On $p$-adic avoidance of words

Svetlana Puzynina

Sobolev Institute of Mathematics, Novosibirsk
University of Turku, Finland
Turku Centre for Computer Science, Finland

Abstract

In this paper a new modification of the notion of avoidance of words is introduced. We say, that a word $w$ $p$-adically avoids a word $u$, if $u$ does not belong to the set of all words whose symbols occur in $w$ at positions which constitute arithmetical progressions with common differences equal to integer powers of $p$. We study the language of words over the binary alphabet $\Sigma = \{a, b\}$, $p$-adically avoiding the word consisting of $k$ successive letters $b$.

1 Introduction

The study of avoidable words and patterns started by A. Thue in the beginning of XX century [9]. A word $v$ avoids a word $u$, if $u$ is not a subword of $v$. A word $v$ arithmetically avoids a word $u$, if $u$ does not occur in arithmetical progressions in $v$. For some results about arithmetic avoidance of patterns see [2, 5]. We propose the notion of $p$-adic avoidance. A word $v$ $p$-adically avoids a word $u$, if $u$ does not occur in arithmetical progressions with common difference equal to an integer power of $p$ in $v$.

In this paper we study the language of words over the alphabet $\Sigma = \{a, b\}$, $p$-adically avoiding the word $b^k$, i.e., the word consisting of $k$ successive letters $b$. The complexity of the language of words avoiding $b^k$ was computed in [7]. The famous Szemerédi theorem (a generalization of Van-der-Vaerden theorem) about arithmetical progressions says that every subset of the set of integers with positive asymptotic frequency contains arithmetic
progressions of arbitrary long length [10, 8]. Szemeredi theorem implies that
the frequency of $b$’s in the infinite word arithmetically avoiding $b^k$ tends to
0. We consider the maximal frequency of $b$’s in infinite words $p$-adically
avoiding the word $b^k$. We prove, that for $p$ prime and $k > p$ asymptotic
maximal frequency of $b$’s is $1 - \frac{1}{k}$ in the case $k$ is not divisible by $p$ and
$1 - \frac{1}{k-1}$ in the case $k$ is divisible by $p$. For some related results on minimal
density of letters for pattern avoidance see [6].

2 Definitions and notation

Let $\Sigma$ be a finite alphabet. As usual, the set of all finite words over $\Sigma$
is denoted by $\Sigma^*$, the set of all right infinite words by $\Sigma^\omega$. A word $u \in \Sigma^*$
is called a subword, or a factor of $w \in \Sigma^* \cup \Sigma^\omega$, if $w = vut$ for some words $v$ and $t$
which may be empty.

Consider a finite or infinite word $w$ over a finite alphabet $\Sigma$. The
subword closure $F(w)$ of the word $w$ is the set of all its subwords. The word $w$
avoids a word $u$, if $F(w)$ does not contain $u$. Denote by $L(u)$ the language of all
finite words avoiding $u$.

Define the operator $A^n_{i,d}$ of arithmetical cutting in the following way:
$A^n_{i,d}(w) = w_i w_{i+d} w_{i+2d} w_{i+3d} \ldots w_{i+(n-1)d}$. This operator cuts from $w$ a word
by arithmetical progression of length $n$ with common difference $d$ starting
from $i$. The arithmetical closure $A(w)$ of $w$ is the set of all its arithmetical
cuttings. The word $w$ arithmetically avoids a word $u$, if $A(w)$ does not contain $u$. Denote by $L_A(u)$ the language of all finite words arithmetically
avoiding $u$. Notice that this language is arithmetically closed, i.e., $L_A(u) = A(L_A(u))$.

Define the operator $P^n_{i,l}$ of $p$-adic cutting to be the operator of arith-
metical cutting for $k = p^l$: $P^n_{i,d} = A^n_{i,p^l}$. The $p$-adic closure $P(w)$ of a
word $w$ is the set of all its $p$-adic cuttings. The word $w$ avoids $p$-adically
a word $u$ if $P(w)$ does not contain $u$. Denote by $L_p(u)$ the language of all
finite words $p$-adically avoiding $u$. Notice that this language is $p$-adically
closed, i.e., $L_p(u) = P(L_p(u))$. It is clear that for every word $u$ it holds
$L_A(u) \subseteq L_p(u) \subseteq L(u)$.

The number of letters $a$ in a word $v \in \Sigma^*$ is denoted by $|a|_v$, as usual.
The prefix of length $n$ of a word $w \in \Sigma^* \cup \Sigma^\omega$ is denoted by $\text{pref}_n(w)$.

Let $w$ be an infinite word in a finite alphabet $\Sigma$. If
\[
\lim_{n \to \infty} \frac{|a|_{\text{pref}_n(w)}}{n}
\]
exists and equals $r$, then the frequency of the letter $a$ is defined as $r$ and denoted by $\text{Freq}_w(a)$.

Let $M$ be a set of infinite words in a finite alphabet $\Sigma$, $M'$ be a subset of words in $M$ such that the frequencies of $a$ exist in these words. Asymptotic maximal frequency $\text{Freq}_{\text{max}}(a)$ of a letter $a \in \Sigma$ in the set $M$ is defined as follows:

$$\text{Freq}_{\text{max}}(a) = \sup_{w \in M'} \text{Freq}_w(a).$$

The subword complexity (or briefly complexity) $f_w(n)$ of an infinite word $w$ is the number of its subwords of length $n$ [3]:

$$f_w(n) = \sharp\{w_iw_{i+1} \ldots w_{i+n-1}| i \geq 1\}.$$

The complexity $f_L(n)$ of a language $L$ is defined to be the function counting the number of words of length $n$ in this language. There exist other modifications of the notion of complexity, for example, maximal pattern complexity [4], arithmetical complexity [1].

Denote by $b^k$ a word that consists of $k$ letters $b$ staying one after another. In this paper we study a binary language $L^k_p$ that avoids $p$-adically a word $b^k$, i.e., $L^k_p = L_p(b^k)$.

### 3 Frequency

In this section we study frequencies of letters in infinite words $p$-adically avoiding $b^k$.

**Theorem 1** Let $k$ be an integer, $p$ prime, $k \neq p$. Then asymptotic maximum frequency of $b$ in infinite binary words $p$-adically avoiding $b^k$ is equal to

$$\text{Freq}_{\text{max}}(b) = \begin{cases} 1 - \frac{1}{k}, & \text{if } k \text{ is not divisible by } p, \\ 1 - \frac{1}{k^l}, & \text{if } k \text{ is divisible by } p. \end{cases}$$

**Proof.** 1. Consider the case $k$ is not divisible by $p$. A word on which the frequency $1 - \frac{1}{k}$ is achieved is $w = (b^{k-1}a)^\omega$. It is clear that the frequency cannot be greater, because every subword of length $k$ contains at least one $a$. Therefore it is sufficient to prove that elements of arithmetical progression of length $k$ with common difference $p^l$ have different residues of division by $k$. Then one of the elements in such progression in $w$ is $a$.

3
Consider arithmetical progression \(i, i + p^l, i + 2p^l, \ldots, i + (k - 1)p^l, i \geq 1, k \geq 0\). Suppose that its elements \(i + j_1p^l\) and \(i + j_2p^l\) have the same residues, i.e., \(i + j_1p^l \equiv i + j_2p^l \equiv 0(\text{mod } k)\). Then \((j_1 - j_2)p^l \equiv 0(\text{mod } k)\). This is impossible since \(0 \leq j_1, j_2 \leq k - 1\) and \(j_1 \neq j_2\).

2. Consider the case \(k\) is divisible by \(p\), \(k \neq p\). A word on which the frequency \(1 - \frac{1}{k} - 1\) is achieved is \(w = (b^{k-2}a)\). Since \(k\) is divisible by \(p\), we have that \(k - 1\) is not divisible by \(p\). We proved that in this case \(w\) \(p\)-adically avoids \(b^{k-1}\), thus it avoids \(b^k\). So it remains to prove that the frequency cannot be greater.

Claim 1. A frequency of a letter in an infinite word is equal to a frequency of this letter in every suffix of this word.

Claim 2. If \(\text{Freq}_n(a) > \rho\), then there exists \(N_0 \in \mathbb{N}\) such that for every \(n > N_0\) it holds \(\frac{|a|_{\text{pref}_n(u)}}{n} > \rho\).

Proofs of these claims follow from definition and simple properties of limit.

Claim 3. Let \(v\) be an infinite word, \(\text{Freq}_n(a) > \rho\). Then there exists a suffix \(u\) of the word \(v\) such that frequency of \(a\)'s in every prefix of \(u\) is greater than \(\rho\).

Proof of Claim 3. Take \(N_1 = \min\{N_0|\forall n > N_0 \frac{|a|_{\text{pref}_n(u)}}{n} > \rho\}\), i.e., \(N_1\) is equal to minimal \(N_0\) from Claim 2. In particular, \(\frac{|a|_{\text{pref}_N(u)}}{N_1} \leq \rho\) if only \(N_1 \neq 0\). Required suffix \(u\) of \(v\) is given by the formula: \(u = v_{N_1+1}v_{N_1+2} \ldots\). Suppose that there exists a prefix of \(u\) such that the frequency of \(a\)'s in it is less than or equal to \(\rho\), i.e., there exists \(N_2 \in \mathbb{N}\) such that \(\frac{|a|_{\text{pref}_{N_2}(u)}}{N_2} \leq \rho\). Consider a prefix of \(v\) of length \(N_1 + N_2\). We have \(\frac{|a|_{\text{pref}_{N_1}(v)}}{N_1} \leq \rho\) and \(\frac{|a|_{v_{N_1+1}v_{N_1+2} \ldots}}{N_2} \leq \rho\), so \(\frac{|a|_{\text{pref}_{N_1+N_2}(v)}}{N_1+N_2} \leq \rho\). This contradicts the choice of \(N_1\).

Claim 3 is proved.

Now we proceed to the proof of the theorem in the second case. Suppose that \(v\) is an infinite word \(p\)-adically avoiding \(b^k\), and frequency of \(b\)'s in it is greater than \(1 - \frac{1}{k} - 1\), i.e., \(\text{Freq}_n(b) > 1 - \frac{1}{k} - 1\). By Claim 3 we have that there exists a suffix \(u\) of \(v\) such that frequency of \(b\)'s in every prefix of \(u\) is greater than \(1 - \frac{1}{k} - 1\). By Claim 1 we have that \(\text{Freq}_n(u) = \text{Freq}_n(b)\).

Now we are going to find out how \(u\) can look like. First, \(u\) should start with \(b^{k-1}\), because otherwise it will have a prefix with \(b\)'s frequency less than or equal to \(1 - \frac{1}{k} - 1\). Secondly, \(u_k = a\), because otherwise \(\text{pref}_{k}(u) = b^k\),
and this is impossible, because $u$ avoids $b^k$. Then we have $u_{k+j} = b$ for $j = 1, \ldots, k - 2$, otherwise $\text{Freq}_{\text{pref}_{k+j}(u)}(b) \leq 1 - \frac{1}{k-1}$. Exactly one of the elements $u_{2k-1}$, $u_{2k}$ equals $a$. They cannot be both $a$, because in this case $\text{Freq}_{\text{pref}_{2k}(u)}(b) = \frac{2k-3}{2k} \leq 1 - \frac{1}{k-1}$ (note that in the case we consider now $k \neq 2$, because $k$ is divisible by $p$ and $k \neq p$), and they cannot be both $b$, because in this case $u_{k+1} \ldots u_{2k} = b^k$. Continuing this line of reasoning, we obtain that $u_{ik+j} = b$ for $i = 1, \ldots, k - 2$, $j = 1, \ldots, k - i - 1$ and for every $i$ in this range exactly one element in a subword $u_{ik+k-i} \ldots u_{ik+k}$ equals $a$.

So the prefix of $u$ looks as follows:

$$u = b^{k-1}a b^{k-2}(.)2b^{k-3}(.)3 \ldots b(\cdot)^{k-1} \ldots, \quad (1)$$

where $(.)^l$ denotes a sequence of length $n$, such that exactly one of these $l$ elements is equal to $a$, all others are equal to $b$.

Now consider the arithmetical cutting $A^k_{i,p}$: $u_{p+1+jp}$, $l = 0, \ldots, k - 1$. In this progression all elements are equal to $b$ except $u_{kp-p+1}$, which belongs to the series $u_{pk-p+1} \ldots u_{pk} = (.)^p$. The word $u$ $p$-adically avoids $b^k$, so $u_{kp-p+1} = a$ and $u_{kp-j} = b$ for $j = 0, \ldots, p - 2$.

Consider the arithmetical progression of length $k$ with common difference $p$ starting in the second position: $u_{2+jp}$, $l = 0, \ldots, k - 1$. In this progression all elements are equal to $b$ except $u_{k(p-2)+2-p}$, which belongs to the series $u_{k(p-2)+2-p} \ldots u_{k(p-2)} = (.)^{p-1}$. The word $u$ $p$-adically avoids $b^k$, so $u_{k(p-2)+2-p} = a$ and $u_{k(p-2)-j} = b$ for $j = 0, \ldots, p - 3$.

Continue this line of reasoning by considering arithmetical cuttings $A^k_{i,p}$: $u_{i+jp}$, $l = 0, \ldots, k - 1$ for $i = 2, \ldots, p - 1$. In this progressions all elements except $u_{k(p-i+1)-p+i}$ are equal to $a$, the element $u_{k(p-i+1)-p+i}$ belongs to the series $u_{k(p-i+1)-p+i} \ldots u_{k(p-i+1)} = (.)^{p-i+1}$. The word $u$ $p$-adically avoids $b^k$, so $u_{k(p-i+1)-p+i} = a$ and $u_{k(p-i+1)-j} = b$ for $j = 0, \ldots, p - i - 1$.

Therefore, the prefix of $u$ looks as follows:

$$u = b(b^{k-2}a)pb^{p-1}b^{k-1}(.)^{p+1}b^{k-p-2}(.)^{p+2} \ldots b(.)^{k-1} \ldots$$

The word $u$ avoids $b^k$, so $u_{2+(k-1)p} \ldots u_{2+(k-1)p+k-1} \neq b^k$. It follows that one of the elements $u_{2+(k-1)p+k-1}$ and $u_{2+(k-1)p+k+2}$ is equal to $a$, therefore, $u_{2+(k-1)p+k+2} = b$, $u_{2+(k-1)p+k+2} = b$.

Next, $u_{2+(k-1)p+k} \ldots u_{2+(k-1)p+2k-1} \neq b^k$, therefore one of the elements $u_{2+(k-1)p+2k-3} \ldots u_{2+(k-1)p+2k-2} \neq b^k$. It follows that $u_{2+(k-1)p+2k-2} = b$, $u_{2+(k-1)p+2k-2} = b$.

Arguing as above, we obtain that for $i = 1, \ldots, k - p - 1$ the condition $u_{2+(k-1)p+ik} \ldots u_{2+(k-1)p+(i+1)k-1} \neq b^k$ implies that one of the elements $u_{2+(k-1)p+ik-1} \ldots, u_{2+(k-1)p+ik}$ is equal to $a$, so $u_{2+(k-1)p+ik+1} = b$, $\ldots$.5
$u_{2+(k-1)p+ik+p-1} = b$. Actually, using obtained $b$’s instead of elements in each series $(.)^j$, we find that in the next series $(.)^{j-1}$ last $(p-1)$ elements are equal to $b$’s:

$$u = b(b^{k-2}a)^p b^{k-2} (.)^2 b^{k-3} (.)^3 \ldots b^p (.)^{k-p} p^{-1} \ldots$$

The frequency of $b$’s in all prefixes of $u$ is greater than $1 - \frac{1}{k-1}$, hence $\text{Freq}_{\text{pref}_{k^2-p+1}(u)}(b) > 1 - \frac{1}{k-1}$, so the subword $u_{k^2-k+1} \ldots u_{k^2-p+1}$ contains at most one $a$. Combining this with the fact that $u_{k^2-k-p+2} \ldots u_{k^2-p+1} \neq b^k$, we get that $u_{k^2-k+1} \ldots u_{k^2-p+1} = (.)^{k-p+1}$, i.e., this subword contains exactly one $a$.

We have $\text{Freq}_{\text{pref}_{k^2-1}(u)}(b) > 1 - \frac{1}{k-1}$, therefore, $u_{k^2-p+2} \ldots u_{k^2-1} = b^{p-2}$. Combining frequency considerations in $\text{pref}_{k^2+k-p+1}(u)$ with the fact that $u_{k^2-p+2} \ldots u_{k^2+k-p+1} \neq b^k$, we get that $u_{k^2} \ldots u_{k^2+k-p+1} = (.)^{k-p+2}$, i.e., this subword contains exactly one $a$.

Using similar considerations of frequencies of $b$’s and avoidance of the word $b^k$, we obtain that the prefix of the word $u$ is of the following form:

$$u = b(b^{k-2}a)^p b^{k-2} (.)^2 b^{k-3} (.)^3 \ldots b(.)^{k-1} \ldots$$

The suffix of $u$ starting in $u_{2+(k-1)(p-1)}$ is the same as the suffix of $u$ starting in $u_2$ that we obtained in (1). So we can apply the same argument to this shifted word, because we did not use the fact that $u_1 = b$. Therefore, we have that

$$u = b(b^{k-2}a)^p b^{k-2} (.)^2 b^{k-3} (.)^3 \ldots b(.)^{k-1} \ldots$$

We proceed in the same way to infinity and obtain that $u = b(b^{k-2}a)^\omega$. It is easy to see that $\text{Freq}_u(b) = 1 - \frac{1}{k-1}$. Claim 1 says that $\text{Freq}_u(b) = \text{Freq}_v(b)$. A contradiction.

\textbf{Remark 1.} Asymptotic maximal frequencies of $b$’s in words $p$-adically avoiding $b^k$ coincide for $k = mp - 1$ and $k = mp$, where $m > 1$. It is quite surprising, because there are less restrictions in the first case.

\textbf{Remark 2.} The theorem holds for every finite alphabet, not necessarily binary. In the proof of the first case of the theorem we actually used only the fact that $p$ and $k$ are coprime, not necessarily that $p$ is prime.

\textbf{Remark 3.} In the case $k = p$ our technique does not work. It is clear, that in this case $1 - \frac{1}{k-1} \leq \text{Freq}_{\text{max}}(b) \leq 1 - \frac{1}{k}$. We suppose that the equality in the left inequality is achieved, except for the case $k = p = 2$ (see the section Examples), and this case seems to be the only exception.
4 Complexity

In this section we obtain some bounds on the complexity of the language $L^k_p$. Notice that if we replace some of $b$’s by $a$’s in a word $p$-adically avoiding $b^k$, then the new word also $p$-adically avoids $b^k$. So we obtain a lower bound on the complexity of the language $L^k_p$:

**Corollary 1** The complexity function $f_{L^k_p}(n)$ of the language $L^k_p$, where $p$ is prime, $k \geq 2$, $k \neq p$, satisfies the following inequalities:

1) $f_{L^k_p}(n) \geq 2^{(1-\frac{1}{k})n}$, if $k$ is not divisible by $p$,

2) $f_{L^k_p}(n) \geq 2^{(1-\frac{1}{k+1})n}$, if $k$ is divisible by $p$.

Notice that asymptotic maximal frequencies of $b$’s in words $p$-adically avoiding $b^k$ coincide for $k = mp - 1$ and $k = mp$. While the combinatorial complexity for $k = mp - 1$ is obviously greater than for $k = mp$: $L^m_p \subseteq L^{mp-1}_p$.

Trivial upper bounds for the complexity can be obtained as follows. The number of words of small length can be easily counted “by hand”. If the number of words of length $l$ in $L^k_p$ is equal to $m$, then we have the following upper bound for the complexity function of the language $L^k_p$: $f_n(L^k_p) \leq m^n$.

5 Examples and concluding remarks

**Example 1.** Consider the frequency of $b$’s in infinite binary words 2-adically avoiding $b^2$. If Theorem 1 held for this case, the asymptotic maximal frequency would be 0, though in fact it is equal to $1/3$. Indeed, the asymptotic maximum frequency of $b$’s in an infinite word 2-adically avoiding $b^2$ cannot exceed $1/3$, because every word of length 3 can contain at most one $b$. On the other hand, the word $(aab)\omega$ 2-adically avoids $b^2$, because the distance between every two $b$’s is divisible by $p$, so it could not be a power of 2.

**Example 2.** Now consider another example. Let $L_2(a^3, b^3)$ be the language of words 2-adically avoiding the words $a^3$ and $b^3$ simultaneously. If we substitute $b$ by $a$ in a word in this language, in the new word there may appear arithmetical progressions with differences $2^l$ consisting of $a$’s. Therefore the maximal asymptotic frequency of $b$’s does not give lower bound for the complexity of the language $L_2(a^3, b^3)$.

Let $*$ denote a letter that could equal either $a$ or $b$. Then the words of the form $(ab*)\omega$ 2-adically avoid $a^3$ and $b^3$. Therefore, the complexity of the language $L_2(a^3, b^3)$ is not less than $2^{n/3}$. 

Note that the same construction can be applied to the language $L_p(a^k, b^k)$, where $k$ and $p$ are coprime: words of the form $(a b^{k-2})^\omega$ $p$-adically avoid $a^k$ and $b^k$. Elements of arithmetic progression of length $k$ with difference $p^l$ have different residues by division by $k$, so every such progression contains one $a$, one $b$ and $(k - 2)$ symbols $*$.  

**Example 3.** Consider a word $(a b^*)^\omega$ from example 2. We modify it as follows: replace $w_1$ from $a$ to $b$, all letters $w_{1+2^i}$ which are equal to $*$ replace by $a$. The word we obtained 2-adically avoids $a^3$ and $b^3$. The same construction can be applied to every element instead of $w_1$, as well as to several $a$-letters (or to several $b$-letters) in $w$ simultaneously.  

This construction can also be easily generalized for the words $p$-adically avoiding $a^k$ and $b^k$, $k$ and $p$ coprime. Replace $w_1$ from $a$ to $b$, in every progression $P_{1,l}^k$ replace one of the letters which are equal to $*$ by $a$.  

**Example 4.** Consider the Thue-Morse word, a fixed point of morphism $\varphi$: $\varphi(a) = ab$, $\varphi(b) = ba$ [9]. This word 2-adically avoids $a^3$ and $b^3$. Indeed, its arithmetic cuttings with differences $2^i$ are subwords of the Thue-Morse word, and Thue-Morse word is cube-free, so it does not contain subwords $a^3$ and $b^3$.  

In this paper we introduced a new notion of $p$-adic avoidance and considered the maximal frequency of $b$’s in words $p$-adically avoiding $b^k$ for all cases except $k = p$. It would be interesting to describe completely this language, in particular, compute its complexity and find maximal frequency for the remaining case.  

The notion of $p$-adic avoidance poses a series of natural problems. It would be interesting to find out which patterns and sets of words are avoidable, to describe corresponding languages.  

**Acknowledgements**

The author is grateful to S. V. Avgustinovich for useful discussions and attention to this work.

**References**


