The complexity of tangent words

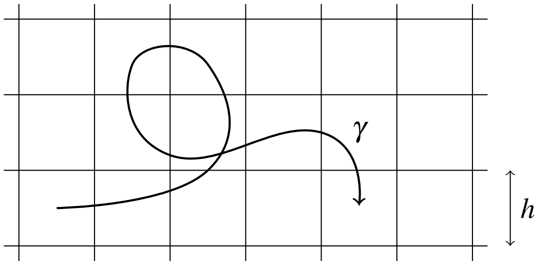
Thierry Monteil
CNRS – Université Montpellier 2
http://www.lirmm.fr/~monteil

In [7], we described the set of words that appear in the coding of smooth (resp. analytic) curves at arbitrary small scale. The aim of this paper is to compute the complexity of those languages.

Keywords: Cutting sequence, symbolic coding, word complexity, multigrid convergence, Sturmian word.

1 Introduction

A *smooth curve* is a map γ from a compact interval I of the real line to the plane, which is C^∞ and such that $||\gamma'(t)|| > 0$ for any $t \in I$ (this last property is called *regularity*). Any such curve can (and will be considered to) be arc-length reparametrised (*i.e.* $\forall t \in I, ||\gamma'(t)|| = 1$).
We can approximate such a curve by drawing a square grid of mesh h on the plane, and look at the sequence of squares that the curve meets. For a generic position of the grid, the curve γ does not hit any corner and crosses the grid transversally, hence the curve passes from a square to a square that is located either *r*ight, *u*p, *l*eft or *d*own of it. We record this sequence of moves and define the *cutting sequence* of the curve γ with respect to this grid as a word w on the alphabet $\{r,u,l,d\}$ which tracks the lines of the grid crossed by the curve γ.
The following picture shows a curve γ with cutting sequence *rruuldrrrd*.



Note that since the grid can be translated, a given curve may have more than one cutting sequence for a given mesh h. Our knowledge of the curve from one of its cutting sequences increases when the mesh h decreases, and when the mesh approaches 0, the local patterns of the cutting sequence play the role of discrete tangents. Such words are called *tangent words*, their first properties were described in [7]. Cutting sequences associated to straight segments are known to be exactly the *balanced words*, which are also the finite factors of Sturmian words. It turns out that the tangent words strictly contain balanced words, and that 2-balanced words strictly contain tangent words. The aim of this note is to count the number of tangent words (resp. tangent analytic words) of a given length, in order to quantify those inclusions.

Preliminary Report. Final version to appear in:
WORDS 2011

2 Tangent words

Tangent words are the finite words that appear in the cutting sequences of some smooth curve for arbitrary small scale. More precisely, let $F(\gamma,G)$ denote the set of factors of the cutting sequence of the curve γ with respect to the square grid G (when the curve hits a corner, the cutting sequence is not defined and we set $F(\gamma,G)=\emptyset$). We define the *asymptotic language* of γ by

$$T(\gamma)=\limsup_{mesh(G)\to 0} F(\gamma,G)=\bigcap_{\varepsilon>0}\bigcup_{mesh(G)\leq\varepsilon} F(\gamma,G).$$

More generally, when X is a set of curves, let us denote by $T(X)$ the set $\bigcup_{\gamma\in X} T(\gamma)$. When X is the set of smooth curves, we denote $T(X)$ by T^∞, and call its elements *tangent words*. When X is the set of analytic curves, we denote $T(X)$ by T^ω, and call its elements *analytic tangent words*. The two languages T^∞ and T^ω are factorial and extendable.

For the sake of simplicity, we will focus on curves going right and up, *i.e.* smooth curves such that both coordinates of $\gamma'(t)$ are positive for any t. Let us rename r and u by 0 and 1 respectively to stick to the usual notation about binary words.

The following results are proved in [7].

2.1 Combinatorial characterisation (desubstitution)

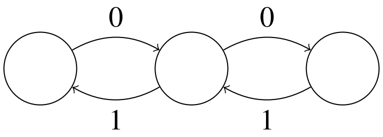
Balanced words are know to have a hierarchical structure, where the morphisms $\sigma_0=(0\mapsto 0, 1\mapsto 10)$ and $\sigma_1=(0\mapsto 01, 1\mapsto 1)$ play a crucial role [8] [5]. The same renormalisation applies to tangent words. Given a finite word w, we can "desubstitute" it by

- removing one 0 per run of 0 if 11 does not appear in w, or
- removing one 1 per run of 1 if 00 does not appear in w.

This desubstitution map (denoted by δ) consists in removing one letter per run of the non-isolated letter. An accelerated version of this desubstitution consists in removing a run equal to the length of the shortest inner run from any run of the non-isolated letter (including possible leading and trailing runs even if they have shorter length).

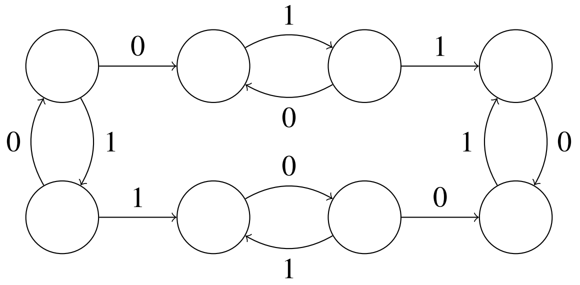
If we repeat this process as much as possible, we get a *derivated word* denoted by $d(w)$. The word w is balanced if, and only if, $d(w)$ is the empty word, and the derivation process is related to the continued fraction development of the slope of the associated straight segment.

A word is said to be *diagonal* if it is recognised by the following automaton with three states, which are all considered as initial and accepting:



A word is said to be *thin diagonal* if it is diagonal and only two states are visited during its recognition.

A word is said to be *non-oscillating diagonal* if it is recognised by the following automaton with eight states, which are all considered as initial and accepting:



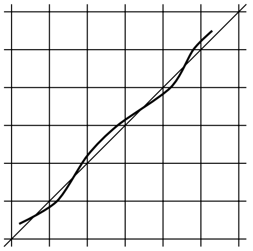
Proposition 1 *A finite word w is tangent if, and only if, $d(w)$ is diagonal.*
A finite word w is tangent analytic if, and only if, $d(w)$ is non-oscillating diagonal.

For example, the word $w = 100100010010010001001000100$ is tangent analytic since it can be desubstituted as $1\cancel{00}1\cancel{00}01\cancel{00}1\cancel{00}1\cancel{00}1\cancel{00}01\cancel{00}1\cancel{00}01\cancel{00} = 110111101101$, and then $\cancel{11}0\cancel{11}110\cancel{11}0\cancel{1} = 01100 = d(w)$, which is non-oscillating diagonal (start from the bottom left state).

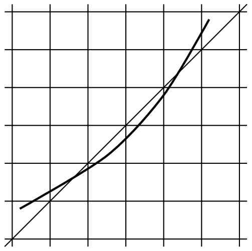
2.2 Geometric characterisation

Proposition 2 *A word w is tangent if, and only if, for any $\varepsilon > 0$, w is the cutting sequence of a smooth curve γ which is ε-close (for the C^1 norm) to a straight segment (the grid is fixed).*
A word w is tangent analytic if, and only if, for any $\varepsilon > 0$, w is the cutting sequence of a smooth curve γ with nowhere zero curvature which is ε-close (for the C^1 norm) to a straight segment (the grid is fixed).

For example, the word 0110100110 is tangent and the word 1001010110 is tangent analytic:



0110100110
tangent



1001010110
tangent analytic

3 Complexity

The *complexity* of a language L is the map that counts, for any integer n, the number of elements of L of length n. It is usually denoted by $p_n(L)$.
The complexity of the balanced words B was studied in [4], [6] and [1], where it was proved to be equal to:

$$p_n(B) = 1 + \sum_{i=1}^{n} \sum_{j=1}^{i} \varphi(j) = 1 + \sum_{i=1}^{n} (n-i+1)\varphi(i) \ ,$$

where φ denotes the Euler totient function: $\varphi(n) = card\{k \leq n \mid \gcd(k,n) = 1\}$.

To compute the complexity of T^∞ and T^ω, we will use the tools introduced by Julien Cassaigne using bispecial factors [3]. They have been used in the context of billiards in [2]. Let L be a factorial and extendable language on the alphabet $\{0,1\}$. A word w in L is said to be *bispecial* if $0w$, $1w$, $w0$, $w1$ are in L. A bispecial factor w is called

- *weak bispecial* if $card\{(a,b) \in \{0,1\}^2 \mid awb \in L\} = 2$,
- *ordinary bispecial* if $card\{(a,b) \in \{0,1\}^2 \mid awb \in L\} = 3$,
- *strong bispecial* if $card\{(a,b) \in \{0,1\}^2 \mid awb \in L\} = 4$.

Let $wb_n(L)$ (resp. $sb_n(L)$) denote the number of weak (resp. strong) bispecial factors of length n in L. Let $s_n(L)$ denote the first difference $p_{n+1}(L) - p_n(L)$. We have:

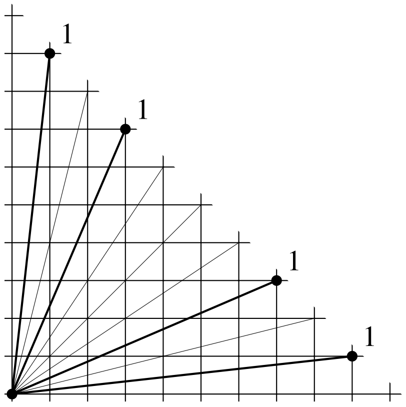
$$s_{n+1}(L) - s_n(L) = sb_n(L) - wb_n(L) \ .$$

Hence, by summing twice, if L is nontrivial, we have:

$$p_n(L) = 1 + n + \sum_{i=0}^{n-1} \sum_{j=0}^{i-1} (sb_j(L) - wb_j(L)) \ .$$

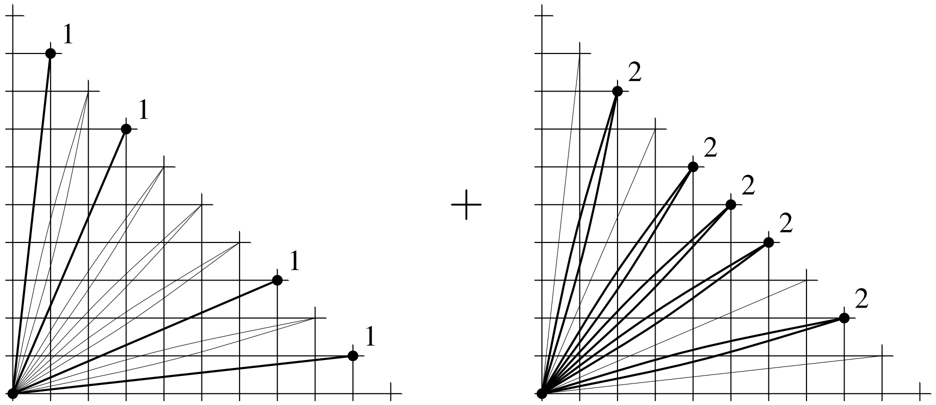
Let us first describe the combinatorial structure of bispecial factors in T^∞. Let w be a bispecial factor. If w is not diagonal, then it can be desubstituted (in a single way) and $\delta(w)$ is a bispecial factor of the same kind. Otherwise, if w is thin diagonal, then it is strong or ordinary bispecial depending on the parity of its length. Otherwise, w is diagonal and the three states are visited during its recognition: w is strong bispecial. Hence, there is no weak bispecial factor in T^∞. This also holds for T^ω.

The geometric characterisation of tangent (resp. tangent analytic) words is convenient to describe and count the strong bispecial factors. We can visualise the strong bispecial factors as follows. Pick a segment from $(0,0)$ to $(p,q) \in \mathbb{Z}^2_{>0}$.
If there is no integer point on the way (which happens precisely when $\gcd(p,q) = 1$), the coding of the corresponding open interval is a bispecial factor of length $p+q-2$ in both T^∞ and T^ω. Those words are also the bispecial factors for balanced words. There are $\varphi(n+2)$ such words of length n, this the geometrical meaning of Lipatov's formula [4].



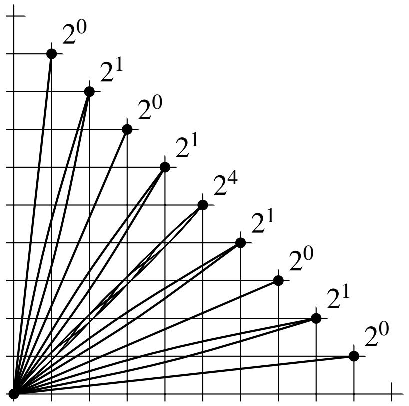
Balanced bispecial factors of length 8

Otherwise, there are $k \geq 1$ points one the way. For tangent analytic words, each such segment corresponds to two bispecial factors of length $p+q-2$: one bending above the k points, another bending under the k points. There are $2(n+2-\varphi(n+2))$ such words of length n.



Tangent analytic bispecial factors of length 8

For tangent words, each such segment corresponds to 2^k bispecial factors of length $p+q-2$ corresponding to all the possibilities of slaloming around the k integer points on the way. Hence, there are $\sum\limits_{\substack{d|n+2\\ d\neq 1}} \varphi(n+2)2^{(n+2)/d-1}$ strong bispecial factors of length n in T^∞.



Tangent bispecial factors of length 8

Proposition 3 *We have:*

$$p_n(T^\omega) = 1+n+\sum_{i=1}^{n}\sum_{j=2}^{i}(2j-\varphi(j)-1)$$

$$p_n(T^\infty) = 1+n+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=2}^{i}\sum_{\substack{d|j\\ d\neq 1}}\varphi(j)2^{j/d}$$

4 Conclusion

Let us recall that a word w is *k-balanced* if:

$$\forall u, v \in Fact(w) \quad |u| = |v| \Rightarrow ||u|_1 - |v|_1| \leq k \ .$$

Each class of words is strictly included in the next one:

- 1-balanced words (digital straight segments)
- tangent analytic words
- tangent words
- 2-balanced words

The complexity of the first two classes, is cubical whereas the complexity of the last two classes is exponential. It can be shown that analytic tangent words can be written as a concatenation of two 1-balanced words. What is the gap between tangent words and 2-balanced words ?

References

[1] Jean Berstel & Michel Pocchiola (1993): *A geometric proof of the enumeration formula for Sturmian words*. Internat. J. Algebra Comput. 3(3), pp. 349–355, doi:10.1142/S0218196793000238.

[2] J. Cassaigne, P. Hubert & S. Troubetzkoy (2002): *Complexity and growth for polygonal billiards*. Ann. Inst. Fourier (Grenoble) 52(3), pp. 835–847. Available at `http://aif.cedram.org/item?id=AIF_2002__52_3_835_0`.

[3] Julien Cassaigne (1997): *Complexité et facteurs spéciaux*. Bull. Belg. Math. Soc. Simon Stevin 4(1), pp. 67–88. Available at `http://projecteuclid.org/getRecord?id=euclid.bbms/1105730624`. Journées Montoises (Mons, 1994).

[4] E. P. Lipatov (1982): *A classification of binary collections and properties of homogeneity classes*. Problemy Kibernet. (39), pp. 67–84.

[5] M. Lothaire (2002): *Algebraic combinatorics on words*. Encyclopedia of Mathematics and its Applications 90, Cambridge University Press, Cambridge. Chapter 3, *Sturmian Words* (by Jean Berstel and Patrice Séébold).

[6] Filippo Mignosi (1991): *On the number of factors of Sturmian words*. Theoret. Comput. Sci. 82(1, Algorithms Automat. Complexity Games), pp. 71–84, doi:10.1016/0304-3975(91)90172-X.

[7] Thierry Monteil (2011): *Another Definition for Digital Tangents*. In: DGCI, Lecture Notes in Computer Science 6607, pp. 95–103, doi:10.1007/978-3-642-19867-0_8.

[8] N. Pytheas Fogg (2002): *Substitutions in dynamics, arithmetics and combinatorics*. Lecture Notes in Mathematics 1794, Springer-Verlag, Berlin, doi:10.1007/b13861. Chapter 6, *Sturmian Sequences* (by Pierre Arnoux).