Symposium on Theoretical Aspects of Computer Science 2008 (Bordeaux), pp. 421-432
www.stacs-conf.org

THE FROBENIUS PROBLEM IN A FREE MONOID

JUI-YI KAO [1], JEFFREY SHALLIT [1], AND ZHI XU [1]

[1] School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, CANADA
E-mail address: JY.academic@gmail.com
E-mail address, Jeffrey Shallit: shallit@graceland.uwaterloo.ca
E-mail address, Zhi Xu: z5xu@cs.uwaterloo.ca

Abstract. The classical Frobenius problem over $\mathbb{N}$ is to compute the largest integer g not representable as a non-negative integer linear combination of non-negative integers $x_1, x_2, \ldots, x_k$, where $\gcd(x_1, x_2, \ldots, x_k) = 1$. In this paper we consider novel generalizations of the Frobenius problem to the noncommutative setting of a free monoid. Unlike the commutative case, where the bound on g is quadratic, we are able to show exponential or subexponential behavior for several analogues of g, with the precise bound depending on the particular measure chosen.

1. Introduction

Let $x_1, x_2, \ldots, x_k$ be positive integers. It is well-known that every sufficiently large integer can be written as a non-negative integer linear combination of the x_i if and only if $\gcd(x_1, x_2, \ldots, x_k) = 1$. The famous *Frobenius problem* (so-called because, according to Brauer [2], "Frobenius mentioned it occasionally in his lectures") is the following:

Given positive integers $x_1, x_2, \ldots, x_k$ with $\gcd(x_1, x_2, \ldots, x_k) = 1$, find the largest positive integer $g(x_1, x_2, \ldots, x_k)$ which *cannot* be represented as a non-negative integer linear combination of the x_i.

Although it seems simple at first glance, the Frobenius problem on positive integers has many subtle and intriguing aspects that continue to elicit study. A recent book by Ramírez Alfonsín [23] lists over 400 references on this problem. Applications to many different fields exist: to algebra [19]; the theory of matrices [11], counting points in polytopes [1]; the problem of efficient sorting using Shellsort [17], the theory of Petri nets [25]; the liveness of weighted circuits [8]; etc.

Generally speaking, research on the Frobenius problem can be classified into three different areas:

1998 ACM Subject Classification: F.4.3.
Key words and phrases: combinatorics on words, Frobenius problem, free monoid.
Reseach supported by a grant from NSERC.

- Formulas or algorithms for the exact computation of $g(x_1, \ldots, x_k)$, including formulas for g where the x_i obey certain relations, such as being in arithmetic progression;
- The computational complexity of the problem;
- Good upper or lower bounds on $g(x_1, \ldots, x_k)$.

For $k = 2$, it is folklore that

$$g(x_1, x_2) = x_1 x_2 - x_1 - x_2; \tag{1.1}$$

this formula is often attributed to Sylvester [24], although he did not actually state it. Eq. (1.1) gives an efficient algorithm to compute g for two elements. For $k = 3$, efficient algorithms have been given by Greenberg [15] and Davison [10]; if $x_1 < x_2 < x_3$, these algorithms run in time bounded by a polynomial in $\log x_3$. Kannan [18] gave a very complicated algorithm that runs in polynomial time in $\log x_k$ if k is fixed, but is wildly exponential in k. However, Ramírez Alfonsín [22] proved that the general problem is NP-hard, under Turing reductions, by reducing from the integer knapsack problem. So it seems very likely that there is no simple formula for computing $g(x_1, x_2, \ldots, x_k)$ for arbitrary k. Nevertheless, recent work by Einstein, Lichtblau, Strzebonski, and Wagon [12] shows that in practice the Frobenius number can be computed relatively efficiently, even for very large numbers, at least for $k \leq 8$.

Another active area of interest is estimating how big g is in terms of $x_1, x_2, \ldots, x_k$ for $x_1 < x_2 < \cdots < x_k$. It is known, for example, that $g(x_1, x_2, \ldots, x_k) < x_k^2$. This follows from Wilf's algorithm [26]. Many other bounds are known.

One can also study variations on the Frobenius problem. For example, given positive integers $x_1, x_2, \ldots, x_k$ with $\gcd(x_1, x_2, \ldots, x_k) = 1$, what is the number $f(x_1, x_2, \ldots, x_k)$ of positive integers not representable as a non-negative integer linear combination of the x_i? Sylvester, in an 1884 paper [24], showed that $f(x_1, x_2) = \frac{1}{2}(x_1 - 1)(x_2 - 1)$.

Our goal in this paper is to generalize the Frobenius problem to the setting of a free monoid. In this framework, we start with a finite, nonempty alphabet Σ, and consider the set of all finite words Σ^*. Instead of considering integers $x_1, x_2, \ldots, x_k$, we consider words $x_1, x_2, \ldots, x_k \in \Sigma^*$. Instead of considering linear combinations of integers, we instead consider the languages $\{x_1, x_2, \ldots, x_k\}^*$ and $x_1^* x_2^* \cdots x_k^*$. Actually, we consider several additional generalizations, which vary according to how we measure the size of the input, conditions on the input, and measures of the size of the result. For an application of the noncommutative Frobenius problem, see Clément, Duval, Guaiana, Perrin, and Rindone [9].

In sections 2 and 3, we introduce the definition of the generalized Frobenius problem. In sections 4 and 5, we discuss the state complexity of this generalized problem. In sections 5 and 6, we will discuss the longest length and number of omitted words, respectively.

In order to motivate our definitions, we consider the easiest case first: where $\Sigma = \{\mathtt{0}\}$, a unary alphabet.

2. The unary case

Suppose $x_i = \mathtt{0}^{a_i}$, where $a_i \in \mathbb{N}$ for $1 \leq i \leq k$. The Frobenius problem is evidently linked to many problems over unary languages. It figures, for example, in estimating the size of the smallest DFA equivalent to a given NFA [7].

If $L \subseteq \Sigma^*$, by $\overline{L}$ we mean $\Sigma^* - L$, the complement of L. If L is a finite language, by $|L|$ we mean the cardinality of L. Evidently we have

Proposition 2.1. *Suppose* $x_i = \mathtt{0}^{a_i}$ *where* $a_i \in \mathbb{N}$ *for* $1 \leq i \leq k$, *and write* $S = \{x_1, x_2, \ldots, x_k\}$. *Then* S^* *is co-finite if and only if* $\gcd(a_1, a_2, \ldots, a_k) = 1$. *Furthermore, if* S^* *is co-finite, then the length of the longest word in* $\overline{S^*}$ *is* $g(a_1, a_2, \ldots, a_k)$, *and* $|\overline{S^*}| = f(a_1, a_2, \ldots, a_k)$.

This result suggests that one appropriate noncommutative generalization of the condition $\gcd(a_1, a_2, \ldots, a_k) = 1$ is that $S^* = \{x_1, x_2, \ldots, x_k\}^*$ be co-finite, and one appropriate generalization of the g function is the length of the longest word not in S^*.

But there are other possible generalizations. Instead of measuring the length of the longest omitted word, we could instead consider the *state complexity* of S^*. By the state complexity of a regular language L, written $\mathrm{sc}(L)$, we mean the number of states in the (unique) minimal deterministic finite automaton (DFA) accepting L. In the unary case, this alternate measure has a nice expression in terms of the ordinary Frobenius function:

Theorem 2.2. *Let* $\gcd(a_1, a_2, \ldots, a_k) = 1$. *Then*

$$\mathrm{sc}(\{\mathtt{0}^{a_1}, \mathtt{0}^{a_2}, \ldots, \mathtt{0}^{a_k}\}^*) = g(a_1, a_2, \ldots, a_k) + 2.$$

Proof. Let $L = \{\mathtt{0}^{a_1}, \mathtt{0}^{a_2}, \ldots, \mathtt{0}^{a_k}\}^*$. Since $\gcd(a_1, a_2, \ldots, a_k) = 1$, every word of length $> g(a_1, a_2, \ldots, a_k)$ is contained in L. Thus we can accept L with a DFA having $g(a_1, \ldots, a_k)+2$ states, using a "tail" of $g(a_1, \ldots, a_k) + 1$ states and a "loop" of one accepting state. Thus $\mathrm{sc}(L) \leq g(a_1, a_2, \ldots, a_k) + 2$.

To see $\mathrm{sc}(L) \geq g(a_1, a_2, \ldots, a_k) + 2$, we show that the words $\epsilon, \mathtt{0}, \mathtt{0}^2, \ldots, \mathtt{0}^{g(a_1,\ldots,a_k)+1}$ are pairwise inequivalent under the Myhill-Nerode equivalence relation. Pick $\mathtt{0}^i$ and $\mathtt{0}^j$, $0 \leq i < j \leq g(a_1, \ldots, a_k) + 1$. Choose $z = \mathtt{0}^{g(a_1,\ldots,a_k)-i}$. Then $\mathtt{0}^i z = \mathtt{0}^{g(a_1,\ldots,a_k)} \notin L$, while $\mathtt{0}^j z = \mathtt{0}^{g(a_1,\ldots,a_k)+j-i} \in L$, since $j > i$. ∎

Corollary 2.3. *Let* $\gcd(a_1, \ldots, a_k) = d$. *Then*

$$\mathrm{sc}(\{\mathtt{0}^{a_1}, \mathtt{0}^{a_2}, \ldots, \mathtt{0}^{a_k}\}^*) = d\Big(g(a_1/d, a_2/d, \ldots, a_k/d) + 1\Big) + 1.$$

Hence it follows that $\mathrm{sc}(\{\mathtt{0}^{a_1}, \mathtt{0}^{a_2}, \ldots, \mathtt{0}^{a_k}\}^*) = O(a^2)$ for $a = \max_{1\leq i\leq k} a_i$. Furthermore, this bound is essentially optimal; since $g(n, n+1) = n^2 - n - 1$, there exist examples with $\mathrm{sc}(\{\mathtt{0}^{a_1}, \mathtt{0}^{a_2}, \ldots, \mathtt{0}^{a_k}\}^*) = \Omega(a^2)$.

3. The case of larger alphabets

We now turn to the main results of the paper. Given as input a list of words $x_1, x_2, \ldots, x_k$, not necessarily distinct, and defining $S = \{x_1, x_2, \ldots, x_k\}$, we can measure the size of the input in a number of different ways:

(a) k, the number of words;
(b) $n = \max_{1\leq i\leq k} |x_i|$, the length of the longest word;
(c) $m = \sum_{1\leq i\leq k} |x_i|$, the total number of symbols;
(d) $\mathrm{sc}(\{x_1, x_2, \ldots, x_k\})$, the state complexity of the language represented by the input.

We may impose various conditions on the input:

(i) Each x_i is defined over the unary alphabet;
(ii) $S^* = \{x_1, x_2, \ldots, x_k\}^*$ is co-finite
(iii) $k = 2$, or k is fixed.

And finally, we can explore various measures on the size of the result:

(1) $\mathcal{L} = \max_{x \in \Sigma^* - S^*} |x|$, the length of the longest word not in S^*;
(2) $\mathcal{K} = \max_{x \in \Sigma^* - x_1^* x_2^* \cdots x_k^*} |x|$, the length of the longest word not in $x_1^* x_2^* \cdots x_k^*$;
(3) $\mathcal{S} = \mathrm{sc}(S^*)$, the state complexity of S^*;
(4) $\mathcal{M} = |\Sigma^* - S^*|$, the number of words not in S^*;
(5) $\mathcal{S}' = \mathrm{sc}(x_1^* x_2^* \cdots x_k^*)$;

Clearly not every combination results in a sensible question to study. In order to study $\mathcal{L}$, the length of the longest word omitted by S^*, we clearly need to impose condition (ii), that S^* be co-finite.

We now study under what conditions it makes sense to study $\mathcal{K} = \max_{x \in \Sigma^* - x_1^* x_2^* \cdots x_k^*} |x|$, the length of the longest word not in $x_1^* x_2^* \cdots x_k^*$.

Theorem 3.1. *Let $x_1, x_2, \ldots, x_k \in \Sigma^+$. Then $Q = x_1^* x_2^* \cdots x_k^*$ is co-finite if and only if $|\Sigma| = 1$ and $\gcd(|x_1|, \ldots, |x_k|) = 1$.*

Proof. If $|\Sigma| = 1$ and $\gcd(|x_1|, \ldots, |x_k|) = 1$, then every sufficiently long unary word can be obtained by concatenations of the x_i, so Q is co-finite.

For the other direction, suppose Q is co-finite. If $|\Sigma| = 1$, let $\gcd(|x_1|, \ldots, |x_k|) = d$. If $d > 1$, Q contains only words of length divisible by d, and so is not co-finite. So $d = 1$.

Hence assume $|\Sigma| \geq 2$, and let a, b be distinct letters in Σ. Let $l = \max_{1 \leq i \leq k} |x_i|$, the length of the longest word. Let $Q' = ((a^{2l} b^{2l})^k)^+$. Then we claim that $Q' \cap Q = \emptyset$. For if none of the x_i consists of powers of a single letter, then the longest block of consecutive identical letters in any word in Q is $< 2l$, so no word in Q' can be in Q. Otherwise, say some of the x_i consist of powers of a single letter. Take any word w in Q, and count the number $n(w)$ of maximal blocks of $2l$ or more consecutive identical letters in w. Clearly $n(w) \leq k$. But $n(w') \geq 2k$ for any word w' in Q'. Thus Q is not co-finite, as it omits all the words in Q'. ∎

4. State complexity results

In this section we study the measures $\mathcal{S} = \mathrm{sc}(S^*)$, and $\mathcal{S}' = \mathrm{sc}(x_1^* x_2^* \cdots x_k^*)$. First we review previous results.

Yu, Zhuang, and Salomaa [27] showed that if L is accepted by a DFA with n states, then L^* can be accepted by a DFA with at most $2^{n-1} + 2^{n-2}$ states. Furthermore, they showed this bound is realized, in the sense that for all $n \geq 2$, there exists a DFA M with n states such that the minimal DFA accepting $L(M)^*$ needs $2^{n-1} + 2^{n-2}$ states. This latter result was given previously by Maslov [21].

Câmpeanu et al. [3, 5] showed that if a DFA with n states accepts a *finite* language L, then L^* can be accepted by a DFA with at most $2^{n-3} + 2^{n-4}$ states for $n \geq 4$. Furthermore, this bound is actually achieved for $n > 4$ for an alphabet of size 3 or more. Unlike the examples we are concerned with in this section, however, the finite languages they construct contain exponentially many words in n.

Holzer and Kutrib [16] examined the nondeterminstic state complexity of Kleene star. They showed that if an NFA M with n states accepts L, then L^* can be accepted by an NFA with $n + 1$ states, and this bound is tight. If L is finite, then $n - 1$ states suffices, and this bound is tight.

Câmpeanu and Ho [4] gave tight bounds for the number of states required to accept a finite language whose words are all bounded by length n.

Proposition 4.1.

(a) $\mathrm{nsc}(\{x_1, x_2, \ldots, x_k\}^*) \leq m - k + 1$.
(b) $\mathrm{sc}(\{x_1, x_2, \ldots, x_k\}^*) \leq 2^{m-k+1}$.
(c) *If no x_i is a prefix of any other x_j, then* $\mathrm{sc}(\{x_1, x_2, \ldots, x_k\}^*) \leq m - k + 2$.

We now recall an example providing a lower bound for the state complexity of $\{x_1, x_2, \ldots, x_k\}^*$ [13]. Let t be an integer ≥ 2, and define words as follows: $y := \mathtt{0}1^{t-1}\mathtt{0}$ and $x_i := 1^{t-i-1}\mathtt{0}1^{i+1}$ for $0 \leq i \leq t-2$. Let $S_t := \{\mathtt{0}, x_0, x_1, \ldots, x_{t-2}, y\}$.

Theorem 4.2. *S_t^* has state complexity $3t2^{t-2} + 2^{t-1}$.*

Corollary 4.3. *There exists a family of sets S_t, each consisting of $t+1$ words of length $\leq t+1$, such that $\mathrm{sc}(S_t^*) = 2^{\Omega(t)}$. If m is the total number of symbols in these words, then $\mathrm{sc}(S_t^*) = 2^{\Omega(\sqrt{m})}$.*

Using similar ideas, we can also create an example achieving subexponential state complexity for $x_1^* x_2^* \cdots x_k^*$.

Theorem 4.4. *Let y and x_i be as defined above. Let $L = (\mathtt{0}^* x_1^* x_2^* \cdots x_{n-1}^* y^*)^e$ where $e = (t+1)(t-2)/2 + 2t$. Then $\mathrm{sc}(L) \geq 2^{t-2}$.*

Proof. Define $A = \{x_0, x_1, \ldots, x_{t-2}, y, 0\}$ and $T = \{x_1, x_2, \ldots, x_{t-2}\}$. For any subset S of T, say $\{s_1, s_2, \ldots, s_j\}$ with $s_1 < s_2 < \cdots s_j$ define

$$x(S) = y x_{t-2} y x_{t-3} x_{t-2} y \cdots y x_1 x_2 \cdots x_{t-2} y x_{s_1} x_{s_2} \cdots x_{s_j} y.$$

Note that $x(S)$ contains t copies of y and at most $(t-2)(t-1)/2 + t - 2 = (t+1)(t-2)/2$ x's. Thus $|x(S)| \leq (t+1)(t + (t+1)(t-2)/2)$ and $|x(S)|_0 \leq 2t + (t+1)(t-2)/2$.

To get the bound $\mathrm{sc}(L) \geq 2^{t-2}$, we exhibit 2^{t-2} pairwise distinct words under the Myhill-Nerode equivalence relation. Let R and S be two distinct subsets of T, and without loss of generality, let $m \in R$, $m \notin S$. By the proof of [13, Theorem 13] we have $x(R)1^{t-m} \in A^*$ but $x(S)1^{t-m} \notin A^*$. Since $L \subseteq A^*$, $x(S)1^{t-m} \notin L$. It remains to see $x(R)1^{t-m} \in L$.

Since $x(R)1^{t-m} \in A^*$, there exists a factorization of $x(R)1^{t-m}$ in terms of elements of A. However, $|x(R)1^{t-m}| \leq |x(R)| + t \leq (t+1)(t + (t+1)(t-2)/2 + t)$ so any factorization of $x(R)1^{t-m}$ into elements of A contains at most $(t+1)(t-2)/2 + 2t$ copies of words other than 0. Similarly $|x(R)1^{t-m}|_0 \leq |x(R)| \leq (t+1)(t-2)/2 + 2t$, so any factorization of $x(R)1^{t-m}$ into elements of A contains at most $(t+1)(t-2)/2 + 2t$ copies of the word 0. Thus a factorization of $x(R)1^{t-m}$ into elements of A is actually contained in L. ∎

Corollary 4.5. *There exists an infinite family of tuples $(x_1, x_2, \ldots, x_k)$ where m, the total number of symbols, is $O(t^4)$, and $\mathrm{sc}(x_1^* \cdots x_k^*) = 2^{\Omega(t)} = 2^{\Omega(m^{1/4})}$.*

We now turn to an upper bound on the state complexity of S^* in the case where the number of words in S is not specified, but we do have a bound on the length of the longest word.

Theorem 4.6. *Let $S = \{x_1, x_2, \ldots, x_k\}$ be a finite set with $\max_{1 \leq i \leq k} |x_i| = n$, that is, the longest word is of length n. Then $\mathrm{sc}(S^*) \leq \frac{2}{2|\Sigma|-1}(2^n|\Sigma|^n - 1)$.*

Proof. The idea is to create a DFA $M = (Q, \Sigma, \delta, q_0, F)$ that records the last $n-1$ symbols seen, together with the set of the possible positions inside those $n-1$ symbols where the factorization of the input into elements of S could end.

The number of states in this DFA is $\sum_{0 \leq i < n} |\Sigma|^i 2^{i+1} = \frac{2}{2|\Sigma|-1}(2^n|\Sigma|^n - 1)$. ∎

5. State complexity for two words

In this section we develop formulas bounding the state complexity of $\{w, x\}^*$ and w^*x^*. Here, as usual, $g(x_1, x_2)$ denotes the Frobenius function introduced in Section 1. We need the following lemma, which is of independent interest and which generalizes a classical theorem of Fine and Wilf [14].

Lemma 5.1. *Let w and x be nonempty words. Let $y \in w\{w, x\}^\omega$ and $z \in x\{w, x\}^\omega$. Then the following conditions are equivalent:*

(a) *y and z agree on a prefix of length $|w| + |x| - \gcd(|w|, |x|)$;*
(b) *$wx = xw$;*
(c) *$y = z$.*

Furthermore, the bound in (a) is optimal, in the sense that for all pairs of lengths (m, n) there exist w, x with $(m, n) = (|w|, |x|)$ such that w^ω and x^ω agree on a prefix of length $|w| + |x| - \gcd(|w|, |x|) - 1$.

Proof. (a) $\implies$ (b): We prove the contrapositive. Suppose $wx \neq xw$. Then we prove that y and z differ at a position $\leq |w| + |x| - \gcd(|w|, |x|)$. The proof is by induction on $|w| + |x|$. The base case is $|w| = |x| = 1$ and is left to the reader.

Now assume the result is true for $|w| + |x| < k$. We prove it for $|w| + |x| = k$. If $|w| = |x|$ then y and z must disagree at the $|w|$'th position or earlier, for otherwise $w = x$ and $wx = xw$; since $|w| \leq d = |w|$, the result follows. So, without loss of generality, assume $|w| < |x|$. If w is not a prefix of x, then y and z disagree at the $|w|$'th position or earlier, and again $|w| \leq d$.

So w is a proper prefix of x. Write $x = wt$ for some nonempty word t. Now any common divisor of $|w|$ and $|x|$ must also divide $|x| - |w| = |t|$, and similarly any common divisor of both $|w|$ and $|t|$ must also divide $|w| + |t| = |x|$. So $\gcd(|w|, |x|) = \gcd(|w|, |t|)$.

Now $wt \neq tw$, for otherwise we have $wx = wwt = wtw = xw$, a contradiction. Then y begins with ww and z begins with wt. By induction (since $|w| + |t| < k$) $w^{-1}y$ and $w^{-1}z$ disagree at position $|w| + |t| - \gcd(|w|, |t|)$ or earlier. Hence y and z disagree at position $2|w| + |t| - \gcd(|w|, |t|) = d$ or earlier.

(b) $\implies$ (c): If $wx = xw$, then by the theorem of Lyndon-Schützenberger, both w and x are powers of a common word u. Hence $y = u^\omega = z$.

(c) $\implies$ (a): Trivial.

For the optimality statement, the words constructed in the paper [6] suffice. ∎

Theorem 5.2. *Let $w, x \in \Sigma^+$. Then*

$$\mathrm{sc}(\{w, x\}^*) \leq \begin{cases} |w| + |x|, & \text{if } wx \neq xw; \\ d\,(g(|w|/d, |x|/d) + 1) + 2, & \text{if } wx = xw \text{ and } d = \gcd(|w|, |x|)\ . \end{cases}$$

Proof. If $wx = xw$, then by a classical theorem of Lyndon and Schützenberger [20], we know there exists a word z and integers $i, j \geq 1$ such that $w = z^i$, $x = z^j$. Thus $\{w, x\}^* = \{z^i, z^j\}^*$. Let $e = \gcd(i, j)$. Then $L = \{z^i, z^j\}^*$ consists of all words of the form z^{ke} for $k > g(i/e, j/e)$, together with some words of the form z^{ke} for $0 \leq k < g(i/e, j/e)$. Thus, as in the proof of Corollary 2.3, we can accept L with a "tail" of $e|z|g(i/e, j/e) + 1$ states and a "loop" of $e|z|$ states. Adding an additional state as a "dead state" to absorb unused transitions gives a total of $(e|z|(g(i/e, j/e)+1)+2$ states. Since $d = e|z|$, the bound follows.

Otherwise, $xw \neq wx$. Without loss of generality, let us assume that $|w| \leq |x|$. Suppose w is not a prefix of x. Let p be the longest common prefix of w and x. Then we can write $w = paw'$ and $x = pbx'$ for $a \neq b$. Then we can accept $\{w, x\}^*$ with a transition diagram that has one chain of nodes labeled p leading from q_0 to a state q, and two additional chains leading from q back to q_0, one labled aw' and one labeled bx'. Since $a \neq b$, this is a DFA. One additional "dead state" might be required to absorb transitions on letters not mentioned. The total number of states is $|p| + 1 + |w'| + |x'| + 1 \leq |w| + |x|$.

Finally, suppose $|w| \leq |x|$ and w is a prefix of x. We claim it suffices to bound the longest common prefix between any word of $w\{w, x\}^*$ and $x\{w, x\}^*$. For if the longest common prefix is of length b, we can distinguish between them after reading $b + 1$ symbols. The $b+1$'th symbol must be one of two possibilities, and we can use back arrows in the transition diagram to the appropriate state. We may need one additional state as a "dead state", so the total number of states needed is $b + 2$. But from Lemma 5, we know $b \leq |w| + |x| - 2$. ∎

Theorem 5.3. *Let $w, x \in \Sigma^+$. Then*

$$\mathrm{sc}(w^*x^*) \leq \begin{cases} |w| + 2|x|, & \text{if } wx \neq wx; \\ d(g(|w|/d, |x|/d) + 1) + 2, & \text{if } wx = xw \text{ and } d = \gcd(|w|, |x|) \ . \end{cases}$$

Proof. Similar to the proof of the previous theorem. Omitted. ∎

6. Longest word omitted

In this section we assume that $S = \{x_1, x_2, \ldots, x_k\}$ for finite words $x_1, x_2, \ldots, x_k$, and S^* is co-finite. We first obtain an upper bound on the length of the longest word not in S^*.

Theorem 6.1. *Suppose $|x_i| \leq n$ for all i. Then if S^* is co-finite, the length of the longest word not in S^* is $< \frac{2}{2|\Sigma|-1}(2^n|\Sigma|^n - 1)$.*

In the rest of this section we show that the length of the longest word not in S^* can be exponentially long in n. We need several preliminary results first.

We say that x is a *proper prefix* of a word y if $y = xz$ for a nonempty word z. Similarly, we say x is a *proper suffix* of y if $y = zx$ for a nonempty word z.

Proposition 6.2. *Let S be a finite set of nonempty words such that S^* is co-finite, and $S^* \neq \Sigma^*$. Then for all $x \in S$, there exists $x' \in S$ such that x is a proper prefix of x, or vice versa. Similarly, for all $x \in S$, there exists $x' \in S$ such that x is a proper suffix of x', or vice versa.*

Proof. Let $x \in S$. Since $S^* \neq \Sigma^*$, there exists $v \in \overline{S^*}$. Since S^* is co-finite, $S^* \cap x^*v$ is nonempty. Let $i \geq 0$ be the smallest integer such that $x^iv \in S^*$; then $i \geq 1$, for otherwise $v \in S^*$. Since $x^iv \in S^*$, there exist $y_1, y_2, \ldots, y_j \in S$ such that $x^iv = y_1y_2 \ldots y_j$. Now $y_1 \neq x$, for otherwise by cancelling an x from both sides, we would have $x^{i-1}v \in S^*$, contradicting the minimality of i. If $|x| < |y_1|$, then x is a proper prefix of y_1, while if $|x| > |y_1|$, then y_1 is a proper prefix of x.

A similar argument applies for the result about suffixes. ∎

Next, we give two lemmas that characterize those sets S such that S^* is co-finite, when S is a set containing words of no more than two distinct lengths.

Lemma 6.3. *Suppose $S \subseteq \Sigma^m \cup \Sigma^n$, $0 < m < n$, and S^* is co-finite. Then $\Sigma^m \subseteq S$.*

Proof. If $S^* = \Sigma^*$, then S must contain every word x of length m, for otherwise S^* would omit x. So assume $S^* \neq \Sigma^*$.

Let $x \in \Sigma^m$. Then $S^* \cap x\Sigma^*$ is nonempty, since S^* is co-finite. Choose v such that $xv \in S^*$; then there is a factorization $xv = y_1 y_2 \cdots y_j$ where each $y_i \in S$. If $y_1 \in \Sigma^m$, then $x = y_1$ and so $x \in S$. Otherwise $y_1 \in \Sigma^n$. By Proposition 6.2, there exists $z \in S$ such that y_1 is a proper prefix of z or vice versa. But since S contains words of only lengths m and n, and $y_1 \in \Sigma^n$, we must have $z \in \Sigma^m$, and z is a prefix of y_1. Then $x = z$, and so $x \in S$. ∎

Lemma 6.4. *Suppose $S \subseteq \Sigma^m \cup \Sigma^n$, with $0 < m < n < 2m$ and S^* is co-finite. Then $\Sigma^l \subseteq S^*$, where $l = m|\Sigma|^{n-m} + n - m$.*

Proof. Let x be a word of length l that is not in S^*. Then we can write x uniquely as

$$x = y_0 z_0 y_1 z_1 \cdots y_{|\Sigma|^{n-m}-1} z_{|\Sigma|^{n-m}-1} y_{|\Sigma|^{n-m}}, \tag{6.1}$$

where $y_i \in \Sigma^{n-m}$ for $0 \leq i \leq |\Sigma|^{n-m}$, and $z_i \in \Sigma^{2m-n}$ for $0 \leq i < |\Sigma|^{n-m}$.

Now suppose that $y_i z_i y_{i+1} \in S$ for some i with $0 \leq i < |\Sigma|^{n-m}$. Then we can write

$$x = \left(\prod_{0 \leq j < i} y_j z_j \right) y_i z_i y_{i+1} \left(\prod_{i+1 \leq k \leq |\Sigma|^{n-m}} z_k y_k \right).$$

Note that $|y_j z_j| = |z_k y_k| = m$. From Lemma 6.3, each term in this factorization is in S. Hence $x \in S^*$, a contradiction. It follows that

$$y_i z_i y_{i+1} \notin S \text{ for all } i \text{ with } 0 \leq i < |\Sigma|^{n-m}. \tag{6.2}$$

Now the factorization of x in Eq. (6.1) uses $|\Sigma|^{n-m} + 1$ y's, and there are only $|\Sigma|^{n-m}$ distinct words of length $n - m$. So, by the pigeonhole principle, we have $y_p = y_q$ for some $0 \leq p < q \leq |\Sigma|^{n-m}$. Now define

$$\begin{aligned} u &= y_0 z_0 \cdots y_{p-1} z_{p-1} \\ v &= y_p z_p \cdots y_{q-1} z_{q-1} \\ w &= y_q z_q \cdots y_{|\Sigma|^{n-m}}, \end{aligned}$$

so $x = uvw$. Since S^* is co-finite, there exists a smallest exponent $k \geq 0$ such that $uv^k w \in S^*$.

Now let $uv^k w = x_1 x_2 \cdots x_j$ be a factorization into elements of S. Then x_1 is a word of length m or n. If $|x_1| = n$, then comparing lengths gives $x_1 = y_0 z_0 y_1$. But by (6.2) we know $y_0 z_0 y_1 \notin S$. So $|x_1| = m$, and comparing lengths gives $x_1 = y_0 z_0$. By similar reasoning we see that $x_2 = y_1 z_1$, and so on. Hence $x_j = y_{|\Sigma|^{n-m}-1} z_{|\Sigma|^{n-m}-1} y_{|\Sigma|^{n-m}} \in S$. But this contradicts (6.2).

Thus, our assumption that $x \notin S^*$ must be false, and so $x \in S^*$. Since x was arbitrary, this proves the result. ∎

Now we can prove an upper bound on the length of omitted words, in the case where S contains words of at most two distinct lengths.

Theorem 6.5. *Suppose $S \subseteq \Sigma^m \cup \Sigma^n$, where $0 < m < n < 2m$, and S^* is co-finite. Then the length of the longest word not in S^* is $\leq g(m,l) = ml - m - l$, where $l = m|\Sigma|^{n-m} + n - m$.*

Proof. Any word in S^* must be a concatenation of words of length m and n. If $\gcd(m,n) = d > 1$, then S^* omits all words whose length is not congruent to 0 (mod d), so S^* is not co-finite, contrary to the hypothesis. Thus $\gcd(m,n) = 1$.

By Lemmas 6.3 and 6.4, we have $\Sigma^m \cup \Sigma^l \subseteq S^*$, where $l = m|\Sigma|^{n-m} + n - m$. Hence S^* contains all words of length m and l; since $\gcd(m,l) = 1$, S^* contains all words of length $> g(m,l)$. ∎

Remark. We can actually improve the result of the previous theorem to arbitrary m and n, thus giving an upper bound in the case where S consists of words of exactly two distinct lengths. Details will appear in a later version of the paper.

Corollary 6.6. *Suppose $S \subseteq \Sigma^m \cup \Sigma^n$, where $0 < m < n < 2m$ and $\gcd(m,n) = 1$. Then S^* is co-finite if and only if $\Sigma^m \subseteq S$ and $\Sigma^l \subseteq S^*$, where $l = m|\Sigma|^{n-m} + n - m$.*

Proof. If S^* is co-finite, then by Lemmas 6.3 and 6.4 we get $\Sigma^m \subseteq S$ and $\Sigma^l \subseteq S^*$. On the other hand, if $\Sigma^m \subseteq S$ and $\Sigma^l \subseteq S^*$, then since $\gcd(m,l) = 1$, every word of length $> g(m,l)$ is contained in S^*, so S^* is co-finite. ∎

We need one more technical lemma.

Lemma 6.7. *Suppose $S \subseteq \Sigma^m \cup \Sigma^n$, where $0 < m < n < 2m$, and S^* is co-finite. Let τ be a word not in S^* where $|\tau| = n + jm$ for some $j \geq 0$. Then $S^* \cap (\tau\Sigma^m)^{i-1}\tau = \emptyset$ for $1 \leq i < m$.*

Proof. As before, since S^* is co-finite we must have $\gcd(m,n) = 1$. Define $L_i = (\tau\Sigma^m)^{i-1}\tau$ for $1 \leq i < m$. We prove that $S^* \cap L_i = \emptyset$ by induction on i.

The base case is $i = 1$. Then $L_i = L_1 = \{\tau\}$. But $S^* \cap \{\tau\} = \emptyset$ by the hypothesis that $\tau \notin S^*$.

Now suppose we have proved the result for some i, $i \leq m-2$, and we want to prove it for $i+1$. First we show that $S^* \cap \Sigma^{n-m}L_i = \emptyset$. Assume that $uw \in S^*$ for some $u \in \Sigma^{n-m}$ and $w \in L_i$. Then there is a factorization

$$uw = y_1 y_2 \cdots y_t \tag{6.3}$$

where $y_h \in S$ for $1 \leq h \leq t$. Now $|uw| = n - m + (n + jm + m)(i-1) + n + jm = n(i+1) + m(ji + i - 2)$. Since $0 < i+1 < m$, m does not divide $|uw|$. Thus at least one of the y_h is of length n, for otherwise (6.3) could not be a factorization of uw into elements of S. Let r be the smallest index such that $|y_r| = n$. Then we have

$$uw = \overbrace{y_1 y_2 \cdots y_{r-1}}^{\text{all of length } m} \quad \overbrace{y_r}^{\text{of length } n} \quad y_{r+1} \cdots y_t.$$

Hence $|y_1 y_2 \cdots y_r| = m(r-1) + n = mr + n - m$. Since, by Lemma 6.3 we have $\Sigma^m \subseteq S$, we can write $y_1 \cdots y_r = u z_1 \cdots z_r$, where $z_h \in S$ for $1 \leq h \leq r$. Thus

$$\begin{aligned} uw &= y_1 \cdots y_{r-1} y_r y_{r+1} \cdots y_t \\ &= u z_1 \cdots z_r y_{r+1} \cdots y_t; \end{aligned}$$

and, cancelling the u on both sides, we get $w = z_1 \cdots z_r y_{r+1} \cdots y_t$. But each term on the right is in S, so $w \in S^*$. But this contradicts our inductive hypothesis that $S^* \cap L_i = \emptyset$.

So now we know that

$$S^* \cap \Sigma^{n-m} L_i = \emptyset; \tag{6.4}$$

we'll use this fact below.

Now assume that $S^* \cap L_{i+1} \neq \emptyset$. Since $L_{i+1} = \tau \Sigma^m L_i$, there exists $\alpha \in \Sigma^m$ and $w \in L_i$ such that $\tau \alpha w \in S^*$. Write $\tau \alpha w = g_1 g_2 \cdots g_p$, where $g_h \in S$ for $1 \leq h \leq p$. We claim that $g_h \in \Sigma^m$ for $1 \leq h \leq j+1$. For if not, let k be the smallest index such that $|g_k| = n$. Then by comparing lengths, we have

$$\tau = \overbrace{g_1 g_2 \cdots g_{k-1}}^{\text{each of length } m} \quad \overbrace{g_k}^{\text{of length } n} \quad \overbrace{g'_1 g'_2 \cdots g'_{j-k+1}}^{\text{each of length } m}$$

for some $g'_1, g'_2, \ldots, g'_{j-k+1} \in \Sigma^m$. But this shows $\tau \in S^*$, a contradiction. We also have $g_{j+1} \notin \Sigma^n$, for otherwise $\tau = g_1 \cdots g_j g_{j+1} \in S^*$, a contradiction.

Now either $g_{j+2} \in \Sigma^m$ or $g_{j+2} \in \Sigma^n$. In the former case, by comparing lengths, we see that $g_{j+3} \cdots g_p \in \Sigma^{n-m} L_i$. But this contradicts (6.4). In the latter case, by comparing lengths, we see $g_{j+3} \cdots g_p \in L_i$, contradicting our inductive hypothesis. Thus our assumption that $S^* \cap L_{i+1} \neq \emptyset$ was wrong, and the lemma is proved. ∎

Now we are ready to give a class of examples achieving the bound in Theorem 6.5. Without lose of generality, let $\Sigma = \{0, 1, \cdots\}$. We define $r(n, k, l)$ to be the word of length l representing n in base k, possibly with leading zeros. For example, $r(11, 2, 5) = \mathtt{01011}$. For integers $0 < m < n$, we define

$$T(m, n) = \{r(i, |\Sigma|, n-m)\mathtt{0}^{2m-n} r(i+1, |\Sigma|, n-m) \ : \ 0 \leq i \leq |\Sigma|^{n-m} - 2\}.$$

For example, over a binary alphabet we have $T(3, 5) = \{\mathtt{00001}, \mathtt{01010}, \mathtt{10011}\}$.

Theorem 6.8. *Let m, n be integers with $0 < m < n < 2m$ and $\gcd(m, n) = 1$, and let $S = \Sigma^m \cup \Sigma^n - T(m, n)$. Then S^* is co-finite and the longest words not in S^* are of length $g(m, l)$, where $l = m|\Sigma|^{n-m} + n - m$.*

Proof. First, let's prove that S^* is co-finite. Since $\Sigma^m \subseteq S$, by Corollary 6.6 it suffices to show that $\Sigma^l \subseteq S^*$, where $l = m|\Sigma|^{n-m} + n - m$.

Let $x \in \Sigma^l$, and write

$$x = y_0 z_0 y_1 z_1 \cdots y_{|\Sigma|^{n-m}-1} z_{|\Sigma|^{n-m}-1} y_{|\Sigma|^{n-m}}$$

where $y_i \in \Sigma^{n-m}$ for $0 \leq i \leq |\Sigma|^{n-m}$, and $z_i \in \Sigma^{2m-n}$ for $0 \leq i < |\Sigma|^{n-m}$.

If $y_i z_i y_{i+1} \in T(m, n)$ for all i, $0 \leq i < |\Sigma|^{n-m}$, then since the base-$k$ expansions are forced to match up, we have $y_i = r(i, |\Sigma|, n-m)$ for $0 \leq i < |\Sigma|^{n-m}$. But the longest such word is of length $m|\Sigma|^{n-m} + n - 2m < l$, a contradiction. Hence $y_i z_i y_{i+1} \in S$ for some i. Thus

$$x = \left(\prod_{0 \leq j < i} y_j z_j \right) y_i z_i y_{i+1} \left(\prod_{i+1 \leq k \leq |\Sigma|^{n-m}} z_k y_k \right).$$

Note that $|y_j z_j| = |z_k y_k| = m$. Since $\Sigma^m \subseteq S$, this gives a factorization of $x \in S^*$. Since x was arbitrary, we have $\Sigma^l \subseteq S^*$.

Now we will prove that $\tau \notin S^*$, where

$$\tau := r(0, |\Sigma|, n-m)\mathtt{0}^{2m-n} r(1, |\Sigma|, n-m)\mathtt{0}^{2m-n} \cdots r(|\Sigma|^{n-m} - 1, |\Sigma|, n-m).$$

Note that $|\tau| = |\Sigma|^{n-m}(n-m) + (|\Sigma|^{n-m}-1)(2m-n) = m|\Sigma|^{n-m} + n - 2m = l - m$. Suppose there exists a factorization $\tau = w_1 w_2 \cdots w_t$, where $w_i \in S$ for $1 \leq i \leq t$. Since $|\tau|$ is not divisible by m, at least one of these terms is of length n. Let k be the smallest index such that $w_k \in \Sigma^n$. then $\tau = w_1 \cdots w_{k-1} w_k w_{k+1} \cdots w_t$. By comparing lengths, we get $w_i = r(i-1, |\Sigma|, n-m)0^{2m-n}$ for $1 \leq i < k$. Thus $w_k = r(k-1, |\Sigma|, n-m)0^{2m-n}r(k, |\Sigma|, n-m) \in S \cap \Sigma^n$. But $r(k-1, |\Sigma|, n-m)0^{2m-n}r(k, |\Sigma|, n-m) \in T(m,n)$, a contradiction. Thus $\tau \notin S^*$.

We may now apply Lemma 6.7 to get that S^* omits words of the form $(\tau\Sigma^m)^{m-2}\tau$; these words are of length $(l-m+m)(m-2)+l-m = lm-l-m = g(m,l)$. ∎

Corollary 6.9. *For each odd integer $n \geq 5$, there exists a set of binary words of length at most n, such that S^* is co-finite and the longest word not in S^* is of length $\Omega(n^2 2^{n/2})$.*

Proof. Choose $m = (n+1)/2$ and apply Theorem 6.8. ∎

Example 6.10. Let $m = 3$, $n = 5$, $\Sigma = \{0, 1\}$. Then $S = \Sigma^3 + \Sigma^5 - \{00001, 01010, 10011\}$. Then a longest word not in S^* is 00001010011 000 00001010011, of length 25.

7. Number of omitted words

Recall that $f(x_1, x_2, \ldots, x_k)$ is the classical function which, for positive integers $x_1, \ldots, x_k$ with $\gcd(x_1, \ldots, x_k) = 1$, counts the number of integers not representable as a non-negative integer linear combination of the x_i. In this section we consider a generalization of this function to the setting of a free monoid, replacing the integers x_i with finite words in Σ^*, and replacing the condition $\gcd(x_1, \ldots, x_k) = 1$ with the requirement that $\{x_1, \ldots, x_k\}^*$ be co-finite.

We have already studied this in the case of a unary alphabet in Section 2, so let us assume that Σ has at least two letters.

Theorem 7.1. *Let $x_1, x_2, \ldots, x_k \in \Sigma^*$ be such that $|x_i| \leq n$ for $1 \leq i \leq n$. Let $S = \{x_1, x_2, \ldots, x_k\}$ and suppose S^* is co-finite. Then*

$$\mathcal{M} = |\Sigma^* - S^*| \leq \frac{|\Sigma|^q - 1}{|\Sigma| - 1},$$

where $q = \frac{2}{2|\Sigma|-1}(2^n|\Sigma|^n - 1)$.

Proof. From Theorem 6.1, we know that if S^* is co-finite, the length of the longest omitted word is $< q$, where $q = \frac{2}{2|\Sigma|-1}(2^n|\Sigma|^n - 1)$. The total number of words $< q$ is $1 + |\Sigma| + \cdots + |\Sigma|^{q-1} = \frac{|\Sigma|^q - 1}{|\Sigma|-1}$. ∎

We now give an example achieving a doubly-exponential lower bound on $\mathcal{M}$.

Theorem 7.2. *Let m, n be integers with $0 < m < n < 2m$ and $\gcd(m,n) = 1$, and let $S = \Sigma^m \cup \Sigma^n - U(m,n)$, where U is defined by*

$$U(m,n) = \{r(i, |\Sigma|, n-m)0^{2m-n}r(j, |\Sigma|, n-m) \ : \ 0 \leq i < j \leq |\Sigma|^{n-m} - 1\}.$$

Then S^* is co-finite and S^* omits at least $2^{|\Sigma|^{n-m}} - |\Sigma|^{n-m} - 1$ words.

Proof. Similar to that of Theorem 6.8. ∎

References

[1] M. Beck, R. Diaz, and S. Robins. The Frobenius problem, rational polytopes, and Fourier-Dedekind sums. *J. Number Theory* **96** (2002), 1–21.

[2] A. Brauer. On a problem of partitions. *Amer. J. Math.* **64** (1942), 299–312.

[3] C. Câmpeanu, K. Culik II, K. Salomaa, and S. Yu. State complexity of basic operations on finite languages. In *WIA'99*, *Lect. Notes in Comp. Science* 2214, pp. 60-70, 2001.

[4] C. Câmpeanu and W. H. Ho. The maximum state complexity for finite languages. *J. Automata, Languages, and Combinatorics* **9** (2004), 189–202.

[5] C. Câmpeanu, K. Salomaa, and S. Yu. State complexity of regular languages: finite versus infinite. In C. S. Calude and G. Păun, eds., *Finite Versus Infinite: Contributions to an Eternal Dilemma*, pp. 53–73. Springer, 2000.

[6] S. Cautis, F. Mignosi, J. Shallit, M.-w. Wang, and S. Yazdani. Periodicity, morphisms, and matrices. *Theoret. Comput. Sci.* **295** (2003), 107–121.

[7] M. Chrobak. Finite automata and unary languages. *Theoret. Comput. Sci.* **47** (1986), 149–158. Errata, **302** (2003), 497–498.

[8] P. Chrzastowski-Wachtel and M. Raczunas. Liveness of weighted circuits and the Diophantine problem of Frobenius. In Z. Ésik, ed., *FCT '93*, *Lect. Notes in Comp. Science* 710, pp. 171–180, 1993.

[9] J. Clément, J.-P. Duval, G. Guaina, D. Perrin, and G. Rindone. Parsing with a finite dictionary. *Theoret. Comput. Sci.* **340** (2005), 432–442.

[10] J. L. Davison. On the linear diophantine problem of Frobenius. *J. Number Theory* **48** (1994), 353–363.

[11] A. L. Dulmage and N. S. Mendelsohn. Gaps in the exponent set of primitive matrices. *Illinois J. Math.* **8** (1964), 642–656.

[12] D. Einstein, D. Lichtblau, A. Strzebonski, and S. Wagon. Frobenius numbers by lattice point enumeration. *Integers* **7** (2007), A15 (electronic).

[13] K. Ellul, B. Krawetz, J. Shallit, and M.-w. Wang. Regular expressions: new results and open problems. *J. Autom. Lang. Combin.* **10** (2005), 407–437.

[14] N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965), 109–114.

[15] H. Greenberg. Solution to a linear Diophantine equation for nonnegative integers. *J. Algorithms* **9** (1988), 343–353.

[16] M. Holzer and M. Kutrib. Nondeterministic descriptional complexity of regular languages. *Internat. J. Found. Comp. Sci.* **14** (2003), 1087–1102.

[17] J. Incerpi and R. Sedgewick. Improved upper bounds on shellsort. *J. Comput. System Sci.* **31** (1985), 210–224.

[18] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica* **12** (1992), 161–177.

[19] E. Kunz. The value-semigroup of a one-dimensional Gorenstein ring. *Proc. Amer. Math. Soc.* **25** (1970), 748–751.

[20] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.

[21] A. N. Maslov. Estimates of the number of states of finite automata. *Dokl. Akad. Nauk. SSSR* **194** (1970), 1266–1268. In Russian. English translation in *Soviet Math. Dokl.* **11** (1970), 1373–1375.

[22] J. L. Ramírez-Alfonsín. Complexity of the Frobenius problem. *Combinatorica* **16** (1996), 143–147.

[23] J. L. Ramírez-Alfonsín. *The Diophantine Frobenius Problem.* Oxford University Press, 2005.

[24] J. J. Sylvester. Problem 7382. *Math. Quest. Sol. Educ. Times* **41** (1884), ix, 21.

[25] E. Teruel, P. Chrzastowski-Wachtel, J. M. Colom, and M. Silva. On weighted T-systems. In K. Jensen, editor, *Applic. and Theory of Petri Nets 1992*, *Lect. Notes in Comp. Science* 616, pp. 348–367, 1992.

[26] H. S. Wilf. A circle-of-lights algorithm for the "money-changing problem". *Amer. Math. Monthly* **85** (1978), 562–565.

[27] S. Yu, Q. Zhuang, and K. Salomaa. The state complexities of some basic operations on regular languages. *Theoret. Comput. Sci.* **125** (1994), 315–328.