

On normality in shifts of finite type

Nicolás Álvarez

Olivier Carton

February 27, 2019

Abstract

In this paper we consider the notion of normality of sequences in shifts of finite type. A sequence is normal if the frequency of each block exists and is equal to the Parry measure of the block. We give a characterization of normality in terms of incompressibility by lossless transducers. The result was already known in the case of the full shift.

1 Introduction

In this paper, we extend to the context of shifts the strong link between normality and incompressibility by finite state machines. This link was known for the full shift, that is the set of all infinite sequences over a fixed alphabet since the work of Schnorr and Stimm [18] and Dai *et al.* [10].

On the one hand, normality is a weak notion of randomness. It has been introduced by Borel in [6] more than a hundred years ago. Roughly speaking, it is required for a sequence to be normal that for each length, all possible blocks of that length occur with the same frequency in the sequence. It has been shown by Borel that almost all sequences (in a measure-theoretic sense) are normal but almost nothing is known about specific sequences coming from number theory like the expansions in some base of fundamental numbers as $\sqrt{2}$ or π . On the other hand, compressibility of sequences, especially by finite state machines, also known as transducers, has been studied since the early days of computer science [13]. This is mainly due to the large range of applications of compression techniques. These two important notions are linked together by the characterization of normality by incompressibility. Normal sequences are exactly those which cannot be compressed by some one-to-one transducer. This is a rather robust characterization as it is valid for many variants of transducers: non-deterministic, two-way [3, 8].

The notion of normality has been extended to broader contexts like the one of dynamical systems and especially shifts of finite type [15]. These extensions are based on the different characterizations of normality, namely uniform distribution [5] and martingales [2]. When sofic shifts are irreducible and aperiodic, they have a measure of maximal entropy and a sequence is then said to be normal if the frequency of each block equals its measure. This extension to shifts meets the original aim of normality to study expansions of numbers in bases when the shift arises from a numerical systems like the β -shifts coming from the numeration in a non-integer base β . Normality can be again interpreted as the good distribution of blocks of digits in the expansion of a number in a base β .

There are two main contributions in this paper. The first one is to give different formulations of the definition of normality and to show that they indeed coincide. These formulations

are similar to the ones used in the classical setting of the full shift and their equivalence in that case is well-known [7]. The second contribution is a characterization theorem that proves that a sequence is normal in a shift of finite type exactly when it is incompressible in the shift by finite state transducers. Again this characterization holds in the case of the full shift [18, 10].

This link between normality and automata opens the question whether selecting digits in a sequence with an oblivious automaton preserves normality in a shift of finite type as it does in the case of the full shift [1]. Oblivious means here that the selection of a digit is based on the prefix of the sequence before the digit but not including it.

The paper is organized as follows. Section 2 introduces all basic notions like shifts and normality. Section 3 is devoted to the equivalence of the different definitions of normality given in the previous section. The notion of incompressibility by transducer is defined in Section 4. The main result is stated and proved in Section 5.

2 Preliminaries

2.1 Notation

We write \mathbb{N} for the set of all natural numbers. An alphabet A is a finite set with at least two symbols. We write A^ω for the set of all infinite words over A and A^k stands for the set of all words of length k . The length of a finite word w is denoted by $|w|$. The positions in finite and infinite words are numbered starting from 1. For a word w and positions $1 \leq i \leq j \leq |w|$, we let $w[i]$ and $w[i..j]$ denote respectively the symbol at position i and the subword of w from position i to position j (inclusive).

For any finite set S we denote its cardinality with $|S|$. We write \log for the logarithm in base 2.

2.2 Shift spaces and subshifts of finite type

In this article we are going to work on shift spaces, in particular subshifts of finite type (SFT). Let A be a given alphabet. The *full shift* is the set A^ω of all (one-sided) infinite sequences $(x_n)_{n \geq 0}$ of symbols in A . The shift σ is the function from A^ω to A^ω which maps each sequence $(x_n)_{n \geq 0}$ to the sequence $(x_n)_{n \geq 1}$ obtained by removing the first symbol.

Let $F \subset A^*$ a set of finite words called *forbidden blocks*. The subshift X_F is the subset of A^ω made of sequences without any occurrences of blocks in F . More formally, it is the set

$$X_F = \{x : x[m..n] \notin F \text{ for each } 1 \leq m \leq n\}.$$

A *shift space* of A^ω or simply a *shift* is a subset X of A^ω which is closed for the product topology and invariant under the shift operator, that is $\sigma(X) = X$. This is equivalent to the existence of a subset $F \subset A^*$ of forbidden blocks such that $X = X_F$. The shift space is said to be of *finite type* if $X = X_F$ for some finite set F of forbidden blocks [14, Def. 2.1.1]. Up to a change of alphabet, any shift space of finite type is the same as a shift space X_F where any forbidden block has length 2, that is $F \subset A^2$. For simplicity, we always assume that each forbidden block has length 2. In that case, the set F is fully determined by the $A \times A$ -matrix $M = (m_{ab})_{a,b \in A}$ where $m_{ab} = 1$ if $ab \notin F$ and $m_{ab} = 0$ otherwise and we write $X = X_M$. The shift X is called *irreducible* if the graph induced by the matrix M is strongly connected, that is, for each symbols $a, b \in A$, there exists an integer n (depending on a and b) such that

$M_{ab}^n > 0$. The shift X is called *irreducible and aperiodic* if there exists an integer n such that $M_{ab}^n > 0$ for each symbols $a, b \in A$.

For a shift X and $n \in \mathbb{N}$, we let $\mathcal{B}(X)$ (resp., $\mathcal{B}_n(X)$) denote the set of all blocks (resp., blocks of length n) that occur in sequences of X . The *(topological) entropy* $h(X)$ of the shift X is defined by

$$h(X) = \lim_{n \rightarrow \infty} \frac{\log |\mathcal{B}_n(X)|}{n}.$$

If $X = X_M$ for some $\{0, 1\}$ matrix M , the entropy $h(X)$ can be computed as follows. By the Perron-Frobenius theory, the non-negative matrix M has a positive eigenvalue λ of greatest modulus [19, Thm 1.5]. The entropy $h(X)$ of $X = X_M$ is then equal to $\log \lambda$ [12, Obs. 1.4.2].

Example 1 (Golden mean shift). *The golden mean shift is the shift space $X_F \subset \{0, 1\}^\omega$ where the set of forbidden blocks is $F = \{11\}$. It is made of all sequences over $\{0, 1\}$ with no two consecutive 1. This subshift is also equal to X_M where $M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and its entropy is therefore $\log \lambda$ where $\lambda = (1 + \sqrt{5})/2$ is the golden mean.*

A probability measure on A^* is a function $\mu : A^* \rightarrow [0, 1]$ such that $\mu(\varepsilon) = 1$ and

$$\sum_{a \in A} \mu(wa) = \mu(w)$$

holds for each word $w \in A^*$. The simplest example of a probability measure is a *Bernoulli measure*. It is a morphism from A^* to $[0, 1]$ (endowed with multiplication) such that $\sum_{a \in A} \mu(a) = 1$. Among the Bernoulli measures there is the *uniform measure* which maps each word $w \in A^*$ to $|A|^{-|w|}$. In particular, each symbol a is mapped to $\mu(a) = 1/|A|$.

By the Carathéodory extension theorem, a measure μ on A^* can be uniquely extended to a probability measure $\hat{\mu}$ on A^ω such that $\hat{\mu}(wA^\omega) = \mu(w)$ holds for each word $w \in A^*$. In the rest of the paper, we use the same symbol for μ and $\hat{\mu}$. A probability measure μ is said to be *(shift) invariant* if the equality

$$\sum_{a \in A} \mu(aw) = \mu(w)$$

holds for each word $w \in A^*$. We now introduce the entropy of a measure [21, Chap. 4]. The *entropy* $h(\mu)$ of a measure μ is defined by

$$h(\mu) = \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{w \in A^n} \mu(w) \log \mu(w)$$

with the usual convention $0 \log 0 = 0$.

For a stochastic matrix P and a stationary distribution π , that is a line vector such that $\pi P = \pi$, the *Markov measure* $\mu_{\pi, P}$ is the invariant measure defined by the following formula [12, Lemma 6.2.1].

$$\mu_{\pi, P}(a_1 a_2 \cdots a_k) = \pi_{a_1} P_{a_1 a_2} \cdots P_{a_{k-1} a_k}$$

A simple computation shows that the entropy $h(\mu_{\pi, P})$ of such a measure is given by the following formula [12, Obs. 6.2.10].

$$h(\mu_{\pi, P}) = - \sum_{i, j \in A} \pi_i P_{ij} \log P_{ij}$$

with the convention $0 \log 0 = 0$.

A measure μ is *compatible* with a shift X if it only puts weight on blocks of X , that is, $\mu(w) > 0$ implies $w \in \mathcal{B}(X)$ for each word w . It is well known that each compatible measure μ satisfies $h(\mu) \leq h(X)$ [12, Obs. 6.2.13]. For a subshift of finite type, there is a unique compatible measure with maximal entropy $h(X)$ [12, Thm. 6.2.20]. This measure is called the *Parry measure* and it is a Markov measure. In the rest of the document we let μ^X denote the Parry measure of an SFT X . This measure can be explicitly given as follows. The Parry measure of an SFT X_M is the (one step) Markov measure given by the stochastic matrix $P = (P_{i,j})$ where $P_{i,j} = M_{i,j}r_j/\lambda r_i$ and the stationary probability distribution π defined by $\pi_i = l_i r_i$, where λ is the Perron eigenvalue of the matrix M and the vectors l and r are respectively left and right eigenvectors for λ normalized so that $\sum_{i=1}^k l_i r_i = 1$.

Example 2 (Parry measure of the golden mean shift). *Consider again the golden mean shift X . Its Parry measure is the Markov measure $\mu_{\pi,P}$ where π is the distribution $\pi = (\lambda^2/(1+\lambda^2), 1/(1+\lambda^2))$ and P is the stochastic matrix $P = \begin{pmatrix} 1/\lambda & 1/\lambda^2 \\ 1 & 0 \end{pmatrix}$ where λ is the golden mean.*

2.3 Normality

We start with the notation for the number of occurrences of a given word within another word.

Definition 3 (Occurrences). *For w and u two words, the number $|w|_u$ of occurrences of u in w and the number $\|w\|_{u,r}$ of aligned occurrences with offset r of u in w are respectively given by*

$$\begin{aligned} |w|_u &= |\{i : w[i..i+|u|-1] = u\}|, \\ \|w\|_{u,r} &= |\{i : w[i..i+|u|-1] = u \text{ and } i = r \bmod |u|\}|. \end{aligned}$$

The number $\|w\|_u$ of aligned occurrences is given by

$$\|w\|_u = \|w\|_{u,1}$$

For example, $|aaaa|_{aa} = 3$, $\|aaaa\|_{aa} = 2$ and $\|aaaa\|_{aa,2} = 1$.

Borel's definition [6] of normality for a sequence $x \in A^\omega$ is that x is *normal* if for each integer $\ell \geq 1$ and each word $w \in A^\ell$ of length ℓ ,

$$\lim_{n \rightarrow \infty} \frac{\|x[1..n\ell]\|_w}{n} = |A|^{-\ell}$$

This definition is extended to the case of an SFT by replacing the uniform measure by the Parry measure of the SFT. A sequence x of an SFT X is called *normal (in X)* if for each integer $\ell \geq 1$ and each word $w \in A^\ell$ of length ℓ ,

$$\lim_{n \rightarrow \infty} \frac{\|x[1..n\ell]\|_w}{n} = \mu^X(w)$$

where μ^X is the Parry measure of X . This definition is based on aligned occurrences. It will be seen in the next section that alternative definitions based on non-aligned occurrences are actually equivalent.

3 Equivalence between definitions of normality

In the literature there are several definitions of normality of a sequence. Some of them are based on aligned occurrences and some others are based on non-aligned occurrences. It is part of the folklore that all these definitions are indeed equivalent. For the classical normality, proofs can be found in [7, Thms 4.2 and 4.5]. For completeness, we provide here a proof for the case of Markov measure.

Theorem 4. *Let μ be a Markov measure on A^* . For each sequence x , the following three statements are equivalent.*

(1) Aligned normality: *for each integer ℓ and each word $w \in A^\ell$,*

$$\lim_{n \rightarrow \infty} \frac{\|x[1..n\ell]\|_w}{n} = \mu(w)$$

(2) Strong aligned normality: *for each $\ell, k \in \mathbb{N}$ and each word $w \in A^\ell$,*

$$\lim_{n \rightarrow \infty} \frac{\|\sigma^k(x)[1..n\ell]\|_w}{n} = \mu(w)$$

(3) Non-aligned normality: *for each word $w \in A^*$,*

$$\lim_{n \rightarrow \infty} \frac{|x[1..n]|_w}{n} = \mu(w).$$

Before proving the theorem, we state two very simple but useful lemmas. The first lemma states that obtaining a proper upper or lower bound for asymptotic frequencies of all words of a given length is sufficient to prove that limiting frequencies will follow the expected measure. The proof follows directly from the equality $\sum_{w \in A^\ell} \mu(w) = 1$ for each integer $\ell \geq 0$.

Lemma 5. *Let μ be a probability measure and ℓ a fixed non-negative integer. For each sequence $x \in A^\omega$, the following three statements are equivalent.*

(1) $\lim_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n = \mu(w)$ *for each $w \in A^\ell$.*

(2) $\limsup_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n \leq \mu(w)$ *for each $w \in A^\ell$.*

(3) $\liminf_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n \geq \mu(w)$ *for each $w \in A^\ell$.*

The next lemma states that it is sufficient to look at lengths which are multiples of a fixed integer k . The proof follows easily from the observation that there are at most k occurrences of w starting between positions kn and $k(n+1)$.

Lemma 6. *Let $k \in \mathbb{N}$ be a fixed positive integer. For each sequence x and each $\ell \in \mathbb{N}$ and each finite word $w \in A^\ell$, the following three statements hold.*

(1) $\liminf_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n = \liminf_{n \rightarrow \infty} \|x[1..nk\ell]\|_w/(nk)$

(2) $\limsup_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n = \limsup_{n \rightarrow \infty} \|x[1..nk\ell]\|_w/(nk)$

(3) $\lim_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n = \lim_{n \rightarrow \infty} \|x[1..nk\ell]\|_w/(nk)$ *if such limits exist.*

Lemma 5 and Lemma 6 are stated for aligned occurrences frequencies but they are also valid for occurrences frequencies.

Proof of Theorem 4. The equivalence between the three definitions of normality is proved as follows. We successively show that (1) implies (2), (2) implies (3) and that (3) implies (1).

(1) implies (2) It is sufficient to prove that if x presents aligned normality then $\sigma(x)$ also presents aligned normality.

For $w \in A^\ell$, $k \geq \ell$ and $1 \leq i \leq k - \ell + 1$ we define $B(k, w, i)$ as the set of words of length k which contains an occurrence of w at position i , that is $B = \{v \in A^k : v[i..i + |w| - 1] = w\}$. Since the Markov measure μ is invariant $\mu(B(k, w, i)) = \mu(w)$ for any $|w| \leq k$ and $1 \leq i \leq k - |w| + 1$.

For any $w \in A^\ell$ and $r \in \mathbb{N}$.

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{\|\sigma(x)[1..n\ell]\|_w}{n} &= \liminf_{n \rightarrow \infty} \frac{\|\sigma(x)[1..nr\ell]\|_w}{nr} \\ &\geq \liminf_{n \rightarrow \infty} \frac{1}{r} \sum_{k=0}^{r-2} \sum_{v \in B(r\ell, w, 2+\ell k)} \frac{\|x[1..nr\ell]\|_v}{n} \\ &= \frac{1}{r} \sum_{k=0}^{r-2} \sum_{v \in B(r\ell, w, 2+\ell k)} \mu(v) \\ &= \frac{r-1}{r} \mu(w) \end{aligned}$$

Since this inequality holds for any $r \in \mathbb{N}$.

$$\liminf_{n \rightarrow \infty} \frac{\|\sigma(x)[1..n\ell]\|_w}{n} \geq \mu(w)$$

and we conclude by Lemma 5.

(2) implies (3) Notice that for any $w \in A^\ell$,

$$|x[1..n]|_w = \sum_{i=0}^{\ell-1} \|\sigma^i(x)[1..n-i]\|_w$$

then

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{|x[1..n]|_w}{n} &= \sum_{i=0}^{\ell-1} \lim_{n \rightarrow \infty} \frac{\|\sigma^i(x)[1..n-i]\|_w}{n} \\ &= \sum_{i=0}^{\ell-1} \mu(w)/\ell = \mu(w) \end{aligned}$$

(3) implies (1) Let w be a finite word of length ℓ . For each word v , we define $\|v\|_{w,*} = \max_{i=1}^{\ell} \|v\|_{w,i}$. And, for a given $\varepsilon > 0$ and $k \in \mathbb{N}$, we define a set $\text{Bad}(w, k, \varepsilon)$ of words of length $k\ell - 1$ where the frequency of aligned occurrences of w is bad:

$$\text{Bad}(w, k, \varepsilon) = \{v \in A^{k\ell-1} : \|v\|_{w,*} > (k-1)(\mu(w) + \varepsilon)\}.$$

By the ergodic theorem for irreducible Markov chains [16, Thm. 1.10.2], for each positive real numbers $\delta, \varepsilon > 0$, there exists k_0 such that for any $k \geq k_0$,

$$\mu(\text{Bad}(w, k, \varepsilon)) < \delta.$$

Now, for any such $k \geq k_0$,

$$\begin{aligned}
\limsup_{n \rightarrow \infty} \frac{\|x[1..n\ell]\|_w}{n} &= \limsup_{n \rightarrow \infty} \frac{\|x[(k-1)\ell + 1..n\ell]\|_w}{n} \\
&\leq \limsup_{n \rightarrow \infty} \frac{1}{n(k-1)\ell} \sum_{t=1}^{(n-1)\ell+1} \|x[t..t + k\ell - 2]\|_{w,2-t} \\
&\leq \limsup_{n \rightarrow \infty} \frac{1}{n(k-1)\ell} \sum_{t=1}^{(n-1)\ell+1} \|x[t..t + k\ell - 2]\|_{w,*} \\
&= \limsup_{n \rightarrow \infty} \sum_{v \in A^{k\ell-1}} \frac{|x[1..(n+k-1)\ell - 1]|_v}{n\ell} \frac{\|v\|_{w,*}}{k-1} \\
&\leq \sum_{v \in A^{k\ell-1}} \left(\limsup_{n \rightarrow \infty} \frac{|x[1..(n+k-1)\ell - 1]|_v}{n\ell} \right) \frac{\|v\|_{w,*}}{k-1} \\
&= \sum_{v \in A^{k\ell-1}} \left(\limsup_{n \rightarrow \infty} \frac{|x[1..n\ell]|_v}{n\ell} \right) \frac{\|v\|_{w,*}}{k-1} \\
&= \sum_{v \in A^{k\ell-1}} \mu(v) \frac{\|v\|_{w,*}}{k-1} \\
&= \sum_{v \in A^{k\ell-1} \setminus \text{Bad}(w,k,\epsilon)} \mu(v) \frac{\|v\|_{w,*}}{k-1} + \sum_{v \in \text{Bad}(w,k,\epsilon)} \mu(v) \frac{\|v\|_{w,*}}{k-1} \\
&\leq (\mu(w) + \epsilon) \sum_{v \in A^{k\ell-1} \setminus \text{Bad}(w,k,\epsilon)} \mu(v) + \sum_{v \in A^{k\ell-1} \setminus \text{Bad}(w,k,\epsilon)} \mu(v) \\
&\leq \mu(w) + \epsilon + \delta
\end{aligned}$$

The inequality on the second line comes from the fact that every aligned occurrence of w in a position $j\ell + 1$ with $k-1 \leq j < n$ is counted $(k-1)\ell$ times as $\|x[t..t + k\ell - 2]\|_{w,2-t}$ for $(j+1-k)\ell + 2 \leq t \leq j\ell + 1$. This technique is due to Cassels [9]. Since the last inequality is true for any $\delta, \epsilon > 0$, it follows that $\limsup_{n \rightarrow \infty} \|x[1..n\ell]\|_w/n \leq \mu(w)$ and we conclude by Lemma 5. \square

4 Finite-state compressibility

In this section, we introduce the automata with output also known as transducers which are used to characterize normality by incompressibility. We consider *non-deterministic transducers* computing functions from sequences in a shift X to sequences in a shift Y , that is, for a given input sequence $x \in X$, there is at most one output sequence $y \in Y$. We focus on transducer that operate in real-time, that is, they process exactly one input alphabet symbol per transition. We start with the definition of a transducer.

Definition 7. A non-deterministic transducer is a tuple $\mathcal{T} = \langle Q, A, B, \delta, I, F \rangle$, where

- Q is a finite set of states,
- A and B are the input and output alphabets, respectively,

- $\delta \subset Q \times A \times B^* \times Q$ is a finite transition relation,
- $I \subseteq Q$ and $F \subseteq Q$ are the sets of initial and final states, respectively.

A transition of such a transducer is a tuple $\langle p, a, v, q \rangle$ in $Q \times A \times B^* \times Q$ which is written $p \xrightarrow{a|v} q$. A finite (respectively infinite) *run* is a finite (respectively infinite) sequence of consecutive transitions,

$$q_0 \xrightarrow{a_1|v_1} q_1 \xrightarrow{a_2|v_2} q_2 \cdots q_{n-1} \xrightarrow{a_n|v_n} q_n \quad (\text{resp. } q_0 \xrightarrow{a_1|v_1} q_1 \xrightarrow{a_2|v_2} q_2 \xrightarrow{a_3|v_3} q_3 \cdots).$$

Its *input and output labels* are $a_1 \cdots a_n$ and $v_1 \cdots v_n$ respectively. A finite run is written $q_0 \xrightarrow{a_1 \cdots a_n | v_1 \cdots v_n} q_n$. An infinite run is *final* if the state q_n is final for infinitely many integers n . In that case, the infinite run is written $q_0 \xrightarrow{a_1 a_2 a_3 \cdots | v_1 v_2 v_3 \cdots} \infty$. An infinite run is *accepting* if it is final and furthermore its first state q_0 is initial. This is the classical Büchi acceptance condition [17]. We always assume that for each sequence x , there is at most one sequence y such that there is an accepting run $q_0 \xrightarrow{x|y} \infty$ and we write $y = \mathcal{T}(x)$. In that case, it can be assumed that there is exactly one accepting run with input label x . By a slight abuse of notation, we write $\mathcal{T}(x[m..n])$ for the output of \mathcal{T} along that run while reading the factor $x[m..n]$. We always assume that all transducers are *trim*: each state can occur in an accepting run.

A transducer \mathcal{T} is called *bounded-to-one* (resp., *one-to-one*) if there is a constant K such that for each sequence y the set $\mathcal{T}^{-1}(y) = \{x : \mathcal{T}(x) = y\}$ has cardinality at most K (resp., at most 1). We call here *compressor* a bounded-to-one transducer. In the literature, *lossless* deterministic transducers are often considered. As it was shown in [3, Prop 2.1], this is an intermediate notion between one-to-one and bounded-to-one. We prefer not to use this notion as it is a structural property of the transducer and not of the function it realizes.

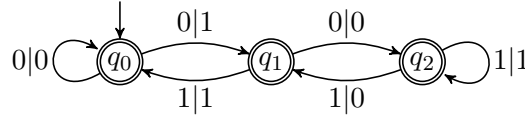


Figure 1: A transducer for the multiplication by 3 in base 2

The transducer pictured in Figure 1 is non-deterministic. It realizes multiplication by 3 on binary expansions of real numbers. If the input x is the binary expansion of some real number $\alpha < 1/3$, then the output is the binary expansion of 3α .

The *compression ratio* $\rho_{\mathcal{C}}(x)$ of a compressor \mathcal{C} on a sequence x is

$$\rho_{\mathcal{C}}(x) = \liminf_{n \rightarrow \infty} \frac{|\mathcal{C}(x[1..n])|}{n}.$$

A sequence x of a shift X is called *compressible in X* if there is a compressor $\mathcal{C} : X \rightarrow X$ such that $\rho_{\mathcal{C}}(x) < 1$.

5 Main Result

It follows from the results in [18, 10] that the words x with compression ratio $\rho(x)$ equal to 1 are exactly the normal words in the full shift. A direct proof of this result appears in [4].

Extensions of this characterization for non-determinism and extra memory appear in [3, 8]. The following theorem extends this result to the context of shifts of finite type.

Theorem 8. *Let X be an irreducible shift of finite type and x a sequence in X . The sequence x is normal in X if and only if it is incompressible in X .*

The following proposition is a very classical result showing that elements of a given shift can be encoded in another shift with a compression ratio close to the ratio of their entropies. To simplify the proof, we assume that the shifts are aperiodic but this is not really necessary. The result can also be extended to sofic shifts.

Proposition 9. *Let X and Y be two irreducible and aperiodic shifts of finite type. For each real number $\varepsilon > 0$, there is a one-to-one transducer \mathcal{T} from X into Y such that for each $x \in X$,*

$$\limsup_{n \rightarrow \infty} \frac{|\mathcal{T}(x[1..n])|}{n} < \frac{h(X)}{h(Y)} + \varepsilon.$$

Proof. Suppose that X and Y are the subshifts $X = X_M$ and $Y = X_N$ for the two $\{0, 1\}$ matrices M and N . Let λ and μ be respectively the eigenvalues of greatest modulus of M and N , so that $h(X) = \log \lambda$ and $h(Y) = \log \mu$. Let p/q be a rational number such that $h(X)/h(Y) < p/q < h(X)/h(Y) + \varepsilon$.

Since the sequence M^n/λ^n converges to some matrix, there is a constant c such that $\sum_{a,b \in A} M_{a,b}^n \leq c\lambda^n$ for each integer $n \geq 0$. Similarly, there is another constant d , a symbol a and an integer n_0 such that $N_{a,a}^n \geq d\mu^n$ for each $n \geq n_0$. Since $q \log \lambda < p \log \mu$, there is an integer n_1 such that $c\lambda^{qn} < d\mu^{pn}$ for each $n \geq n_1$. Let us recall that $M_{a,b}^n$ is the number of words w of length $n-1$ such that awb is a block of X and the sum $\sum_{a,b \in A} M_{a,b}^n$ is thus the number of blocks of length $n+1$ in X .

From the previous inequalities, it follows that, for $n \geq \max(n_0, n_1)$, the number $|\mathcal{B}_{qn+1}(X)|$ of blocks of length $qn+1$ in X is less than the number of words w of length $pn-1$ such that awa is a block of Y . Let us choose an integer n such that $n \geq \max(n_0, n_1)$. Let f be a one-to-one function which maps each block u of length $qn+1$ of X to a word $w = f(u)$ such that awa is a block of Y . The transducer \mathcal{T} reads each word $x \in X$ by blocks of length $qn+1$. For each read block u of length $qn+1$, it outputs aw where $w = f(u)$. Since f is one-to-one, the function realized by \mathcal{T} is also one-to-one. Furthermore, $\lim_{N \rightarrow \infty} |\mathcal{T}(x[1..N])|/N = pn/(qn+1) < h(X)/h(Y) + \varepsilon$. \square

The following corollary allows us to work with transducers from a specific shift into the full-shift and adapt those results to the case where the transducer has the same shift as domain and image.

Corollary 10. *There is a compressor $\mathcal{C} : X \rightarrow X$ such that $\rho_{\mathcal{C}}(x) < 1$ if and only if there is a compressor $\mathcal{C}' : X \rightarrow 2^\omega$ such that $\rho_{\mathcal{C}'}(x) < h(X)$.*

Proof. Suppose that there is a compressor $\mathcal{C} : X \rightarrow X$ such that $\rho_{\mathcal{C}}(x) < 1$. Let ε be a positive real number such that $(h(X) + \varepsilon)\rho_{\mathcal{C}}(x) < h(X)$. By the previous proposition with $Y = 2^\omega$, there is a transducer \mathcal{T} from X to 2^ω such that $\limsup_{n \rightarrow \infty} |\mathcal{T}(x[1..n])|/n \leq h(X) + \varepsilon$. The composition $\mathcal{C}' = \mathcal{T} \circ \mathcal{C}$ gives the required compressor. The converse is proved similarly by using the previous proposition with $X = 2^\omega$ and $Y = X$. \square

The following result is a classical generalization of Kraft's inequality. It is the key lemma used to prove that normal sequences cannot be compressed by finite state machines.

Lemma 11. *Let \mathcal{C} be a compressor from X to Y with $|Q|$ states. For each word w , let $L_{\mathcal{C}}(w)$ be the minimum number of symbols written by a finite run in \mathcal{C} with input label w . Then*

$$\sum_{w \in A^\ell} 2^{-L_{\mathcal{C}}(w)} \leq K|Q|^2(1 + \ell r_{\mathcal{C}})$$

where $|\mathcal{C}^{-1}(y)| \leq K$ for each $y \in Y$ and $r_{\mathcal{C}}$ is the maximum number of symbols written by a single transition of \mathcal{C} .

Proof. We claim that for each integer k the cardinality of the set $\{w \in A^\ell : L_{\mathcal{C}}(w) = k\}$ is at most $K|Q|^2 2^k$. Let p and q be two states of \mathcal{C} and v a word of length k . We claim that the set $\{w \in A^\ell : p \xrightarrow{w|v} q\}$ has cardinality at most K . Suppose that there are n distinct words w_1, \dots, w_n in this set. Since the transducer is trim, there is a finite run $i \xrightarrow{u|v'} p$ from an initial state i and a final run $q \xrightarrow{x|y} \infty$. It follows that $\mathcal{C}(uw_jx) = v'vy$ for each $1 \leq j \leq n$ and thus $n \leq K$. If a word w belongs to $\{w \in A^\ell : L_{\mathcal{C}}(w) = k\}$, then there are two states p, q and a word v of length k such that w belongs to $\{w \in A^\ell : p \xrightarrow{w|v} q\}$. This proves the upper bound for the cardinality of $\{w \in A^\ell : L_{\mathcal{C}}(w) = k\}$ since there are $|Q|^2$ possible choices for p and q and 2^k possible choices for v .

$$\begin{aligned} \sum_{w \in A^\ell} 2^{-L_{\mathcal{C}}(w)} &= \sum_{k=0}^{\ell r_{\mathcal{C}}} |\{w \in A^\ell : L_{\mathcal{C}}(w) = k\}| 2^{-k} \\ &\leq \sum_{k=0}^{\ell r_{\mathcal{C}}} K|Q|^2 = K|Q|^2(1 + \ell r_{\mathcal{C}}) \end{aligned}$$

□

Let $u \in A^n$ and $w \in A^\ell$ be two finite words of length n and ℓ and let x be an infinite word. First define the relative frequency $P(w, u)$ by $P(w, u) = \ell \|u\|_w / n$. This is just the number of aligned occurrences of w in u normalized by the factor ℓ/n such that $\sum_{w \in A^\ell} P(w, u) = 1$. The ℓ -block entropy $h_\ell(u)$ of u is then defined $h_\ell(u) = -\frac{1}{\ell} \sum_{w \in A^\ell} P(w, u) \log P(w, u)$. This ℓ -block entropy is extended to infinite words by setting $h_\ell(x) = \liminf_{k \rightarrow \infty} h_\ell(x[1..k\ell])$. The block entropy $h(x)$ of x is then defined by $h(x) = \liminf_{\ell \rightarrow \infty} h_\ell(x)$.

It should be noted that the block entropy $h_\ell(x)$ has been defined using aligned occurrences. This is the same as \hat{H}_ℓ in [20] but not the same as \tilde{H}_ℓ in [13] where entropy is defined using non-aligned occurrences. Therefore the existence of the limit $\lim_{\ell \rightarrow \infty} h_\ell(x)$ does not follow from the results in [13] and $h(x)$ is defined as $h(x) = \liminf_{\ell \rightarrow \infty} h_\ell(x)$.

Lemma 12 (Proof of Theorem 3 in [13]). *Given an alphabet A and a sequence $x \in A^\omega$. For any compressor $\mathcal{C} : A^\omega \rightarrow \{0, 1\}^\omega$:*

$$\rho_{\mathcal{C}}(x) \geq h(x)$$

For completeness, we present the proof of this theorem as given in [20].

Proof. Let us consider a bounded-to-one compressor \mathcal{C} with $|Q|$ states. Suppose that for each $y \in \{0, 1\}^\omega$, $|\mathcal{C}^{-1}(y)| \leq K$. For a word $w \in A^\ell$, \mathcal{C} produces an output depending on its current state. Let us denote as $L_{\mathcal{C}}(w)$ the length of the shortest output that \mathcal{C} produces when reading w , where the minimum is taken over all possible finite runs with w as input label.

$$\rho_{\mathcal{C}}(x[1..k\ell]) \geq \frac{1}{\ell} \sum_{w \in A^\ell} P(x[1..k\ell], w) \cdot L_{\mathcal{C}}(w)$$

Then,

$$h_\ell(x[1..k\ell]) - \rho_{\mathcal{C}}(x[1..k\ell]) \leq \frac{1}{\ell} \sum_{w \in A^\ell} P(x[1..k\ell], w) \log \left(\frac{2^{-L_{\mathcal{C}}(w)}}{P(x[1..k\ell], w)} \right)$$

By Jensen inequality applied to the log function,

$$h_\ell(x[1..k\ell]) - \rho_{\mathcal{C}}(x[1..k\ell]) \leq \frac{1}{\ell} \log \left(\sum_{w \in A^\ell} 2^{-L_{\mathcal{C}}(w)} \right)$$

By the generalized Kraft's inequality of Lemma 11,

$$h_\ell(x[1..k\ell]) - \rho_{\mathcal{C}}(x[1..k\ell]) \leq \frac{1}{\ell} \log(K|Q|^2(1 + \ell r_{\mathcal{C}}))$$

and taking first the limit when $k \rightarrow \infty$ and then the limit when $\ell \rightarrow \infty$ yields the required inequality $h(x) \leq \rho_{\mathcal{C}}(x)$. \square

We now come to the proof of the main theorem.

Proof of Theorem 8. Assume that the sequence x is normal in the shift X . Let $\phi : [0, 1] \rightarrow \mathbb{R}$ be defined as $\phi(p) = -p \log p$ with the usual convention that $0 \log 0 = 0$. Since ϕ is a continuous function and for every word $w \in A^\ell$, $\lim_{k \rightarrow \infty} P(w, x[1..k\ell]) = \mu(w)$

$$h_\ell(x) = \frac{1}{\ell} \sum_{w \in A^\ell} \phi(\mu(w)).$$

Therefore $h(x) = \liminf_{\ell \rightarrow \infty} h_\ell(x) = h(\mu) = h(X)$. By Lemma 12, there is no compressor $\mathcal{C}' : X \rightarrow \{0, 1\}^\omega$ with a compression ratio better than $h(X)$. By Corollary 10 we conclude that there is no compressor $\mathcal{C} : X \rightarrow X$ such that $\rho_{\mathcal{C}}(x) < 1$.

Now suppose that the sequence x is not normal. By definition, there is a finite word $w_0 \in A^*$ such that either

$$\lim_{n \rightarrow \infty} \frac{\|x[1..n\ell]\|_{w_0}}{n} \neq \mu^X(w_0)$$

or this limit does not exist where $\ell = |w_0|$ is the length of w_0 .

It is possible to choose a subsequence of positions $1 \leq n_1 < n_2 < n_3 < \dots$ such that the ratio $\|x[1..n_i\ell]\|_w / n_i$ converges for every $w \in A^\ell$ and such that the limit of this ratio is different from $\mu(w_0)$ for $w = w_0$.

Let $M = |\mathcal{B}_\ell(X)|$ be the number of blocks of length ℓ in X and let $B = \{1, 2, \dots, M\}$ be an alphabet of cardinality M . We can encode x into a sequence $y \in B^\omega$ by taking aligned words of length ℓ in x and representing them as a single symbol of B using a bijective mapping $f : \mathcal{B}_\ell(X) \rightarrow B$. The sequence y belongs to a subshift of finite type Y with entropy $h(Y) = \ell h(X)$.

For every $b \in B$, the limit $\lim_{i \rightarrow \infty} |y[1..n_i]|_b / n_i$ does exist, and for $b_0 = f(w_0)$, it satisfies $\lim_{i \rightarrow \infty} |y[1..n_i]|_{b_0} / n_i \neq \mu^Y(b_0) = \mu^X(w_0)$.

Let n'_1, n'_2, \dots be a subsequence of n_1, n_2, \dots such that the ratio $|y[1..n'_i]|_{ab}/n'_i$ converges for each $a, b \in B$. Define the distribution vector $\pi = (\pi_a)_{a \in B}$ and the stochastic matrix $P = (P_{ab})_{a, b \in B}$ by

$$\pi_a = \lim_{i \rightarrow \infty} \frac{|y[1..n'_i]|_a}{n'_i} \quad \text{and} \quad P_{ab} = \begin{cases} \lim_{i \rightarrow \infty} \frac{|y[1..n'_i]|_{ab}}{|y[1..n'_i]|_a} & \text{if } \pi_a \neq 0 \\ \frac{1}{M} & \text{otherwise.} \end{cases}$$

The stochastic matrix P is used to define a measure ν on A^* by setting for each word $a_1 a_2 \dots a_n \in B^*$

$$\nu(a_1 a_2 \dots a_n) = \frac{1}{M} \prod_{i=1}^{n-1} P_{a_i a_{i+1}}.$$

with the convention that $\nu(a) = 1/M$ for each symbol $a \in B$. Note that this measure might be not invariant because the vector $(1, \dots, 1)$ might be not a left eigenvector of the matrix P .

Let k be an integer to be fixed later. We construct an appropriate encoding of B^k based on the values of ν . Some care must be taken for words where ν takes the value 0. Let $S = \{u \in B^k : \nu(u) = 0\}$ be the subset of words of length k mapped to 0 by ν and $T = B^k \setminus S$ be its complement. Note that if $u = a_1 a_2 \dots a_k$ belongs to S , there is then some index $1 \leq i \leq k-1$ such that $P_{a_i a_{i+1}} = 0$, which means that $\lim_{i \rightarrow \infty} |y[1..n'_i]|_{a_i a_{i+1}}/n'_i = 0$, and in turn $\lim_{i \rightarrow \infty} |y[1..n'_i]|_u/n'_i = 0$.

If S is non-empty, define a one-to-one mapping

$$C_S : S \rightarrow \{0, 1\}^L \text{ where } L = \lceil \log |S| \rceil$$

For T , we define a prefix-free code

$$C_T : T \rightarrow \{0, 1\}^* \text{ such that } |C_T(u)| = \lceil -\log \nu(u) \rceil$$

The existence of such a code is guaranteed by Kraft's inequality since $\sum_{u \in T} \nu(u) = 1$. The functions C_S and C_T are now used to define a unique function $C_k : B^k \rightarrow \{0, 1\}^*$ as follows.

$$C_k(u) = \begin{cases} 0C_S(u) & \text{if } u \in S \\ 1C_T(u) & \text{if } u \in T \end{cases}$$

Since both functions C_S and C_T are one-to-one, the function C_k is also one-to-one. This latter function is now used to define a transducer $\mathcal{C} : Y \rightarrow \{0, 1\}^\omega$ which reads each sequence in Y by blocks of length k and for each read block $u \in B^k$ outputs $C_k(u)$. Since the function C_k is one-to-one, the transducer \mathcal{C} is also one-to-one. We now estimate its compression ratio $\rho_{\mathcal{C}}(y)$ on the input y .

$$\begin{aligned}
\rho_{\mathcal{C}}(y) &= \liminf_{n \rightarrow \infty} \frac{|\mathcal{C}(y[1..n])|}{n} \\
&\leq \liminf_{i \rightarrow \infty} \frac{|\mathcal{C}(y[1..n'_i])|}{n'_i} \\
&= \liminf_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{u \in B^k} \|y[1..n'_i]\|_u |C_k(u)| \\
&= \liminf_{i \rightarrow \infty} \frac{1}{n'_i} \left(\sum_{u \in S} \|y[1..n'_i]\|_u (L+1) + \sum_{u \in T} \|y[1..n'_i]\|_u (1 + |C_T(u)|) \right) \\
&= \liminf_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{u \in T} \|y[1..n'_i]\|_u (1 + |C_T(u)|) \\
&= \liminf_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{u \in T} \|y[1..n'_i]\|_u (1 + \lceil -\log \nu(u) \rceil) \\
&\leq \liminf_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{u \in T} \|y[1..n'_i]\|_u \left(2 + \log \frac{M}{\prod_{j=1}^{k-1} P_{u_i u_{i+1}}} \right) \\
&= \frac{(2 + \log M) \lfloor n'_i/k \rfloor}{n'_i} - \limsup_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{u \in T} \|y[1..n'_i]\|_u \sum_{j=1}^{k-1} \log(P_{u_i u_{i+1}}) \\
&\leq \frac{2 + \log M}{k} - \limsup_{i \rightarrow \infty} \frac{1}{n'_i} \sum_{j=1}^{n'_i-1} \log P_{y_i y_{i+1}} \\
&= \frac{2 + \log M}{k} - \limsup_{i \rightarrow \infty} \sum_{a,b \in B} \frac{\lfloor y[1..n'_i] \rfloor_{ab}}{n'_i} \log P_{ab} \\
&= \frac{2 + \log M}{k} - \sum_{a,b \in B} \pi_a P_{ab} \log P_{ab}
\end{aligned}$$

Since the last inequality is valid for any $k \in \mathbb{N}$, and

$$- \sum_{a,b \in B} \pi_a P_{ab} \log P_{ab} = h(\mu_{\pi,P}) < h(\mu^Y) = h(Y) = \ell h(X)$$

We conclude that there is a compressor $\mathcal{C} : Y \rightarrow \{0,1\}^\omega$, such that $\rho_{\mathcal{C}}(y) < \ell h(X)$. Now, define the compressor $\mathcal{C}' : X \rightarrow \{0,1\}^\omega$, which takes blocks of ℓ symbols from the input, maps them into B using the bisection $f : \mathcal{B}_\ell(X) \rightarrow B$ and then simulates the transducer \mathcal{C} to produce a binary output. Its compression ratio on the input x is given by

$$\rho_{\mathcal{C}'}(x) = \rho_{\mathcal{C}}(y)/\ell < h(X).$$

This inequality implies, by Corollary 10, that there is a compressor $\mathcal{C}'' : X \rightarrow X$, such that $\rho_{\mathcal{C}''}(x) < 1$. \square

6 Outlook

The main theorem (Theorem 8) is stated and proved for shifts of finite type for simplicity. We would like to provide some evidence that the result can be generalized to the case of sofic

shifts. The Parry measure of an irreducible sofic shift does exist and it is an hidden Markov chain (see [22, Thm 1], [11, Thm 4] and [14, p. 444]).

In the proof of Theorem 4, the fact that the measure μ is a Markov chain is only used through the ergodic theorem. Since this latter result also holds for hidden Markov chains, Theorem 4 can be lifted to hidden Markov chains.

The proof of Proposition 9 can be adapted to sofic shifts. The rest of the proof of Theorem 8 does not really use the fact that the shift X is of finite type.

Acknowledgments

The authors would like to thank the anonymous referee for useful suggestions and Verónica Becher for very fruitful discussions. Alvarez's work was financed through a postgraduate scholarship from CONICET (National Scientific and Technical Research Council of Argentina) and his stays at France to collaborate with Carton were possible thanks to financial support from Laboratoire International Associé Infinis, the Ministry of Education of Argentina and the French agency Campus France. Carton is member of the Laboratoire International Associé INFINIS. He is partially supported by the ECOS project PA17C04 and by the DeLTA project (ANR-16-CE40-0007).

References

- [1] V. N. Agafonov. Normal sequences and finite automata. *Soviet Mathematics Doklady*, 9:324–325, 1968.
- [2] J. Almarza and S. Figueira. Normality in non-integer bases and polynomial time randomness. *Journal of Computer and System Sciences*, 81:1059–1087, 2015.
- [3] V. Becher, O. Carton, and P. A. Heiber. Normality and automata. *Journal of Computer and System Sciences*, 81(8):1592–1613, 2015.
- [4] V. Becher and P. A. Heiber. Normal numbers and finite automata. *Theoretical Computer Science*, 477:109–116, 2013.
- [5] A. Bertrand-Mathis. Nombres normaux. *Journal de théorie des nombres de Bordeaux*, 8:397–412, 1996.
- [6] É. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*, 27:247–271, 1909.
- [7] Y. Bugeaud. *Distribution Modulo One and Diophantine Approximation*. Cambridge Tracts in Mathematics. Cambridge University Press, 2012.
- [8] O. Carton and P. A. Heiber. Normality and two-way automata. *Information and Computation*, 241:264–276, 2015.
- [9] J. W. S. Cassels. On a paper of Niven and Zuckerman. *Pacific Journal of Mathematics*, 2(4):555–557, 1952.
- [10] J. Dai, J. Lathrop, J. Lutz, and E. Mayordomo. Finite-state dimension. *Theoretical Computer Science*, 310:1–33, 2004.

- [11] R. Ficher. Sofic systems and graphs. *Monatshefte für Mathematik*, 80:179–186, 1975.
- [12] B. P. Kitchens. *Symbolic Dynamics*. Springer, 1998.
- [13] A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.
- [14] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1992.
- [15] M. Madritsch. Normal numbers and symbolic dynamics. In *Sequences*, chapter 8. Cambridge University Press, 2018.
- [16] J.R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- [17] D. Perrin and J.-É. Pin. *Infinite Words*. Elsevier, 2004.
- [18] C. P. Schnorr and H. Stimm. Endliche Automaten und Zufallsfolgen. *Acta Informatica*, 1:345–359, 1972.
- [19] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, 2006.
- [20] D. Sheinwald. On the Ziv-Lempel proof and related topics. *Proceedings of the IEEE*, 82(6):866–871, 1994.
- [21] P. Walter. *An introduction to Ergodic Theory*. Springer, 1982.
- [22] B. Weiss. Subshifts of finite type and sofic systems. *Monatshefte für Mathematik*, 77:462–474, 1977.

Nicolás Alvarez
 ICIC - Universidad Nacional del Sur, CONICET
 Departamento de Ciencias en Ingeniería de la Computación
 naa@cs.uns.edu.ar

Olivier Carton
 Institut de Recherche en Informatique Fondamentale
 Université Paris Diderot
 Olivier.Carton@irif.fr