Quasiperiods, Subword Complexity and Pisot Numbers

Ronny Polley*, and Ludwig Staiger†

June 2014

Abstract

A quasiperiod of a word or an infinite string is a word which covers every part of the string. A word or an infinite string is referred to as quasiperiodic if it has a quasiperiod. It is obvious that a quasiperiodic infinite string cannot have every word as a subword (factor). Therefore, the question arises how large the set of subwords of a quasiperiodic infinite string can be [3].

Here we show that on the one hand the maximal subword complexity of quasiperiodic infinite strings and on the other hand the size of the sets of maximally complex quasiperiodic infinite strings both are intimately related to the smallest Pisot number t_P (also known as *plastic constant*).

We provide an exact estimate on the maximal subword complexity for quasiperiodic infinite words.

Keywords: quasiperiodic words, subword complexity, Hausdorff measure

In his tutorial [3] Solomon Marcus discussed some open questions on quasiperiodic infinite words. Soon after its publication Levé and Richomme [2] gave answers on some of the open problems. In connection with Marcus' Question 2 they presented a quasiperiodic infinite word (with quasiperiod aba) of exponential subword complexity, and they posed the new question of what is the maximal complexity of a quasiperiodic infinite word.

In a recent paper [5] we estimated the maximal asymptotic (in the sense of [9]) subword complexity of quasiperiodic infinite words. More precisely, it is shown in [5] that every quasiperiodic infinite word ξ has at most $f(\xi, n) \leq O(1) \cdot t_P^n$ factors (subwords) of length n, where t_P is the smallest Pisot number, that is, the unique positive root of the polynomial $t^3 - t - 1$. Moreover, the general construction of [8, Section 5] yields quasiperiodic infinite words achieving this bound. In fact, also Levé's and Richomme's [2] example meets this upper bound.

Surprisingly, it turned out in [5] that there are also infinite words meeting this bound having $aabaa$—a different word—as quasiperiod. Moreover, it was shown that all other quasiperiods yield infinite words asymptotically below this bound.

The aim of this paper is to compare these two maximal quasiperiods aba and $aabaa$ in order to obtain an answer which one of them yields infinite words of greater complexity. Here we compare the quasiperiods aba and $aabaa$ in two respects.

1. Which one of the words aba or $aabaa$ generates the larger set (ω-language) of infinite words having q as quasiperiod, and
2. which one of the words aba or $aabaa$ generates an ω-word ξ_q having a maximal subword function $f(\xi_q, n)$?

*itCampus Software- und Systemhaus GmbH, Leipzig, D-04229 Leipzig, Germany

†Corresponding author, Martin-Luther-Universität Halle-Wittenberg, Institut für Informatik, von-Seckendorff-Platz 1, D–06099 Halle (Saale), Germany

As a measure of ω-languages in Item 1 we use the Hausdorff dimension and Hausdorff measure of a subset of the Cantor space of infinite words (ω-words). We obtain that, when neglecting the fixed prefix q of quasiperiodic ω-words having this quasiperiod q, for both words, the sets of ω-words having quasiperiod aba or $aabaa$ have the same Hausdorff dimension $\log t_P$ and the same Hausdorff measure t_p.

A difference for these quasiperiods appears when we consider the constant in the bound on $f(\xi, n)$. It turns out that the bounding constants c_{aba} and c_{aabaa} satisfy $c_{aba} < c_{aabaa}$, thus $aabaa$ is the quasiperiod having the maximally achievable subword complexity for quasiperiodic ω-words.

1 Notation

In this section we introduce the notation used throughout the paper. By $\mathbb{N} = \{0, 1, 2, \ldots\}$ we denote the set of natural numbers. Let X be an alphabet of cardinality $|X| = r \geq 2$. By X^* we denote the set of finite words on X, including the *empty word* e, and X^ω is the set of infinite strings (ω-words) over X. Subsets of X^* will be referred to as *languages* and subsets of X^ω as *ω-languages.*

For $w \in X^*$ and $\eta \in X^* \cup X^\omega$ let $w \cdot \eta$ be their *concatenation.* This concatenation product extends in an obvious way to subsets $L \subseteq X^*$ and $B \subseteq X^* \cup X^\omega$. For a language L let $L^* := \bigcup_{i \in \mathbb{N}} L^i$, and by $L^\omega := \{w_1 \cdots w_i \cdots : w_i \in L \setminus \{e\}\}$ we denote the set of infinite strings formed by concatenating words in L. Furthermore $|w|$ is the *length* of the word $w \in X^*$ and $\mathbf{pref}(B)$ is the set of all finite prefixes of strings in $B \subseteq X^* \cup X^\omega$. We shall abbreviate $w \in \mathbf{pref}(\eta)$ ($\eta \in X^* \cup X^\omega$) by $w \sqsubseteq \eta$.

We denote by $B/w := \{\eta : w \cdot \eta \in B\}$ the *left derivative* of the set $B \subseteq X^* \cup X^\omega$. As usual, a language $L \subseteq X^*$ is *regular* provided it is accepted by a finite automaton. An equivalent condition is that its set of left derivatives $\{L/w : w \in X^*\}$ is finite.

The sets of infixes of B or η are $\mathbf{infix}(B) := \bigcup_{w \in X^*} \mathbf{pref}(B/w)$ and $\mathbf{infix}(\eta) := \bigcup_{w \in X^*} \mathbf{pref}(\{\eta\}/w)$, respectively. In the sequel we assume the reader to be familiar with basic facts of language theory.

2 Quasiperiodicity

2.1 General properties

A finite or infinite word $\eta \in X^* \cup X^\omega$ is referred to as *quasiperiodic* with quasiperiod $q \in X^* \setminus \{e\}$ provided for every $j < |\eta| \in \mathbb{N} \cup \{\infty\}$ there is a prefix $u_j \sqsubseteq \eta$ of length $j - |q| < |u_j| \leq j$ such that $u_j \cdot q \sqsubseteq \eta$, that is, for every $w \sqsubseteq \eta$ the relation $u_{|w|} \sqsubset w \sqsubseteq u_{|w|} \cdot q$ is valid (cf. [2, 3]).

Next we introduce the finite language P_q which generates the set of quasiperiodic ω-words having quasiperiod q. We set

$$P_q := \{v : e \sqsubset v \sqsubseteq q \sqsubset v \cdot q\} . \tag{1}$$

The following characterisation of ω-words having quasiperiod q is found in [5].

$$\{\xi : \xi \in X^\omega \wedge \xi \text{ has quasiperiod } q\} = P_q^\omega = \{\xi : \xi \in X^\omega \wedge \mathbf{pref}(\xi) \subseteq \mathbf{pref}(P_q^*)\} \tag{2}$$

3 Hausdorff Dimension and Hausdorff Measure

3.1 General properties

First, we shall briefly describe the basic formulae needed for the definition of Hausdorff measure and Hausdorff dimension of a subset of X^ω. For more background and motivation see Section 1 of [4].

In the setting of languages and ω-languages this can be read as follows (see [4, 8]). For $F \subseteq X^\omega$, $r = |X| \geq 2$ and $0 \leq \alpha \leq 1$ the equation

$$\mathbb{L}_\alpha(F) := \lim_{l\to\infty} \inf\Big\{ \sum_{w\in W} r^{-\alpha\cdot|w|} : F \subseteq W \cdot X^\omega \wedge \forall w(w \in W \Rightarrow |w| \geq l)\Big\} \tag{3}$$

defines the α-dimensional metric outer measure on X^ω. The measure $\mathbb{L}_\alpha$ satisfies the following properties (see [4, 8]).

Proposition 1 *Let $F \subseteq X^\omega$, $V \subseteq X^*$ and $\alpha \in [0,1]$.*

1. *If $\mathbb{L}_\alpha(F) < \infty$ then $\mathbb{L}_{\alpha+\varepsilon}(F) = 0$ for all $\varepsilon > 0$.*
2. *It holds the scaling property $\mathbb{L}_\alpha(w \cdot F) = r^{-\alpha\cdot|w|} \cdot \mathbb{L}_\alpha(F)$.*

Then the *Hausdorff dimension* of F is defined as

$$\dim F := \sup\{\alpha : \alpha = 0 \vee \mathbb{L}_\alpha(F) = \infty\} = \inf\{\alpha : \mathbb{L}_\alpha(F) = 0\}\,.$$

It should be mentioned that dim is countably stable and invariant under scaling, that is, for $F_i \subseteq X^\omega$ we have

$$\dim \textstyle\bigcup_{i\in\mathbb{N}} F_i = \sup\{\dim F_i : i \in \mathbb{N}\} \quad \text{and} \quad \dim w \cdot F_0 = \dim F_0\,. \tag{4}$$

Lemma 2 *Let $V \subseteq X^*$ be regular language and $\dim V^\omega = \alpha$. Then $\mathbb{L}_\alpha(V^\omega) > 0$.*

3.2 The Hausdorff measure of P^ω_{aba} and P^ω_{aabaa}

In order to estimate the Hausdorff dimension and Hausdorff measure of the sets P^ω_{aba} and P^ω_{aabaa} we use the approach of [4]. To this end we consider for $F = P^\omega_q$ the adjacency matrix $\mathcal{A}_q$: Let $\{F/w : w \in \mathbf{pref}(F)\} = \{F_0 = F, F_1, \ldots, F_k\}$ (without repetitions) and $\mathcal{A}_q = (a_{i,j})^k_{i,j=0}$ where $a_{i,j} := |\{x : x \in X \wedge F_i/x = F_j\}|$. Then, according to [4, Section 3] $\dim P^\omega_q = \log_r \lambda_q$ where λ_q is the maximal eigenvalue of $\mathcal{A}_q$ and, for $\alpha = \dim P^\omega_q$, the value $\mathbb{L}_\alpha(P^\omega_q)$ is the topmost entry of a non-negative eigenvector $\vec{a}_q$ of $\mathcal{A}_q$ corresponding to λ_q having a 1 at specified positions (for more details see [4, Section 3]). Using this procedure we obtain $\dim P^\omega_{aba} = \dim P^\omega_{aabaa} = \log_r t_P$, $\mathbb{L}_\alpha(P^\omega_{aba}) = t_P^{-3}$ and $\mathbb{L}_\alpha(P^\omega_{aabaa}) = t_P^{-5}$.

This estimate, however, does not seem to represent the 'real' size of the sets P^ω_{aba} and P^ω_{aabaa}: All ω-words in P^ω_{aba} start with aba and all ω-words in P^ω_{aabaa} start with the longer word $aabaa$. Thus, in view of Proposition 1.2, these prefixes contribute the factors t_P^{-3} and t_P^{-5}, respectively, to the Hausdorff measure.

In order to eliminate the influence of the prefixes we consider instead the sets $\widehat{P}^\omega_q := \{\zeta : \exists v(v \in X^* \wedge v \cdot \zeta \in P^\omega_q)\}$ of all tails (suffixes) of ω-words in P^ω_q. Here the above procedure is likewise applicable. We obtain the adjacency matrices (see also Section 4.2)

$$\widehat{\mathcal{A}}_{aba} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad \widehat{\mathcal{A}}_{aabaa} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

and the values $\dim \widehat{P}_q^\omega = \log_r t_P$ and $\mathbb{L}_\alpha(\widehat{P}_q^\omega) = t_P$, for $q \in \{aba, aabaa\}$ and $\alpha = \log_r t_P$.

Remark 3 The sets of tails $\widehat{P}_{aba}^\omega$ and $\widehat{P}_{aabaa}^\omega$ can also be characterised via forbidden subwords: $\widehat{P}_{aba}^\omega = \{a,b\}^\omega \setminus \{a,b\}^* \cdot \{aaa, bb\} \cdot \{a,b\}^\omega$ and $\widehat{P}_{aabaa}^\omega = \{a,b\}^\omega \setminus \{a,b\}^* \cdot \{aaaaa, bab, bb\} \cdot \{a,b\}^\omega$. Here their Hausdorff dimension can also be obtained by Volkmann's [10] approach.

4 Subword Complexity

4.1 The subword complexity of quasiperiodic ω-words

In this section we investigate upper bounds on the subword complexity function $f(\xi, n)$ for quasiperiodic ω-words. If $\xi \in X^\omega$ is quasiperiodic with quasiperiod q then Eq. (2) shows $\mathbf{infix}(\xi) \subseteq \mathbf{infix}(P_q^*)$. Thus

$$f(\xi, n) \leq |\mathbf{infix}(P_q^*) \cap X^n| \text{ for } \xi \in P_q^\omega\,. \tag{6}$$

Similarly to the proof of Proposition 5.5 of [8] let $\xi_q := \prod_{v \in P_q^* \setminus \{e\}} v$ where the order of the factors $v \in P_q^* \setminus \{e\}$ is an arbitrary but fixed well-order, e.g. the length-lexicogrephical order. This implies $\mathbf{infix}(\xi) = \mathbf{infix}(P_q^*)$. Consequently, the tight upper bound on the subword complexity of quasiperiodic ω-words having a certain quasiperiod q is $f_q(n) := |\mathbf{infix}(P_q^*) \cap X^n|$.

The following facts are known from the theory of formal power series (cf. [1, 6]). As $\mathbf{infix}(P_q^*)$ is a regular language the power series $\mathfrak{s}_q^*(t) := \sum_{n \in \mathbb{N}} f_q(n) \cdot t^n$ is a rational series and, therefore, f_q satisfies a recurrence relation

$$f_q(n+k) = \textstyle\sum_{i=0}^{k-1} m_i \cdot f_q(n+i) \tag{7}$$

with integer coefficients $m_i \in \mathbb{Z}$. Thus $f_q(n) = \sum_{i=0}^{k'-1} g_i(n) \cdot \lambda_i^n$ where $k' \leq k$, λ_i are pairwise distinct roots of the polynomial $\chi_q(t) = t^n - \sum_{i=0}^{k-1} a_i \cdot t^i$ and g_i are polynomials of degree not larger than k.

The growth of $f_q(n)$ mainly depends on the (positive) root λ_q of largest modulus among the λ_i and the corresponding polynomial g_i. Using Corollary 4 of [7] (see also [5, Eq. (8)]) one can show—without explicitly inspecting the polynomials $\chi_q(t)$—that the polynomial g_i corresponding to the maximal root λ_q is constant.

Lemma 4 ([5, Lemma 16]) *Let $q \in X^* \setminus \{e\}$. Then there are constants $c_{q,1}, c_{q,2} > 0$ and a $\lambda_q \geq 1$ such that*

$$c_{q,1} \cdot \lambda_q^n \leq |\mathbf{infix}(P_q^*) \cap X^n| \leq c_{q,2} \cdot \lambda_q^n\,.$$

Next we are looking for those quasiperiods q which yield the largest value of λ_q among all quasiperiods.

Lemma 5 ([5, Lemma 18]) *Let X be an arbitrary alphabet containing at least the two letters a, b. Then the maximal value λ_q is obtained for $q = aba$ or $aabaa$.*
This value is $\lambda_{aba} = \lambda_{aabaa} = t_P$ where t_P is the positive root of the polynomial $t^3 - t - 1$.

Remark 6 The bound in Lemma 5 is independent of the size of the alphabet X. And indeed, quasiperiodic ω-words of maximal subword complexity have quasiperiods of the form aba or $aabaa$, $a, b \in X,\ a \neq b$, thus consist of only two different letters.

4.2 Quasiperiods of maximal subword complexity

We have seen that the quasiperiods aba and $aabaa$ yield quasiperiodic ω-words of maximal asymptotic subword complexity. In this section we investigate which one of these two quasiperiods yields ω-words $\xi \in \{a, b\}^\omega$ of larger subword complexity $f(\xi, n)$, that is, forces the larger constant $c_{q,2}$ ($q \in \{aba, aabaa\}$) in the upper bound of Lemma 4.

From the deterministic automata $\mathcal{B}_{aba}$ and $\mathcal{B}_{aabaa}$ (see Table 1) accepting the languages $\mathbf{infix}(P^*_{aba})$ and $\mathbf{infix}(P^*_{aabaa})$, respectively, we obtain the adjacency matrices $\widehat{\mathcal{A}}_{aba}$ and $\widehat{\mathcal{A}}_{aabaa}$ of Eq. (5) and their characteristic polynomials $\chi_{aba}(t) = t \cdot (t^3 - t - 1)$ and $\chi_{aabaa}(t) = t^2 \cdot (t^3 - t - 1) \cdot (t^2 + 1) = t^7 - t^4 - t^3 - t^2$.

$\mathcal{B}_{aba}$	z_0	z_1	z_2	z_3
a	z_3		z_3	z_1
b	z_2	z_2		z_2

$\mathcal{B}_{aabaa}$	s_0	s_1	s_2	s_3	s_4	s_5	s_6
a	s_1	s_5		s_4	s_5	s_6	s_2
b	s_3	s_3	s_3			s_3	s_3

Table 1: Automata $\mathcal{B}_{aba}$ and $\mathcal{B}_{aabaa}$ accepting $\mathbf{infix}(P^*_{aba})$ and $\mathbf{infix}(P^*_{aabaa})$, respectively

So both sequences $(|\mathbf{infix}(P^*_{aba}) \cap X^n|)_{n\in\mathbb{N}}$ and $(|\mathbf{infix}(P^*_{aabaa}) \cap X^n|)_{n\in\mathbb{N}}$ satisfy the recurrence relation $f_q(n+7) = f_q(n+4) + f_q(n+3) + f_q(n+2)$ with the initial values $(9, 7, 5, 4, 3, 2, 1)$ for $q = aba$ (see also [2]) and $(10, 8, 6, 4, 3, 2, 1)$ for $q = aabaa$ which shows already that the growth of $(|\mathbf{infix}(P^*_{aabaa}) \cap X^n|)_{n\in\mathbb{N}}$ is the larger one.

Finally we turn to the above mentioned constants $c_{q,2}$ for $q \in \{aba, aabaa\}$. The characteristic polynomials χ_{aba} and χ_{aabaa} have as root of maximal modulus the smallest Pisot number $t_P > 1$. The other roots satisfy $|t| < 1$ or, additionally, $t = \pm\sqrt{-1}$ in case of χ_{aabaa}.

Using the standard methods of recurrent relations one obtains for a quasiperiodic ω-word ξ with quasiperiod aba the largest achievable subword complexity $f(\xi, n) = \mathrm{INT}(\frac{2t_P^2+3t_P+2}{2t_P+3} \cdot t_P^n)$, for large n, where $\mathrm{INT}(\alpha)$ is the integer closest to the real α.

Similarly, for a quasiperiodic ω-word ξ with quasiperiod $aabaa$ the largest achievable subword complexity satisfies $|f(\xi, n) - \mathrm{INT}(\frac{13t_P^2+16t_P+9}{10t_P+15} \cdot t_P^n)| \leq 1$, for large n. Observe that for the constants it holds $\frac{2t_P^2+3t_P+2}{2t_P+3} < \frac{13t_P^2+16t_P+9}{10t_P+15}$.

References

[1] J. Berstel and D. Perrin, *Theory of Codes*, Academic Press Inc., Orlando, 1985.

[2] F. Levé and G. Richomme, *Quasiperiodic infinite words: Some answers*, Bulletin of the EATCS, 84:128–138, 2004.

[3] S. Marcus, *Quasiperiodic infinite words*, Bulletin of the EATCS, 82:170–174, 2004.

[4] W. Merzenich and L. Staiger, *Fractals, dimension, and formal languages*, RAIRO Inform. Théor. Appl., 28(3-4):361–386, 1994.

[5] R. Polley and L. Staiger, *The maximal subword complexity of quasiperiodic infinite words*, In I. McQuillan and G. Pighizzini, editors, Proceedings Twelfth Annual Workshop on Descriptional Complexity of Formal Systems, volume 31 of Electronic Proceedings in Theoretical Computer Science, 169–176, 2010.
`http://rvg.web.cse.unsw.edu.au/eptcs/content.cgi?DCFS2010`

[6] A. Salomaa and M. Soittola, *Automata-theoretic Aspects of Formal Power Series*, Springer-Verlag, New York, 1978.

[7] L. Staiger, *The entropy of finite-state ω-languages*, Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 14(5):383–392, 1985.

[8] L. Staiger, *Kolgomorov complexity and Hausdorff dimension*, Inf. Comput., 103(2):159–194, 1993.

[9] L. Staiger, *Asymptotic subword complexity*, In H. Bordihn, M. Kutrib and B. Truthe, editors, Languages Alive, volume 7300 of Lecture Notes in Computer Science, Springer-Verlag, Heidelberg, 236–245, 2012.

[10] B. Volkmann, *Über Hausdorffsche Dimensionen von Mengen, die durch Zifferneigenschaften charakterisiert sind V*, Math. Zeitschr. 65:389–413, 1953.