\begin{aligned}&|e_{2^n}(v_n) e_{2^{n+1}}(v_{n+1,1}) \dots e_{2^{n+1}}(v_{n+1,j})|_u \leq \\ &(L_{n,j} - L_{n-1,0}) \left(\frac{|u|-1}{2^n} + \frac{1}{(b+1)^{|u|}} + \frac{(b+1)^{2^n}}{(b+1)^{2^n} n} \right) - (|u|-1).\end{aligned}$$

Since $n > i_0$ we get

$$|e_{2^n}(v_n) e_{2^{n+1}}(v_{n+1,1}) \dots e_{2^{n+1}}(v_{n+1,j})|_u \leq (L_{n,j} - L_{n-1,0}) \left(\frac{2}{(b+1)^{|u|}} \right). \tag{3.0.12}$$

Using (3.0.11) and (3.0.12) in (3.0.8) we get

$$|\widehat{v}[1,M]|_u \leq \delta + (L_{n,j} - L_{i_0,0}) \frac{2}{(b+1)^{|u|}}.$$

Dividing both sides by $|\widehat{v}[1,M]| = M$ we get

$$\frac{|\widehat{v}[1,M]|_u}{M} \leq \frac{\delta}{M} + \frac{L_{n,j} - L_{i_0,0}}{M} \frac{2}{(b+1)^{|u|}}. \tag{3.0.13}$$

By (3.0.7) we have that

$$L_{n,j} - M \leq L_{n,j} - L_{n,j-1} = \widehat{\ell}_{2^{n+1}}.$$

By construction of v_n,

$$\ell_{2^{n+1}} \leq |v_n|.$$

Then, since $e_{2^n}(v_n)$ is a substring of $\widehat{v}[1, M]$ we get that

$$\widehat{\ell}_{2^{n+1}} \leq |e_{2^n}(v_n)| \leq M.$$

Which gives us the bound $L_{n,j} \leq 2M$. Using this in (3.0.13) we get

$$\frac{|\widehat{v}[1, M]|_u}{M} \leq \frac{\delta}{M} + \frac{2M - L_{i_0,0}}{M} \frac{2}{(b+1)^{|u|}} < \frac{\delta}{M} + \frac{4}{(b+1)^{|u|}}. \tag{3.0.14}$$

Taking limit superior as $M \to \infty$ and since δ does not depend on M we get

$$\limsup_{M \to \infty} \frac{|\widehat{v}[1, M]|_u}{M} \leq \frac{4}{(b+1)^{|u|}}. \tag{3.0.15}$$

Since this bound is valid for all $u \in \widehat{A}^*$, using the Lemma 15 (Hot Spot Lemma) with $C = 4$ follows that $\widehat{v}$ is normal. Therefore, we constructed a normal word $\widehat{v}$ such that $r(\widehat{v}) = v$ as desired. This completes the proof of Theorem 1.

4. SOME REMARKS ABOUT THE PROOF OF THEOREM 1

4.1 On the choice of w_n

We can study how flexible is the construction of the proof on the choice of the sequence (w_n). We wonder what other sequences we can choose so that the proof remains valid. Looking at the proof, the only places where we use the explicit construction of (w_n) is in Lemma 8 and Lemma 9. The property of the sequence we are using is that

$$\Delta_{\widehat{A},n} w_n = 0$$

This means that we can change the w_n for some other sequence satisfying this property. In fact, if we only have that the discrepancy of w_n is small, namely

$$\Delta_{\widehat{A},n} w_n < \delta_n$$

we can, with a little more of work, obtain a bound similar to that of to Lemma 9 but also involving the δ_n. Then, we can use this bound in the proof of the Theorem 1. If we choose the δ_n to be small enough (and maybe depending of the ε_n) we can adapt the proof to work for this new sequence (w_n). Any normal word $z \in \widehat{A}^\omega$ can be split it in consecutive strings $z_1, z_2, \ldots, z_n, \ldots$ such that

$$z = z_1 z_2 \ldots z_n \ldots$$

and for each n, the word z_n satisfies

$$\Delta_{\widehat{A},n}(z_n) < \delta_n.$$

If the lengths of z_n do not grow larger than exponential on n, we can use this sequence (z_n) as an alternative for (w_n) to expand normal words. This means that we can do the process to expand a normal word to $\widehat{A}$ with substrings of any normal word in $\widehat{A}$ that has a partition into substrings with this property.

4.2 On the computability of the construction

If we know the convergence rates of the normal word to expand, we can calculate e_n for all $n \in \mathbb{N}$ and we can easily compute the expanded word. If we don't know anything about the convergence rates, we can still compute the expanded word with a finite-injury priority method [6], but we will not know how good will be our approximation at each step of the algorithm.

BIBLIOGRAPHY

[1] V. N. Agafonov. Normal sequences and finite automata. *Soviet Mathematics Doklady*, 9:324–325, 1968.

[2] Verónica Becher and Olivier Carton. Normal numbers and computer science. In Valérie Berthé and Michel Rigó, editors, *Sequences, Groups, and Number Theory*, Trends in Mathematics Series. Birkhauser/Springer, 2017.

[3] Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*, 27(1):247–271, 1909.

[4] Yann Bugeaud. *Distribution modulo one and Diophantine approximation*, volume 193. Cambridge University Press, 2012.

[5] David Champernowne. The Construction of Decimals Normal in the Scale of Ten. *The Journal of the London Mathematical Society*, s1-8(4):254–260, 1933.

[6] Jr. Hartley Rogers. *Theory of recursive functions and effective computability.* MIT Press, Cambridge, MA, second edition, 1987.

[7] Teturo Kamae and Benjamin Weiss. Normal numbers and selection rules. *Israel Journal of Mathematics*, 21(2):101–110, 1975.

[8] I. I. Piatetski-Shapiro. On the law of distribution of the fractional parts of the exponential function. *Izv. Akad. Nauk SSSR Ser. Mat.*, 15(1):47–52, 1951.

[9] Joseph Vandehey. Uncanny subsequence selections that generate normal numbers. *arXiv:1607.03531*, 2016.