Unequal letter frequencies in ternary square-free words

Pascal Ochem
LaBRI – Université Bordeaux 1
351 cours de la Libération
33405 Talence Cedex FRANCE
ochem@labri.fr

September 11, 2007

Abstract

We consider the set S of triples (x, y, z) corresponding to the frequency of each alphabet letter in some infinite ternary square-free word (so $x + y + z = 1$). We conjecture that this set is convex. We obtain bounds on S by with a generalization of our method to bound the extremal frequency of one letter. This method uses weights on the alphabet letters. Finally, we obtain positive results, that is, explicit triples in S lying close to its boundary.

1 Introduction and preliminary results

A *square* is a repetition of the form xx, where x is a nonempty word; an example in English is `hotshots`.
Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. It is easy to see that every word of length ≥ 4 over Σ_2 must contain a square, so squares cannot be avoided in infinite binary words. However, Thue showed [1, 9, 10] that there exist infinite words over Σ_3 that avoid squares.

Let $|w|_i$ denote the number of occurrences of the letter i in the finite word w. The frequency of the letter i in the finite word w is thus $\frac{|w|_i}{|w|}$. In the case of infinite words, we say that the letter i has frequency q in the infinite word w if for every $\epsilon > 0$, there exists an integer n_ϵ such that for every factor v of length at least n_ϵ, $\left|\frac{|v|_i}{|v|} - q\right| < \epsilon$.

Various authors have considered letter frequencies [2, 4, 5, 6, 7, 8] in infinite words avoiding some repetition. Most results concern the minimal or maximal frequency of one letter. The aim of this paper is to present an extension of our methods [6] to the general case, that is, when the frequency of every letter in Σ_k is considered. Of course, the case of the binary alphabet is irrevelant in this

context since minimizing the frequency of one letter is the same as maximizing the frequency of the other. We chose to study ternary square-free words.

Let S denote the set of triples (x, y, z) corresponding to the frequency of each alphabet letter in some infinite ternary square-free word. We thus have $x+y+z = 1$. We conjecture that S is convex, that is, if (x, y, z) and (x', y', z') both belong to S then $(x+t(z'-z), y+t(z'-z), z+t(z'-z))$ belongs to S for every $t \in [0, 1]$.

Using symmetries between the alphabet letters, we focus on the triples of S of the form $(x, y, 1-x-y)$ such that $x \leq y \leq 1-x-y$ or on the corresponding set S' of points (x, y) satisfying

$$x \leq y \tag{1}$$

and

$$x + 2y \leq 1. \tag{2}$$

In the following, x and y denote coordinates of points in S'. The first results on S' are obtained from the known bounds on the minimal frequency $f_{\min}$ and the maximal frequency $f_{\max}$ of a letter in an infinite ternary square-free word.

Theorem 1.

- *[4]* $f_{\min} = \frac{883}{3215}$.
- *[6]* $f_{\max} = \frac{255}{653}$.

From the proofs of Theorem 1 we deduce, in our notations, that

Corollary 2.

- $$x \geq \frac{883}{3215} \tag{3}$$

- $$x + y \geq 1 - \frac{255}{653} = \frac{398}{653} \tag{4}$$

- *S' contains the point* $P_1 = \left(\frac{883}{3215}, \frac{1166}{3215}\right)$.
- *S' contains the point* $P_2 = \left(\frac{199}{653}, \frac{199}{653}\right)$.

The square-free morphism $0 \mapsto 012,\ 1 \mapsto 02,\ 2 \mapsto 1$ shows that S' contains the point $P_0 = \left(\frac{1}{3}, \frac{1}{3}\right)$.

To our knowledge every infinite ternary square-free word constructed in the litterature correspond to a point inside the triangle (P_0, P_1, P_2). This triangle, which is inside the convex hull of S', already occupies most of the area of the region bounded by equations 1 to 4, which contains S'.

In Section 2, we give a way to obtain new bounds on S'. In Section 3, we obtain two new points in S' that extend the triangle (P_0, P_1, P_2). The C sources of the programs used in this paper are available at
`http://dept-info.labri.fr/~ochem/morphisms/` .

2 Negative results

We extend the method [6] by putting weights on the alphabet letters. A *weight function* $\omega : \Sigma_k^* \to \mathbb{R}$ is a mapping satisfying $\omega(uv) = \omega(u) + \omega(v)$ for every $u,\ v \in \Sigma_k^*$. It is thus completely defined by the k-tuple $(\omega(0), \ldots, \omega(k-1))$. The *average weight* of a finite word $w \in \Sigma_k^*$ is $\alpha_\omega(w) = \frac{\omega(w)}{|w|}$. Alternatively, we have $\alpha_\omega(w) = \sum_{i=0}^{k-1} \frac{|w|_i}{|w|}\omega(i)$, which allows to extend the definition of the average weight to infinite words such that the frequency of every alphabet letter is defined.

A word w is said to be t-biprolongable in a factorial language L if there exists a word $lwr \in L$ such that $|l| = |r| = t$. A *suffix cover* of L is a set S of finite words in L such that every finite word that is t-biprolongable in L and of length at least $\max_{u\in S} |u|$ has a suffix that belongs to S, for some finite number t. Taking $t = 20$ is sufficient for every negative result in this paper. For a weight function ω and a word $u \in S$, let

$$A_{\omega,u}(q) = \{w \in L \mid uw \in L \text{ and for every prefix } w' \text{ of } w,\ \alpha_\omega(w') < q\}.$$

Lemma 3. *Let L be a factorial language and S one of its suffix covers. Let $q \in \mathbb{Q}$. If $A_{\omega,u}(q)$ is finite for every word $u \in S$, then $\alpha_\omega(w) \geq q$ for every infinite word $w \in L$.*

Proof. Assume $A_{\omega,u}(q)$ is finite for every word $u \in S$. We show that every right-infinite word $w \in L$ has a decomposition into finite factors $w = pv_0v_1v_2v_3\cdots$ such that $|p| = t$, $|v_0| = \max_{u\in S} |u|$, and $\alpha_\omega(v_i) \geq q$ for every $i \geq 1$. Notice that for every $i \geq 0$, the factor $f_i = v_0 \cdots v_i$ is t-biprolongable in L and is such that $|f_i| \geq \max_{u\in S} |u|$. Thus, for every $i \geq 0$, f_i has a suffix $s_i \in S$, and since $A_{\omega,s_i}(q)$ is finite, there exists a finite factor v_{i+1} at the right of f_i such that $\alpha_\omega(v_{i+1}) \geq q$. □

It is easy to see that Lemma 3 and the definition of $A_{\omega,u}(q)$ can be modified to provide bounds of the form $\alpha_\omega(w) > q$.

For two weight functions ω and ω' respectively defined by the k-tuples $(\omega(0), \ldots, \omega(k-1))$ and $(\omega'(0) = a\omega(0) + b, \ldots, \omega'(k-1) = a\omega(k-1) + b)$, we have $\alpha_{\omega'}(w) = a\alpha_\omega(w) + b$. Our goal is to minimize the average weight of a word in L, we thus put weights in decreasing order of frequency. In the case of ternary square-free words, we can thus consider only weight functions ω_c with weights $(0, c, 1-c)$ with $0 \leq c \leq \frac{1}{2}$. So 0 is the most frequent letter with frequency z and weight 0, letter 1 has frequency y and weight c, and 2 is the least frequent letter with frequency x and weight $1 - c$.

Lemma 3 enables us to obtain bounds on S'. We first choose $q \in Q$, a suffix cover S for ternary square-free words, and a weight functions ω_c. Then we check by computer that $A_{\omega_c,u}(q)$ is finite for every $u \in S$. This shows that for every infinite ternary square-free word w, we have $(1-c) \times x + c \times y + 0 \times z = (1-c)x + cy \geq q$. Equations 3 and 4 would have corresponded to the cases $c = 0$ and $c = \frac{1}{2}$ respectively.

There does not seem to be "smart choices" for the values of c, so we simply took c in $\left\{\frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{2}{5}\right\}$. In each case, the suffix cover was $\{0, 1, 202, 102, 2012, 21012, 0212\}$. Notice that since alphabet letters play distinct roles, there is no notion of reduced suffix cover as in [6].

Theorem 4. *If w is an infinite ternary square-free word with letter frequencies (x, y, z) such that $x \leq y \leq z$, then:*

$$y > \frac{4230}{1493} - 9x \tag{5}$$

$$y > \frac{729}{500} - 4x \tag{6}$$

$$y > \frac{380}{381} - \frac{7}{3}x \tag{7}$$

$$y > \frac{340}{447} - \frac{3}{2}x \tag{8}$$

3 Positive results

To prove that a point $(x, y) \in \mathbb{Q}^2$ belongs to S', we construct an infinite ternary square-free word w as the image of any infinite ternary square-free word by a suitable square-free morphism. We write $x = \frac{n_x}{d}$ and $y = \frac{n_y}{d}$. For increasing values of t, we look for a square-free $(t \times d)$-uniform morphism h such that $|h(i)|_2 = t \times n_x$ and $|h(i)|_1 = t \times n_y$ for every $i \in \Sigma_3$. The square-freeness of h is checked thanks a result of Crochemore [3] saying that a uniform morphism is square-free if the image of every square-free word of length 3 is square-free.

Theorem 5. *There exists infinite ternary square-free words with letter frequencies $\left(\frac{13}{45}, \frac{1}{3}, \frac{17}{45}\right)$ and $\left(\frac{2}{7}, \frac{19}{56}, \frac{3}{8}\right)$.*

Proof. Ternary square-free words with letter frequencies $\left(\frac{13}{45}, \frac{1}{3}, \frac{17}{45}\right)$ can be constructed with the following square-free (2×45)-uniform morphism

$$\begin{aligned}
0 &\to 010201202101201020121012010210121020102101201020120210120\\
&\quad 210120210201021012010201202101 2\\
1 &\to 010210121020102101201020120210201021012010201210120102101 2\\
&\quad 021020102101201020120210201210 12\\
2 &\to 010210121020102101201020120210201021012010201210201021012 0\\
&\quad 210201021012102010212010201210 12.
\end{aligned}$$

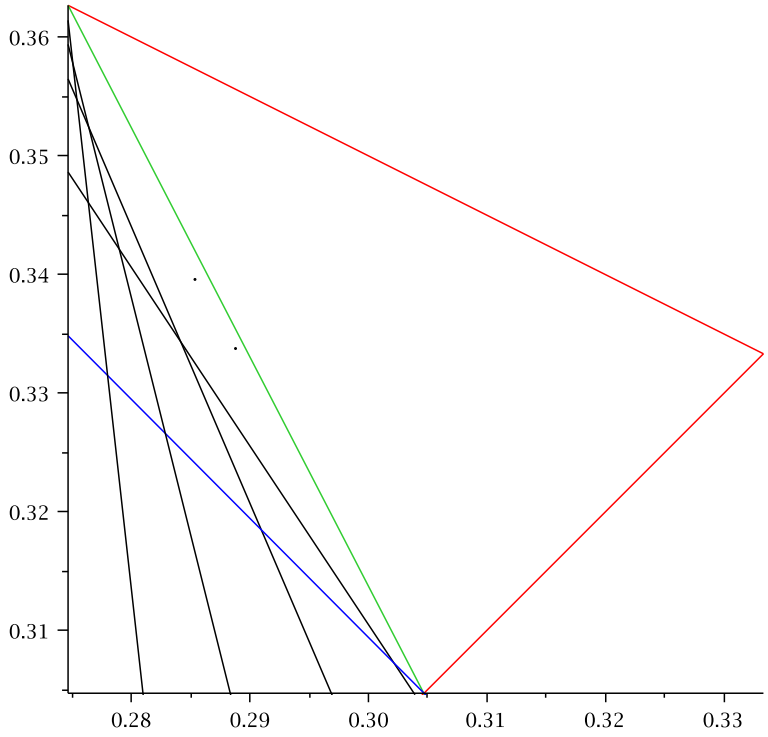
Ternary square-free words with letter frequencies $\left(\frac{2}{7}, \frac{19}{56}, \frac{3}{8}\right)$ can be con-

structed with the following square-free (2×56)-uniform morphism

$$
\begin{aligned}
0 &\to 0102012101201020120210120102012101201021012102010210120102 \\
&\quad 0120210120102101202102010210120102012021012010210121102 \\
1 &\to 0102101201020120210201021012010201210120102101210201021012 \\
&\quad 0102012021012010210121020102101201020121020102101201202 \\
2 &\to 0102101201020121012010210120210201021012010201210120102101 \\
&\quad 2102010210120102012021020102101201020121020102101201202.
\end{aligned}
$$

□

Theorem 5 means that S' also contains $P_3 = \left(\frac{13}{45}, \frac{1}{3}\right)$ and $P_4 = \left(\frac{2}{7}, \frac{19}{56}\right)$. The figure below is meant to visualize the known results about the set S'.



References

[1] J. Berstel. *Axel Thue's Papers on Repetitions in Words: a Translation.* Number 20 in Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal, February 1995.

[2] J. Chalopin and P. Ochem. Dejean's conjecture and letter frequency, *Mons Days of Theoretical Computer Science*, Rennes, August 30 - September 2 2006.

[3] M. Crochemore. Sharp characterizations of squarefree morphisms. *Theoret. Comput. Sci.* **18** (1982), 221–226.

[4] A. Khalyavin. The minimal density of a letter in an infinite ternary square-free word is $\frac{883}{3215}$. *J. Integer Sequences* 10(6):Article 07.6.5 (2007).

[5] R. Kolpakov, G. Kucherov, and Y. Tarannikov. On repetition-free binary words of minimal density, *Theoret. Comput. Sci.* **218** (1999), 161–175.

[6] P. Ochem. Letter frequency in infinite repetition-free words, *Theoret. Comput. Sci.* 380 (2007), 388–392.

[7] C. Richard and U. Grimm. On the entropy and letter frequencies of ternary square-free words, *Electron. J. Comb.* **11** (2004), #R14

[8] Y. Tarannikov. The minimal density of a letter in an infinite ternary square-free word is 0.2746..., *J. Integer Sequences* 5(2):Article 02.2.2 (2002).

[9] A. Thue. Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[10] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.