The Word Entropy and How to Compute It

Sébastien Ferenczi[1(✉)], Christian Mauduit[1], and Carlos Gustavo Moreira[2]

[1] Aix Marseille Université, CNRS, Centrale Marseille, Institut de Mathématiques de Marseille, I2M - UMR 7373, 163, avenue de Luminy, 13288 Marseille Cedex 9, France
ssferenczi@gmail.com, mauduit@iml.univ-mrs.fr
[2] Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ 22460-320, Brazil
gugu@impa.br

Abstract. The complexity function of an infinite word counts the number of its factors. For any positive function f, its *exponential rate of growth* $E_0(f)$ is $\lim\limits_{n\to\infty} \inf \frac{1}{n} \log f(n)$. We define a new quantity, the *word entropy* $E_W(f)$, as the maximal exponential growth rate of a complexity function smaller than f. This is in general smaller than $E_0(f)$, and more difficult to compute; we give an algorithm to estimate it. The quantity $E_W(f)$ is used to compute the Hausdorff dimension of the set of real numbers whose expansions in a given base have complexity bounded by f.

Keywords: Word complexity · Positive entropy

1 Definitions

Let A be the finite alphabet $\{0, 1, \ldots, q-1\}$, If $w \in A^{\mathbb{N}}$, and $L(w)$ the set of finite factors of w; for any non-negative integer n, we write $L_n(w) = L(w) \cap A^n$. The classical complexity function is described for example in [2].

Definition 1. *The complexity function of $w \in A^{\mathbb{N}}$ is defined for any non-negative integer n by $p_w(n) = |L_n(w)|$.*

Our work concerns the study of infinite words w the complexity function of which is bounded by a given function f from $\mathbb{N}$ to $\mathbb{R}^+$. More precisely, if f is such a function, we put

$$W(f) = \{w \in A^{\mathbb{N}}, p_w(n) \leq f(n), \forall n \in \mathbb{N}\}.$$

Definition 2. *If f is a function from $\mathbb{N}$ to $\mathbb{R}^+$, we call exponential rate of growth of f the quantity*

$$E_0(f) = \lim_{n\to\infty} \inf \frac{1}{n} \log f(n)$$

and word entropy of f the quantity

$$E_W(f) = \sup_{w \in W(f)} E_0(p_w).$$

S. Brlek et al. (Eds.): WORDS 2017, LNCS 10432, pp. 157–163, 2017.
DOI: 10.1007/978-3-319-66396-8_15

Of course, if $E_0 = 0$ then E_W is zero also. Thus the study of E_W is interesting only when f has exponential growth: we are in the little-explored field of *word combinatorics in positive entropy*, or exponential complexity. For an equivalent theory in zero entropy, see [3,4].

2 First Properties of E_0 and E_W

The basic study of these quantities is carried out in [5], where the following results are proved.

If f is itself a complexity function (i.e. $f = p_w$ for some $w \in A^{\mathbb{N}}$), then $E_W(f) = E_0(f)$. But *in general E_W may be much smaller than E_0.*

We define mild regularity conditions for f: f is said to satisfy $(\mathcal{C})$ if the sequence $(f(n))_{n\geq 1}$ is strictly increasing, there exists $n_0 \in \mathbb{N}$ such that $\forall n \geq n_0 \Rightarrow f(2n) \leq f(n)^2$, $f(n+1) \leq f(1)f(n)$, and the sequence $\left(\frac{1}{n}\log f(n)\right)_{n\geq 1}$ converges.

But for each $1 < \theta \leq q$, and $n_0 \in \mathbb{N}$ such that $\theta^{n_0+1} > n_0 + q - 1$, we define the function f by $f(1) = q$, $f(n) = n + q - 1$ for $1 \leq n \leq n_0$ and $f(n) = \theta^n$ for $n > n_0$. We have $E_0(f) = \log\theta$ and it is proved that

$$E_W(f) \leq \frac{1}{n_0}\log(n_0 + q - 1),$$

which can be made arbitrarily small, independently of θ, while f satisfies $(\mathcal{C})$.

We define stronger regularity conditions for f.

Definition 3. *We say that a function f from $\mathbb{N}$ to $\mathbb{R}^+$ satisfies the conditions $(\mathcal{C}^*)$ if (i) for any $n \in \mathbb{N}$ we have $f(n+1) > f(n) \geq n+1$; (ii) for any $(n, n') \in \mathbb{N}^2$ we have $f(n+n') \leq f(n)f(n')$.*

But even with $(\mathcal{C}^*)$ we may have $E_W(f) < E_0(f)$. Indeed, let f be the function defined by $f(n) = \lceil 3^{n/2} \rceil$ for any $n \in \mathbb{N}$. Then it is easy to check that f satisfies conditions $(\mathcal{C}^*)$ and that $E_0(f) = \lim\limits_{n\to\infty} \frac{1}{n}\log f(n) = \log(\sqrt{3})$. On the other hand, we have $f(1) = 2$, $f(2) = 3$; thus the language has no 00 or no 11, and this implies that $E_W(f) \leq \log(\frac{1+\sqrt{5}}{2}) < E_0(f)$.

At least, under these conditions, we have the important

Theorem 4. *If f is a function from $\mathbb{N}$ to $\mathbb{R}^+$ satisfying the conditions $(\mathcal{C}*)$, then $E_W(f) > \frac{1}{2}E_0(f)$.*

It is also shown in [5] that the constant $\frac{1}{2}$ is optimal.

Finally, it will be useful to know that

Theorem 5. *For any function f from $\mathbb{N}$ to $\mathbb{R}^+$, there exists $w \in W(f)$ such that for any $n \in \mathbb{N}$ we have $p_w(n) \geq \exp(E_W(f)n)$.*

3 Algorithm

In general $E_W(f)$ is much more difficult to compute than $E_0(f)$; now we will give an algorithm which allows us to estimate with arbitrary precision $E_W(f)$ from finitely many values of f, if we know already $E_0(f)$ and have some information on the speed with which this limit is approximated.

We assume that f satisfies conditions $\mathcal{C}^*$. We don't loose too much generality with this assumption, since if the function f which satisfies the weaker conditions $\mathcal{C}$, we can replace it by the function $\tilde{f}$ given recursively by

$$\tilde{f}(n) := \min\{f(n), \min_{1\le k<n} \tilde{f}(k)\tilde{f}(n-k)\},$$

which satisfies conditions $\mathcal{C}^*$, such that $\tilde{f}(n) \le f(n), \forall n \in \mathbb{N}$ and $W(\tilde{f}) = W(f)$.

Theorem 6. *There is an algorithm which gives, starting from f and ε, a quantity h such that $(1-\varepsilon)h \le E_W(f) \le h$. h depends explicitely on ε, $E_0(f)$, N, $f(1)$, ..., $f(N)$, for an integer N which depends explicitely on ε, $E_0(f)$, and an integer n_0, larger than an explicit function of ε and $E_0(f)$, and such that*

$$\frac{\log f(n)}{n} < (1 + \frac{E_0(f)\varepsilon}{210(4+2E_0(f))})E_0(f), \quad \textit{for} \quad n_0 \le n < 2n_0.$$

We shall now give the algorithm. f is given and henceforth we omit to mention it in $E_0(f)$ and $E_W(f)$. Also given is $\varepsilon \in (0,1)$.

Description of the algorithm

– Let

$$\delta := \frac{E_0\varepsilon}{105(4+2E_0)} < \frac{\varepsilon}{210}.$$

– Let

$$K := \lceil \delta^{-1} \rceil + 1.$$

– Choose a positive integer

$$n_0 \ge K \vee \frac{4K^2}{420^3 E_0}$$

such that

$$\frac{\log f(n)}{n} < (1 + \frac{\delta}{2})E_0, \forall n \ge n_0;$$

in view of conditions $\mathcal{C}^*$, this last condition is equivalent to $\frac{\log f(n)}{n} < (1+\frac{\delta}{2})E_0, n_0 \le n < 2n_0$.

– Choose intervals so large that all the lengths of words we manipulate stay in one of them. Namely, for each $t \geq 0$, let

$$n_{t+1} := \exp(K((1+\delta)^2 E_0 n_t + E_0)).$$

We take

$$N := n_K.$$

– Choose a set $Y \subset A^N$: for each possible Y, we define $L_n(Y) = \cup_{\gamma \in Y} L_(\gamma)$, $q_n(Y) := |L_n(Y)|$, for $1 \leq n \leq N$. We look at those Y for which $q_n(Y) \leq f(n), \forall n \leq N$, and choose one among them such that

$$\min_{1 \leq n \leq N} \frac{\log q_n(Y)}{n}$$

is maximum.

– By Lemma 7 below, on one of the large intervals we have defined, namely $[n_r, n_{r+1}]$, $\frac{\log q_n(Y)}{n}$ will be almost constant. Let

$$h := \frac{\log q_{n_r}(Y)}{n_r}.$$

Here is the lemma we needed; henceforth, Y is fixed and we omit to mention it in the $q_n(Y)$:

Lemma 7. *There exists $r < K$, such that*

$$\frac{\log q_{n_r}}{n_r} < (1+\delta)\frac{\log q_{n_{r+1}}}{n_{r+1}}.$$

Proof. Otherwise $\frac{\log q_{n_0}}{n_0} \geq (1+\delta)^K \frac{\log q_{n_K}}{n_K}$: as $K > \frac{1}{\delta}$, $(1+\delta)^K$ would be close to e for δ small enough, and is larger than $\frac{9}{4}$ as $\delta < \frac{1}{2}$; thus, as $\frac{\log q_{n_K}}{n_K} \geq E_W$ by the proof of Proposition 8 below, we have $\frac{\log q_{n_0}}{n_0} \geq \frac{9}{4} E_W$, but $q_{n_0} \leq f(n_0)$ hence $\frac{\log q_{n_0}}{n_0} < (1+\frac{\delta}{2})E_0$, and this contradicts $E_0 \leq 2E_W$, which is true by Theorem 4.

We prove now that indeed h is a good approximation of the word entropy.

Proposition 8.

$$h \geq E_W.$$

Proof. We prove that

$$\min_{1 \leq n \leq N} \frac{\log q_n}{n} \geq E_W.$$

We know by Theorem 5 that there is $\hat{w} \in W(f)$ with $p_n(\hat{w}) \geq \exp(E_W n)$, for all $n \geq 1$. For such a word $\hat{w}$, let $X := L_N(\hat{w}) \subset A^N$. We have, for each n with $1 \leq n \leq N$, $L_n(X) = L_n(\hat{w})$ and $f(n) \geq \#L_n(\hat{w}) = p_n(\hat{w}) \geq \exp(E_W n)$. Thus X is one of the possible Y, and the result follows from the maximality of $\min_{1 \leq n \leq N} \frac{\log q_n}{n}$.

What remains to prove is the following proposition (which, understandably, does not use the maximality of $\min_{1\le n\le N} \frac{\log q_n}{n}$).

Proposition 9.

$$(1-\varepsilon)h \le E_W.$$

Proof. Our strategy is to build a word w such that, for all $n \ge 1$,

$$\exp((1-\varepsilon)hn) \le p_n(w) \le f(n),$$

which gives the conclusion by definition of E_W. To build the word w, we shall define an integer m, and build successive subsets of $L_m(Y)$; for such a subset Z, we order it (lexicographically for example) and define $w(Z)$ to be the *Champernowne word* on Z: namely, if $Z = \{\beta_1, \beta_2, ..., \beta_t\}$, we build the infinite word

$$w(Z) := \beta_1\beta_2\dots\beta_t\beta_1\beta_1\beta_1\beta_2\beta_1\beta_3\dots\beta_{t-1}\beta_t\beta_1\beta_1\beta_1\dots\beta_t\beta_t\beta_t\dots$$

made by concatenation of all words in Z followed by the concatenations of all pairs of words of Z followed by the concatenations of all triples of words of Z, etc.

The word $w(Z)$ will satisfy $\exp((1-\varepsilon)hn) \le p_n(w(Z))$ for all n as soon as

$$|Z| \ge \exp((1-\varepsilon)hm),$$

since, for every positive integer k, we will have at least $|Z|^k$ factors of length km in $w(Z)$.

The successive (decreasing) subsets Z of $L_m(Y)$ we build will all have cardinality at least $\exp((1-\varepsilon)hm)$, and the words $w(Z)$ will satisfy $p_n(w(Z)) \le f(n)$ for n in an interval which will increase at each new set Z we build, and ultimately contains all the integers.

We give only the main ideas of the remaining proof. In the first stage we define two lengths of words, $\hat{n}$ and $m > \frac{\hat{n}}{2\varepsilon}$, which will be both in the interval $[n_r, n_{r+1}]$, and a set Z_1 of words of length m of the form $\gamma\theta$, for words γ of length $\hat{n}$, such that the word $\gamma\theta\gamma$ is in $L_{m+\hat{n}}(Y)$. This is done by looking precisely at twin occurrences of words.

Let $\tilde{\varepsilon} = \frac{\varepsilon}{15} = \frac{7(4+2E_0)\delta}{E_0} > 14\delta$; then we can get such a set Z_1 with $|Z_1| \ge \exp((1-\tilde{\varepsilon})h(m+\hat{n}))$.

In the second stage, we define a new set $Z_2 \subset Z_1$ in which all the words have the same prefix γ_1 of length $6\tilde{\varepsilon}hm$, and all the words have the same suffix γ_2 of length $6\tilde{\varepsilon}hm$, with $|Z_2| \ge |Z_1|\exp(-12\tilde{\varepsilon}hm - 2\delta h\hat{n})$, and $2\delta h\hat{n} \le (1-\tilde{\varepsilon})\hat{n}$, thus

$$|Z_2| \ge \exp((1-13\tilde{\varepsilon})hm).$$

As a consequence of the definition of Z_2, all words of Z_2 have the same prefix of length $\hat{n}$, which is a prefix γ_0 of γ_1; as Z_2 is included in Z_1, any word of Z_2 is of the form $\gamma_0\theta$, amd the word $\gamma_0\theta\gamma_0$ is in $L_{m+\hat{n}}(Y)$.

At this stage we can prove

Claim. $p_{w(Z_2)}(n) \leq f(n)$ for all $1 \leq n \leq \hat{n}+1$.

Let us shrink again our set of words.

Lemma 10. *For a given subset Z of Z_2, there exists $Z' \subset Z$, $|Z'| \geq (1-\exp(-(j-1)\frac{E_0}{2}))^j|Z|$, such that the total number of factors of length $\hat{n}+j$ of all words $\gamma_0\theta\gamma_0$ such that $\gamma_0\theta$ is in Z' is at most $f(\hat{n}+j)-j$.*

We start from Z_2 and apply successively Lemma 10 from $j=2$ to $j=6\tilde{\varepsilon}m$, getting $6\tilde{\varepsilon}m-1$ successive sets Z'; at the end, we get a set Z_3 such that the total number of factors of length $\hat{n}+j$ of words $\gamma_0\theta\gamma_0$ for $\gamma_0\theta$ in Z_3 is at most $f(\hat{n}+j)-j$ for $j=2,\ldots,6\tilde{\varepsilon}m$, and $\frac{|Z_3|}{|Z_2|}$ is at least

$$\Pi_{2\leq j\leq 6\tilde{\varepsilon}m-\hat{n}}(1-\exp(-(j-1)\frac{E_0}{2}))^j \geq \Pi_{j\geq 2}(1-\exp(-(j-1)\frac{E_0}{2}))^j,$$

which implies after computations that

$$|Z_3| \geq \exp((1-14\tilde{\varepsilon})hm).$$

We can now bound the number of short factors by using the factors we have just deleted and properties of γ_0, γ_1 and γ_2.

Claim. $p_{w(Z_3)}(n) \leq f(n)$ for all $1 \leq n \leq 6\tilde{\varepsilon}m$.

We shrink our set again.

Let $m \geq n > 6\tilde{\varepsilon}m$; in average a factor of length n of a word in Z_3 occurs in at most $\frac{m|Z_3|}{f(n)}$ elements of Z_3. We consider the $\frac{f(n)}{mn^2}$ factors of length n which occur the least often. In total, these factors occur in at most $\frac{m|Z_3|}{f(n)}\frac{f(n)}{mn^2} = \frac{|Z_3|}{n^2}$ elements of Z_3. We remove these words from Z_3, for all $m \geq n > 6\tilde{\varepsilon}m$, obtaining a set Z_4 with $|Z_4| \geq \exp((1-15\tilde{\varepsilon})hm)$.

We can now control medium length factors, using again the missing factors we have just created, and γ_1 and γ_2, but not γ_0.

Claim. $p_{w(Z_4)}(n) \leq f(n)$ for all $1 \leq n \leq m$.

Finally we put $Z_5 = Z_4$ if $|Z_4| \leq \exp((1-4\tilde{\varepsilon})hm)$, otherwise we take for Z_5 any subset of Z_4 with $\lceil \exp((1-4\tilde{\varepsilon})hm) \rceil$ elements. In both cases we have

$$|Z_5| \geq \exp((1-\varepsilon)hm).$$

For the long factors, we use mainly the fact that there are many missing factors of length m, but we need also some help from γ_1 and γ_2.

Claim. $p_{w(Z_5)}(n) \leq f(n)$ for all n.

In view of the considerations at the beginning of the proof of Proposition 9, Claim 3 completes the proof of that proposition, and thus of Theorem 6.

4 Application

We define

$$C(f) = \{x = \sum_{n \geq 0} \frac{w_n}{q^{n+1}} \in [0,1], w(x) = w_0 w_1 \cdots w_n \cdots \in W(f)\}.$$

We are interested in the Hausdorff dimensions of this set, see [1] for definitions; indeed, the main motivation for studying the word entropy is Theorem 4.8 of [5]:

Theorem 11.
The Hausdorff dimension of $C(f)$ *is equal to* $E_W(f)/\log q$.

References

1. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. Wiley, Chichester (1990)
2. Ferenczi, S.: Complexity of sequences and dynamical systems. Discrete Math. **206**(1–3), 145–154 (1999). http://dx.doi.org/10.1016/S0012-365X(98)00400-2, (Tiruchirappalli 1996)
3. Mauduit, C., Moreira, C.G.: Complexity of infinite sequences with zero entropy. Acta Arith. **142**(4), 331–346 (2010). http://dx.doi.org/10.4064/aa142-4-3
4. Mauduit, C., Moreira, C.G.: Generalized Hausdorff dimensions of sets of real numbers with zero entropy expansion. Ergodic Theor. Dynam. Syst. **32**(3), 1073–1089 (2012). http://dx.doi.org/10.1017/S0143385711000137
5. Mauduit, C., Moreira, C.G.: Complexity and fractal dimensions for infinite sequences with positive entropy (2017)