

# FROM RANDOMNESS IN TWO SYMBOLS TO RANDOMNESS IN THREE SYMBOLS

ARIEL ZYLBER

Universidad de Buenos Aires, Buenos Aires, ARGENTINA

**ABSTRACT.** 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 of 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 infinitely many positions of a given word, in such a way that normality is preserved. Specifically, given a symbol  $b$  that is not present in the original alphabet  $A$  and given a word  $x$  that is normal to the alphabet  $A$  we solve how to insert the symbol  $b$  in infinitely many positions of the word  $x$  such that the resulting word is normal to the expanded alphabet  $A \cup \{b\}$ .

*Communicated by Friedrich Pillichshammer*

## 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

---

© 2021 BOKU-University of Natural Resources and Life Sciences and Mathematical Institute, Slovak Academy of Sciences.

2010 Mathematics Subject Classification: Primary 11K16; 05A05.

Keywords: Normal numbers, combinatorics on words, Champernowne number.



Licensed under the Creative Commons BY-NC-ND 4.0 International Public License.

examples of normal words have been found. Champernowne [5] in 1933 was the first to show an explicit and simple instance,

$$12345678910111213141516171819202122232425262728 \dots$$

the concatenation of all the natural numbers in the natural order is a normal word for the alphabet  $A = \{0, 1, \dots, 9\}$ . 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 to this question.

We call a finite set of symbols an alphabet. 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  $A^\omega$  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  $k$ .

For an alphabet  $A$ ,  $A^*$  is the free monoid generated by  $A$  with concatenation as its operation. Given two alphabets  $A$  and  $B$ , a function  $f : A \rightarrow B^*$  induces a morphism of monoids  $f^* : A^* \rightarrow B^*$  that extends  $f$  by concatenation. Namely, given a word  $v \in A^*$ ,

$$v = v_1 v_2 \dots v_k \quad \text{with} \quad v_1, v_2, \dots, v_k \in A,$$

then

$$f^*(v) = f(v_1) f(v_2) \dots f(v_k).$$

This function  $f$  can be extended in a similar way to  $A^\omega$ .

Fix an alphabet  $A$  and a new symbol  $b$ . Let  $\widehat{A}$  be  $A \cup \{b\}$ . We write  $\lambda$  for the empty word. We define the reduction function  $r : \widehat{A} \rightarrow (A \cup \{\lambda\})$  such that if  $s \in \widehat{A}$ ,

$$r(s) = \begin{cases} \lambda & \text{if } s = b, \\ s, & \text{otherwise.} \end{cases}$$

The function  $r$  extends to  $\widehat{A}^* \cup \widehat{A}^\omega$  as the morphism that removes the symbol  $b$  from a word.

We can now state the main result of this note.

**THEOREM 1.1.** *Let  $v \in A^\omega$  be a normal word. Then there exists some normal word  $\widehat{v} \in \widehat{A}^\omega$  such that  $r(\widehat{v}) = v$ .*

For any given normal word  $x$  in  $A^\omega$  the proof of Theorem 1.1 gives a way of inserting occurrences of the new symbol  $b$  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 [4,8] also known as the Hot Spot Lemma. The main idea is to use a Champernowne-like word in the expanded alphabet as a reference for insertion of the new symbol  $b$  in the given normal word  $x$ . We define the discrepancy of a finite word  $w$  with respect to the length  $\ell$  as the maximum difference between the expected frequency and the actual frequency in  $w$  of any block of  $\ell$  digits. This is essentially the same concept of discrepancy defined by Besicovitch [10].

The key ingredient of the proof of Theorem 1.1 is given by Lemma 2.13 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 again with 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 1.1 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.

## 2. Tools and lemmas

The length of a finite word  $v$  is denoted by  $|v|$ . Given two words  $u$  and  $v$  with  $u$  finite, we denote by  $uv$  the word resulting by concatenating  $u$  and  $v$ . The position of symbols in words are numbered starting from 1. For a word  $v$ , we denote by  $v[i, j]$  the substring of  $v$  from position  $i$  to position  $j$ . We call  $v[i]$  the symbol corresponding to the  $i$ th position of  $v$ . A substring of a word  $v$  is 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 a subsequence of  $v$  is a word of the form

$$v[i_1]v[i_2] \dots v[i_k] \quad \text{for some } i_1, i_2, \dots, i_k \in \mathbb{N} \quad \text{with } i_1 < i_2 < \dots < i_k \leq |v|.$$

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

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

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 to solve our problem. A thorough presentation of normality can be read from the monographs [2, 4].

**DEFINITION 2.1** (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  $\ell$  if every  $v \in A^\ell$  verifies that

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

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

From now on, we fix a base  $b \geq 2$  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.

### 2.1. Champernowne-like words

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

**DEFINITION 2.2** (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  $\widehat{A} = \{0, 1\}$ ,  $w_3 = 000\ 001\ 010\ 011\ 100\ 101\ 110\ 111$ . The spaces have been added just for clarity.

**DEFINITION 2.3** (The wildcard morphism). Let  $B = \{b, \star\}$  the alphabet consisting of only the symbols ‘ $b$ ’ and ‘ $\star$ ’. We define the *wildcard function*  $(\star) : \widehat{A} \rightarrow B$  such that if  $s \in \widehat{A}$ ,

$$(\star)(s) = \begin{cases} b & \text{if } s = b, \\ \star, & \text{otherwise.} \end{cases}$$

The function  $(\star)$  extends to  $\widehat{A}^*$  as the morphism that replaces all the symbols different from ‘ $b$ ’ from a word with a wildcard ‘ $\star$ ’. Given a word  $v \in \widehat{A}^*$ , we write  $v^\star = (\star)(v)$ .

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  $v \downharpoonright n$  to denote the word that results from removing the first  $n$  symbols of  $v$ .

**DEFINITION 2.4** (The expansion of order  $n$  of a given word). For each  $n \in \mathbb{N}$  we let

$$\begin{aligned} \ell_n &= \|w_n^*\|_\star, \\ \widehat{\ell}_n &= |w_n^*|. \end{aligned}$$

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

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

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

The expansion  $e_n : A^{\ell_n} \rightarrow \widehat{A}^{\widehat{\ell}_n}$  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 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.

The expansion  $e_n$  extends to  $(A^{\ell_n})^*$  by concatenation. 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 2.5.**  $\widehat{\ell}_n = n(b+1)^n$  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

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

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

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

**OBSERVATION 2.6.** 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 2.7.** 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 2.8.** Given  $v, w \in \widehat{A}^*$ , then  $v = w$  if and only if  $v^* = w^*$  and  $r(v) = r(w)$ .

**OBSERVATION 2.9.** If  $B = \{b, \star\}$  and  $v \in B^n$ , then  $\|w_n^*\|_v = b^{\|v\|_\star}$ .

**OBSERVATION 2.10.** If  $v \in A^{\ell_n}$  and  $w \in (A^{\ell_n})^*$ , then  $\|w\|_v = \|e_n(w)\|_{e_n(v)}$ .

**LEMMA 2.11.** Given  $w \in \widehat{A}^n$  then

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

*Proof.* By Observation 2.7 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 2.8 we get  $\|e_n(u)\|_w$  is equal to

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

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

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

By Observation 2.9 we conclude that

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

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

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

Let

$$I = \{0 \leq i < \ell_n/n : \mathbb{1}\left(\left(w_n^*\right)[in + 1, in + n] = w^*\right) = 1\}$$

be the set of indexes where the first term of the product of equation (1) does not vanish. Notice that  $I$  does not depend on  $u$ .

Analyzing the second term of the product of equation (1), we observe that

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

Applying this we reduce (1) to

$$\|e_n(u)\|_w = \sum_{i \in I} \mathbb{1}\left(u[m(n, in) + 1, m(n, in + n)] = r(w)\right). \quad (2)$$

For  $i \in I$  we have that

$$\left(e_n(u)[in + 1, in + n]\right)^* = w^*,$$

which implies that

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

Summing (2) over all  $u \in A^{\ell_n}$  we get

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

And applying Observation 2.6 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

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

□

## 2.2. 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 1.1.

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  $\ell$  in  $u$  is

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

We can measure how far this frequency is from the case where all words of length  $\ell$  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  $\ell$  is the maximum of this distance among all  $v \in A^\ell$  and we denote it by  $\Delta_{A,\ell}(u)$ .

**DEFINITION 2.12** (Discrepancy of a finite word for a given length  $\ell$ ).

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

An easy equivalence is that  $u$  is simply normal to length  $\ell$  if and only if

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

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  $\ell$  be a length and let  $\varepsilon$  be a real number 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 2.13** (Main Lemma). *For each  $n \in \mathbb{N}$  there exists a constant  $c_n > 0$  such that for every  $\varepsilon > 0$  and every word  $v \in (A^{\ell_n})^*$  we have that if*

$$\Delta_{A,\ell_n}(v) < \varepsilon, \tag{3}$$

then

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

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

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

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 2.10 this reduces to

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

Applying (3) 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 Lemma 2.11 we get

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

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

$$\frac{|w|}{|e_n(v)|} \|e_n(v)\|_w < \frac{n|v|}{\ell_n |e_n(v)|} (1 + b^{\ell_n} \varepsilon). \quad (4)$$

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  $1 \leq i \leq t$ . So, we conclude that  $|e_n(v)| = t\widehat{\ell}_n$ .

Using this on (4) we get

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

using Observation 2.5 we can replace the value of  $\widehat{\ell}_n$  and get

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

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

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

□

### 2.3. Other useful results

**PROPOSITION 2.14.** *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*

$$\Delta_{C,mn}(v) < \varepsilon \tag{5}$$

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 (5) we get

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

Using Observation 2.7 we get

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

Using Observation 2.6 we get

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

**PROPOSITION 2.15.** *Given a finite alphabet  $C$ , some  $n \in \mathbb{N}$  and  $u, v \in (C^n)^*$ , if*

$$\Delta_{C,n}(u) < \varepsilon \quad (6)$$

and

$$\Delta_{C,n}(uv) < \varepsilon, \quad (7)$$

then

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

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

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

Using (6) and (7) 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

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

Thus

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

In a similar way we can conclude that

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

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

**PROPOSITION 2.16.** *Given a finite alphabet  $C$ , some  $n \in \mathbb{N}$  and  $u, v \in (C^n)^*$ , if*

$$\Delta_{C,n}(u) < \varepsilon \quad (8)$$

and

$$\Delta_{C,n}(v) < \frac{|uv| + |u|}{|v|} \varepsilon \quad (9)$$

then

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

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

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

Using (8) and (9) 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

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

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.  $\square$

Our analysis so far focuses in aligned occurrences of a given word in an expanded word. For a technical reason the proof of Theorem 1.1 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 2.17.** *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*

$$\Delta_{C,n}(u) < \varepsilon, \tag{10}$$

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 (10) we get,

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

Since

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

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 2.6 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,

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

as desired. □

The first paragraph in the proof above yields the following.

**OBSERVATION 2.18.** 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 2.19** (Hot Spot Lemma, Piatetski-Shapiro 1951 [8]). *Let  $x \in A^\omega$ . Then,  $x$  is normal if and only if there is positive constant  $C$  such that for all lengths  $\ell$  and for every word  $u \in A^\ell$ ,*

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

### 3. Proof of Theorem 1.1

We construct inductively a sequence of nonempty finite substrings  $\{v_i\}_{i \in \mathbb{N}}$  of  $v$  that verifies that  $v_1 v_2 \dots v_k$  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  $L_{n-1} = |v_1 v_2 \dots v_{n-1}|$  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 2.13, 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  $v_1 v_2 \dots v_n = v[1, L_{n-1} + t_n \ell_{2^n}]$  and thus it 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  $\ell_{2^i}$ , 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  $0 \leq j \leq t_{n+1}$ , and define

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

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

$$\Delta_{A,\ell_{2^{n+1}}}(v_n) < \varepsilon_n \quad (11)$$

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

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

Using Proposition 2.15 we have that

$$\Delta_{A,\ell_{2^{n+1}}}(v'_{n+1}) < \frac{|v_n v'_{n+1}| + |v_n|}{|v'_{n+1}|} \varepsilon_n. \quad (12)$$

Now, by (11) and Proposition 2.14 we have that

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

and applying Lemma 2.13 we get

$$\Delta_{\widehat{A},2^n}(e_{2^n}(v_n)) < b^{\ell_{2^n}} c_{2^n} \varepsilon_n \quad (13)$$

Similarly, applying Lemma 2.13 to (12) we get

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

and by Proposition 2.14 we conclude

$$\Delta_{\widehat{A},2^n}(e_{2^{n+1}}(v'_{n+1})) < \frac{|v_n v'_{n+1}| + |v_n|}{|v'_{n+1}|} (b+1)^{2^n} c_{2^{n+1}} \varepsilon_n. \quad (14)$$

Using Proposition 2.16 with (13) and (14) we get that

$$\begin{aligned} \Delta_{\widehat{A},2^n}(e_{2^n}(v_n) e_{2^{n+1}}(v'_{n+1})) &< 3 \max(b^{\ell_{2^n}} c_{2^n}, (b+1)^{2^n} c_{2^{n+1}}) \varepsilon_n \\ &< \frac{1}{(b+1)^{2^n} n}. \end{aligned} \quad (15)$$

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

$$\Delta_{\widehat{A},2^n}(e_{2^n}(v_n)) < \frac{1}{(b+1)^{2^n} n}. \quad (16)$$

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  $0 \leq j \leq t_n$ , we define

$$L_{n,j} = |v_1 v_2 \dots v_n v_{n+1,1} v_{n+1,2} \dots v_{n+1,j}|.$$

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

$$L_{n,j-1} \leq M \leq L_{n,j}. \quad (17)$$

By Observation 2.18 we get

$$\begin{aligned} |\widehat{v}[1, M]|_u &\leq |\widehat{v}[1, L_{n,j}]|_u \\ &\leq \left( \sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u \right) + |e_{2^n}(v_n) e_{2^{n+1}}(v_{n+1,1}) \dots \\ &\quad \dots e_{2^{n+1}}(v_{n+1,j})|_u + (n-1)(|u| - 1). \end{aligned} \quad (18)$$

Given that we have (16) for each term of the sum, we can apply Proposition 2.17 and we get the bound

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

Noticing that

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

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

$$\frac{|u| - 1}{2^i} + \frac{1}{i} \leq \frac{1}{(b+1)^{|u|}}. \quad (19)$$

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

$$\begin{aligned} \sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u &\leq \sum_{i=1}^{i_0} |e_{2^i}(v_i)| \left( \frac{|u| - 1}{2^i} + \frac{1}{(b+1)^{|u|}} + \frac{1}{i} \right) \\ &\quad + \sum_{i=i_0+1}^{n-1} |e_{2^i}(v_i)| \left( \frac{|u| - 1}{2^i} + \frac{1}{(b+1)^{|u|}} + \frac{1}{i} \right) \\ &\quad - (n-1)(|u| - 1). \end{aligned} \quad (20)$$

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  $\delta$  does not depend on  $M$ ) and using (19) 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

$$\sum_{i=1}^{n-1} |e_{2^i}(v_i)|_u \leq \delta + (L_{n-1,0} - L_{i_0,0}) \frac{2}{(b+1)^{|u|}} - (n-1)(|u|-1). \quad (21)$$

Having (15) and using Proposition 2.17 with the second term of (18) 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

$$\begin{aligned} |e_{2^n}(v_n) e_{2^{n+1}}(v_{n+1,1}) \dots e_{2^{n+1}}(v_{n+1,j})|_u &\leq \\ (L_{n,j} - L_{n-1,0}) &\left( \frac{2}{(b+1)^{|u|}} \right). \end{aligned} \quad (22)$$

Using (21) and (22) in (18) 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

$$\frac{|\widehat{v}[1, M]|_u}{M} \leq \frac{\delta}{M} + \frac{L_{n,j} - L_{i_0,0}}{M} \frac{2}{(b+1)^{|u|}}. \quad (23)$$

By (17) 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 (23) we get

$$\begin{aligned} \frac{|\widehat{v}[1, M]|_u}{M} &\leq \frac{\delta}{M} + \frac{2M - L_{i_0,0}}{M} \frac{2}{(b+1)^{|u|}} \\ &< \frac{\delta}{M} + \frac{4}{(b+1)^{|u|}}. \end{aligned} \tag{24}$$

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

$$\limsup_{M \rightarrow \infty} \frac{|\widehat{v}[1, M]|_u}{M} \leq \frac{4}{(b+1)^{|u|}}. \tag{25}$$

Since this bound is valid for all  $u \in \widehat{A}^*$ , using the Hot Spot Lemma (Lemma 2.19) with  $C = 4$  we have 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 1.1.

## 4. Remarks about the proof of Theorem 1.1

### 4.1. On the choice of $w_n$

We chose the sequence  $(w_n)$  of Champernowne like words to replicate the pattern of appearances of the new symbol in  $\widehat{v}$ . 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 2.11 and Lemma 2.13. 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 Lemma 2.13 but also involving the  $\delta_n$ . Then, we can use this bound in the proof of the Theorem 1.1. If we choose the  $\delta_n$  to be small enough (and maybe depending of the  $\varepsilon_n$ ), we can adapt the proof to work for this new sequence  $(w_n)$ . Any normal word  $z \in \widehat{A}^\omega$  can be split in consecutive strings  $z_1, z_2, \dots, z_n, \dots$  such that

$$z = z_1 z_2 \dots z_n \dots$$

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

Recall that a computable real number is a number that can be produced by an algorithm, that outputs each of the digits in its expansion in a given integer base, one after the other. 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 our approximation will be at each step of the algorithm.

**ACKNOWLEDGEMENT.** I am grateful to Verónica Becher for posing the problem and for her comments during the development of this work at the University of Buenos Aires.

#### REFERENCES

- [1] AGAFONOV, V. N.: *Normal sequences and finite automata*. (English. Russian original) Zbl 0242.94040 Sov. Math., Dokl. **9** (1968), 324–325. (Translation from Dokl. Akad. Nauk SSSR 179, 255-256 (1968)).
- [2] BECHER, V.—CARTON, O.: *Normal Numbers and Computer Science*. In: (Valérie Berthé and Michel Rigó, ed.) *Sequences, groups, and number theory, Trends in Mathematics Series*. Birkhauser/Springer, Cham, 2018. pp. 233–269.
- [3] BOREL, É.: *Les probabilités dénombrables et leurs applications arithmétiques*, Rend. Circ. Mat. Palermo **27** (1909), no. 1. 247–271.
- [4] BUGEAUD, Y.: *Distribution Modulo One and Diophantine Approximation*. In: *Cambridge Tracts in Mathematics Vol. 193*, Cambridge University Press, Cambridge, 2012.
- [5] CHAMPERNOWNE, D.: *The Construction of Decimals Normal in the Scale of Ten*. J. London Math. Soc. **8** (1933), no. 4, 254–260.
- [6] ROGERS, H. JR.: *Theory of recursive functions and effective computability* (second edition). MIT Press, Cambridge, MA, 1987.

ARIEL ZYLBER

- [7] KAMAE, T.—WEISS, B.: *Normal numbers and selection rules*, Israel J. Math. **21** (1975), no. 2, 101–110.
- [8] PIATETSKI-SHAPIRO, I. I.: *On the law of distribution of the fractional parts of the exponential function*, Izv. Akad. Nauk SSSR Ser. Mat. **15** (1951), no. 1, 47–52.
- [9] VANDEHEY, J.: *Uncanny subsequence selections that generate normal numbers*, Unif. Distrib. Theory **12** (2018), no. 2, 65–75.
- [10] BESICOVITCH, A. S.: *The asymptotic distribution of the numerals in the decimal representation of the squares of the natural numbers*, Math. Z. **39** (1935), no. 1, 146–156.

Received April 18, 2021

Accepted November 8, 2021

**Ariel Zylber**

*Departamento de Computación*

*Facultad de Ciencias Exactas y Naturales*

*Universidad de Buenos Aires*

*Intendente Güiraldes 2160*

*C1428EGA Buenos Aires*

*ARGENTINA*

*E-mail: azylber@dc.uba.ar*