Parameterized Prefix Distance between Regular Languages

Martin Kutrib, Katja Meckel, and Matthias Wendlandt

Institut für Informatik, Universität Giessen
Arndtstr. 2, 35392 Giessen, Germany
{kutrib,meckel,matthias.wendlandt}@informatik.uni-giessen.de

Abstract. We investigate the parameterized prefix distance between regular languages. The prefix distance between words is extended to languages in such a way that the distances of all words up to length n to the mutual other language are summed up. Tight upper bounds for the distance between unary as well as non-unary regular languages are derived. It is shown that there are pairs of languages having a constant, degree k polynomial, and exponential distance. Moreover, for every constant and every polynomial, languages over a binary alphabet are constructed that have exactly that distance. From the density and census functions of regular languages the orders of possible distances between languages are derived and are shown to be decidable.

1 Introduction

Finite state devices are used in several applications and implementations in software engineering, programming languages and other practical areas in computer science. They are one of the first and most intensely investigated computational models. Due to several applications and implementations of transducers in theoretical and practical areas of computer science, their fault-tolerance or even usability in the presence of failures is a natural question of crucial importance. The applications are widely spread. For example, finite state transducers are currently used for compiler constructions [1], language and speech processing [7], and even for the design of controllability systems in aircraft design [9]. Much of the underlying theory has originated from linguistics. In natural language and speech processing transducers with more than one hundred million states may be used [8]. All of the components involved may be subject to failure. However, not all faults necessarily incapacitate the automaton entirely. In several applications small aberrations are tolerable. From this point of view the questions of what are tolerable aberrations arise immediately. We consider the distance between the languages accepted by the original and the faulty machine as measure for this purpose. So, even in the case of transducers we regard the accepting part of the computation only.

Inspired by these considerations, here we start to investigate the parameterized prefix distance between regular languages. In [4] several notions of distances

V. Geffert et al. (Eds.): SOFSEM 2014, LNCS 8327, pp. 419–430, 2014.

have been extended from distances between strings to distances between languages (see also [6]). To this end, a relative distance between a language L_1 and a language L_2 is defined to be the supremum of the minimal distances of all words from L_1 to L_2. The distance between L_1 and L_2 is defined as the maximum of their mutual relative distances. Since here we are interested in computations of faulty finite state devices that are still tolerable, we stick with the prefix distance and consider a parameterized extension. For words w_1 and w_2 the prefix distance sums up the number of all letters of w_1 and w_2 that do not belong to a common prefix of these words. One can suppose that on the common prefixes the computations of both machines are the same until a faulty component comes into play and the computations diverge. The parameterized prefix distance between languages sums up the distances of all words up to length n from one language to their closest words from the other language, and vice versa.

Since the distance between identical words should always be 0, for the distances between languages, the number of words in their symmetric difference plays a crucial role. In this connection we utilize the density and census functions that count the number of words in a language. The study of densities of regular languages has a long history (see, for example, [3,5,10,11,12,13]). Restricted to the number of unary words in a binary language the census function has been shown to be log-space many-one complete for $\#L$ in [2].

In particular, in the present paper tight upper bounds for the parameterized prefix distance between unary as well as non-unary regular languages are derived. It is shown that there are pairs of languages having a constant, degree k polynomial, and exponential distance. Moreover, for every constant and every polynomial, languages over a binary alphabet are constructed that have exactly that distance. From practical as well as theoretical point of view, it is important to decide this order. Here, the orders of possible distances between regular languages are derived and are shown to be decidable.

2 Preliminaries

We write Σ^* for the set of all words over the finite alphabet Σ, and $\mathbb{N}$ for the set $\{0, 1, 2, \dots\}$ of non-negative integers. The *empty word* is denoted by λ and the *reversal* of a word w by w^R. For the *length* of w we write $|w|$ and for the number of occurrences of a symbol a in w we use the notation $|w|_a$. We use $\subseteq$ for *inclusions* and $\subset$ for *strict inclusions*, and write 2^S for the powerset of a set S.

In general, a *distance* over Σ^* is a function $d : \Sigma^* \times \Sigma^* \to \mathbb{N} \cup \{\infty\}$ satisfying, for all $x, y, z \in \Sigma^*$, the conditions $d(x, y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$, and $d(x, y) \leq d(x, z) + d(z, y)$.

For example, words w_1 and w_2 over Σ^* can be compared by summing up the number of all letters of w_1 and w_2 that do not belong to a common prefix of these words. This so-called prefix distance $d_{\text{pref}} : \Sigma^* \times \Sigma^* \to \mathbb{N}$ between words is defined to be $d_{\text{pref}}(w_1, w_2) = |w_1| + |w_2| - 2\max\{\, |v| \mid w_1, w_2 \in v\Sigma^* \,\}$. Clearly, $d_{\text{pref}}(w_1, w_2) = 0$ if and only if $w_1 = w_2$, and $d_{\text{pref}}(w_1, w_2) = |w_1| + |w_2|$ if and only if the first letters of w_1 and w_2 are different. Moreover, the prefix distance between two words can be large if their length difference is large.

Distances over Σ are extended to distances between a word and a language by taking the minimum of the distances between the word and the words belonging to the language. For the prefix distance we obtain pref-$d : \Sigma^* \times 2^{\Sigma^*} \to \mathbb{N} \cup \{\infty\}$ which is defined to be

$$\text{pref-}d(w, L) = \begin{cases} \min\{\, d_{\text{pref}}(w, w') \mid w' \in L \,\} & \text{if } L \neq \emptyset \\ \infty & \text{otherwise} \end{cases}.$$

Clearly, pref-$d(w, L) = 0$ if $w \in L$.

The next step is to extend the distance between a word and a language to a distance between two languages $L_1, L_2 \subseteq \Sigma^*$. This can be done by taking the maximum of the suprema of the distances of all words from L_1 to L_2 and vice versa. However, here we are interested in a parameterized definition, where the distance additionally depends on the length of the words. So, the *parameterized prefix distance* between languages pref-$D : \mathbb{N} \times 2^{\Sigma^*} \times 2^{\Sigma^*} \to \mathbb{N} \cup \{\infty\}$ is defined by

$$\text{pref-}D(n, L_1, L_2) = \sum_{\substack{w \in L_1,\\ 0 \leq |w| \leq n}} \text{pref-}d(w, L_2) + \sum_{\substack{w \in L_2,\\ 0 \leq |w| \leq n}} \text{pref-}d(w, L_1).$$

In general, one cannot expect to obtain a convenient description of the parameterized prefix distance for *all* n. So, in the following, if not stated otherwise, it is understood that pref-$D(n, L_1, L_2) = f(n)$ means pref-$D(n, L_1, L_2) = f(n)$, for all n greater than some constant n_0.

The following technical proposition is a useful tool for the analysis and construction of regular languages having a certain distance.

Proposition 1. *Let $L_1, L_2 \subseteq \Sigma^*$ be two languages so that $L_1 \subseteq L_2$.*

1. *For a word $v \in L_2 \setminus L_1$, let $L_1' = L_1 \cup \{v\}$ and $L_2' = L_2 \setminus \{v\}$. Then*

$$\begin{aligned} &\text{pref-}D(n, L_1, L_2) > \text{pref-}D(n, L_1', L_2) \text{ and} \\ &\text{pref-}D(n, L_1, L_2) > \text{pref-}D(n, L_1, L_2'). \end{aligned}$$

2. *For a word $v \in \Sigma^* \setminus L_2$, let $L_2' = L_2 \cup \{v\}$. Then*

$$\text{pref-}D(n, L_1, L_2) < \text{pref-}D(n, L_1, L_2').$$

3. *For a word $v \in L_1$, let $L_1' = L_1 \setminus \{v\}$. Then*

$$\text{pref-}D(n, L_1, L_2) < \text{pref-}D(n, L_1', L_2).$$

Example 2. We consider the two regular languages $L_1 = \{a, b\}^*\{ab, ba\}\{a, b\}^*$, that is, the language of all words over $\{a, b\}$ containing the factor ab or ba, and $L_2 = \{a, b\}^* b\{a, b\}^* b\{a, b\}^*$, that is, the language of all words over $\{a, b\}$ containing at least two symbols b.

In order to compute their parameterized prefix distance, first the distances of all words from L_1 to L_2 are determined. All words of L_1 that belong to L_2 are of the forms $\{a, b\}^*\{ab, ba\}\{a, b\}^* b\{a, b\}^*$ or $\{a, b\}^* b\{a, b\}^*\{ab, ba\}\{a, b\}^*$. So, we only have to compute the prefix distances of words w from $a^*\{ab, ba\}a^*$ to L_2, which is 1 since $wb \in L_2$ and pref-$d(w, L_2) = |w| + |wb| - 2|w|$.

Second, all words w from L_2 that are not included in L_1 are of the form b^2b^*. Their prefix distance to L_1 is also always 1, since $wa \in L_1$.

Together, the prefix distance between L_1 and L_2 is

$$\text{pref-}D(n, L_1, L_2) = \sum_{\substack{w \in a^*ba^*,\\ 2 \le |w| \le n}} \text{pref-}d(w, L_2) + \sum_{\substack{w \in b^2b^*,\\ 2 \le |w| \le n}} \text{pref-}d(w, L_1).$$

These sums can be reformulated by summing up over the sizes of the words and multiplying by their prefix distance to the languages they are not contained in. So, we obtain

$$\text{pref-}D(n, L_1, L_2) = \sum_{i=2}^{n} |\{w \in a^*ba^* \mid |w| = i\}| \cdot 1 + \sum_{i=2}^{n} \left|\{w \in b^2b^* \mid |w| = i\}\right| \cdot 1.$$

The set $\{\, w \in a^*ba^* \mid |w| = i \,\}$ of the first sum contains i words. The set $\{\, w \in b^2b^* \mid |w| = i \,\}$ of the third sum has a size of 1. Therefore, the result is

$$\text{pref-}D(n, L_1, L_2) = \sum_{i=2}^{n} i + \sum_{i=2}^{n} 1 = \frac{(n+1)n}{2} - 1 + n - 1 = \frac{n^2}{2} + \frac{3}{2}n - 2.$$

□

3 Upper and Lower Bounds for the Prefix Distance

We turn to investigate the range of possible parameterized distances between regular languages. We are interested in upper bounds and whether these upper bounds are tight, that is, whether there are witness languages showing that the upper bound is, in fact, the best possible.

To determine the upper bound of the prefix distance between two languages $L_1, L_2 \subseteq \Sigma^*$ we consider some word $w \in L_1$ and the shortest word $s \in L_2$. In any case we have pref-$d(w, L_2) \le |w| + |s|$ and, thus, the word w contributes in a maximal way to the distance if it does not have a common prefix with s. In this case, we have pref-$d(w, L_2) = |w| + |s|$. This observation leads to a general upper bound as follows.

Proposition 3. *Let $L_1, L_2 \subseteq \Sigma^*$ be two non-empty languages, $m_1 = \min\{\, |w| \mid w \in L_1 \,\}$ and $m_2 = \min\{\, |w| \mid w \in L_2 \,\}$ be the lengths of the shortest words of L_1 and L_2, respectively, $m = \min\{m_1, m_2\}$, and $M = \max\{m_1, m_2\}$. Then*

$$\text{pref-}D(n, L_1, L_2) \le \sum_{i=m}^{n} |\Sigma|^i \cdot (i + M).$$

The next lemma identifies properties that are necessary for two languages to match the upper bound.

Lemma 4. *Let L_1, L_2 be languages with*

$$m = \min\{\min\{\, |w| \mid w \in L_1 \,\}, \min\{\, |w| \mid w \in L_2 \,\}\} \text{ and}$$

$$M = \max\{\min\{\, |w| \mid w \in L_1 \,\}, \min\{\, |w| \mid w \in L_2 \,\}\}.$$

Then the upper bound of Proposition 3 is met only if (i) each word $w \in L_1 \cup L_2$ contributes $|w| + M$ to the prefix distance, (ii) $L_1 \cap L_2 = \emptyset$ if $m \geq 1$, and $L_1 \cap L_2 \subseteq \{\lambda\}$ if $m = 0$, and (iii) $L_1 \cup L_2 = \{\, w \in \Sigma^* \mid |w| \geq m \,\}$.

Proof. We assume $m \geq 1$ and $L_1 \cap L_2 \neq \emptyset$, or $m = 0$ and $L_1 \cap L_2$ is not a subset of $\{\lambda\}$. In both cases there exists at least one word w of length greater than or equal to $\max\{1, m\}$ that does not contribute to pref-$D(|w|, L_1, L_2)$. So there must be a word in $L_1 \cup L_2$ that contributes more than $|w| + M$ to the prefix distance. However, this is a contradiction to the choice of M to be the maximum of the sizes of the shortest words. So, (ii) is a necessary condition.

If $L_1 \cup L_2 \neq \{\, w \in \Sigma^* \mid |w| \geq m \,\}$, then there is a word not in $L_1 \cup L_2$ whose length is at least $\max\{1, m\}$. This word can be added to both languages L_1 and L_2 without affecting pref-$D(n, L_1, L_2)$. Since in this case the intersection $L_1 \cap L_2$ contains a non-empty word, we have a contradiction to (ii). This shows (iii).

At last we assume that there exists a word $w \in L_1 \cup L_2$ that contributes less than $|w| + M$ to pref-$D(|w|, L_1, L_2)$. Then there must be a word in $L_1 \cup L_2$ that contributes more than $|w| + M$ to the prefix distance. The same contradiction as for (ii) shows case (i). □

Lemma 4 particularly shows that the upper bound cannot be reached if $m < M$. Let in this case w with $|w| = M$ be a shortest word in its language, say L_2. Then pref-$d(w, L_1) \leq |w| + m < |w| + M$. So, condition (i) of the lemma is violated. Next we turn to show that the upper bound of Proposition 3 is the best possible, in the sense that there are worst case languages for which it is matched. These languages necessarily satisfy the conditions of Lemma 4.

Proposition 5. *For any $M = m \geq 0$, there are binary regular languages $L_1, L_2 \subseteq \{a, b\}^*$ so that* pref-$D(n, L_1, L_2) = \sum_{i=m}^{n} |\Sigma|^i \cdot (i + M)$, *where m is the minimum and M is the maximum of the lengths of the shortest words in L_1 and L_2.*

Proof. For any $m \geq 1$, we use the disjoint regular witness languages $L_1 = a\{a, b\}^{m-1}\{a, b\}^*$ and $L_2 = b\{a, b\}^{m-1}\{a, b\}^*$. In particular, no two words of L_1 and L_2 have a common prefix.

Let $w \in L_1$ be some word. Its prefix distance to L_2 is $|w| + m$. Similarly, the prefix distance of every word $w \in L_2$ to the language L_1 is $|w| + m$. So we have

$$\text{pref-}D(n, L_1, L_2) = \sum_{\substack{w \in L_1 \setminus L_2, \\ m \leq |w| \leq n}} |w| + m + \sum_{\substack{w \in L_2 \setminus L_1, \\ m \leq |w| \leq n}} |w| + m.$$

Since $L_1 \cup L_2 = \{\, w \in \Sigma^* \mid |w| \geq m \,\}$ and $L_1 \cap L_2 = \emptyset$ this in turn is

$$\text{pref-}D(n, L_1, L_2) = \sum_{i=m}^{n} |\Sigma|^i \cdot (i+m).$$

If $m = 0$, then the empty word belongs to both languages. In this case we set $L_1 = \{\lambda\} \cup a\{a,b\}^*$ and $L_2 = \{\lambda\} \cup b\{a,b\}^*$ and obtain

$$\text{pref-}D(n, L_1, L_2) = \sum_{i=1}^{n} |\Sigma|^i \cdot i = \sum_{i=0}^{n} |\Sigma|^i \cdot (i+0) = \sum_{i=m}^{n} |\Sigma|^i \cdot (i+m).$$

□

So far, we considered languages over alphabets with at least two letters. For unary languages the situation changes significantly. An immediate observation is, that every two words have a distance to each other which is given by their length difference only.

Proposition 6. *Let $L_1 \subseteq \{a\}^*$ and $L_2 \subseteq \{a\}^*$ be two non-empty unary languages. Then* pref-$D(n, L_1, L_2) \leq \frac{n(n+1)}{2} + 1$.

As for the general case, the upper bound for the parameterized prefix distance of unary regular languages is tight. However, the witness languages of the following proof are the only ones whose distance meets the upper bound.

Proposition 7. *There are unary regular languages $L_1, L_2 \subseteq \{a\}^*$ so that their prefix distance is* pref-$D(n, L_1, L_2) = \frac{n(n+1)}{2} + 1$.

Proof. Let $L_1 = aa^*$ and $L_2 = \{\lambda\}$. These languages are unary, regular, and disjoint. Therefore, the prefix distance of each word in $w \in L_1$ to L_2 is $|w|$. For the only word λ in L_2 its distance to L_1 is $d_{\text{pref}}(\lambda, a) = 1$. So we have pref-$D(n, L_1, L_2) = 1 + \sum_{i=1}^{n} i = \frac{n(n+1)}{2} + 1$. □

4 Distances Below the Upper Bound

So far, we have explored the upper and lower bounds for parameterized prefix distances. Here we are interested in the question which functions are possible to obtain by considering the prefix distance of two regular languages. The next proposition gives an example for regular languages whose parameterized prefix distance is superpolynomial.

Proposition 8. *There are regular languages L_1 and L_2 even over a binary alphabet so that* pref-$D(n, L_1, L_2) \in \Theta(n2^n)$.

Proof. Here we can use the witness languages L_1 and L_2 from the case $m = 0$ in the proof of Proposition 5. There,

$$\text{pref-}D(n, L_1, L_2) = \sum_{i=1}^{n} |\Sigma|^i \cdot i = \sum_{i=1}^{n} 2^i \cdot i.$$

has been shown. This sum is equal to $n2^{n+2} - (n+1)2^{n+1} + 2 \in \Theta(n2^n)$. □

Next we give evidence that, for any constant $c \geq 1$, there are regular languages having parameterized prefix distance c.

Proposition 9. *Let $c \geq 1$ be an integer. Then there are unary regular languages L_1 and L_2 so that* pref-$D(n, L_1, L_2) = c$*, for all $n \geq c$.*

Proof. We use the languages $L_1 = \{\lambda\}$ and $L_2 = \{\lambda, a^c\}$ as witnesses. Since $\lambda \in L_1 \cap L_2$ the empty word in L_1 and L_2 does not contribute to the distance between L_1 and L_2. Clearly, pref-$d(a^c, L_1) = c$ and, thus, pref-$D(n, L_1, L_2) = c$, for all $n \geq c$. □

Now we turn to the main part of this section. Given an arbitrary polynomial p with integer coefficients whose leading coefficient is positive, we show how to construct two regular languages over a binary alphabet having exactly the parameterized prefix distance p. Clearly, a negative leading coefficient does not make sense since it would yield a negative distance.

Theorem 10. *Let $p(n) = x_k \cdot n^k + x_{k-1} \cdot n^{k-1} + \cdots + x_0$ be a polynomial of degree $k \geq 0$ with integer coefficients x_i, $0 \leq i \leq k$, and $x_k \geq 1$. Then two regular languages L_1 and L_2 over the alphabet $\{a, b\}$ can effectively be constructed so that* pref-$D(n, L_1, L_2) = p(n)$*, for all $n \geq n_0$, where n_0 is some constant.*

Proof. Proposition 9 already shows the special case $k = 0$. Therefore, we assume $k \geq 1$. The basic idea of the construction is to start with two languages whose distance is already a polynomial of degree k, but its coefficients may be incorrect. Subsequently, the coefficients are corrected one after the other, from x_k to x_0. When coefficient x_i is corrected, the coefficients x_k to x_{i+1} are not affected while the coefficients x_{i-1} to x_0 may be changed.

In general, language L_1 will always be a subset of L_2. In this way, the words from L_1 never contribute to the distance.

For the corrections of the coefficients a set of equally long prefixes is used. So, we define $P \subseteq \{a, b\}^l$, for some constant l, with $P = \{p_0, p_1, \ldots, p_m\}$. Assume for a moment that $l \geq k$ is large enough to perform the following constructions. Later we will give evidence that it always can be chosen appropriately.

We consider auxiliary languages

$$L_{r,-1} = \{\, p_r \,\} \text{ and } L_{r,-1,b} = L_{r,-1} \cup L_{r,-1}b,$$
$$L_{r,s} = \{\, p_r v \mid v \in \{a, b\}^*, |v|_b = s \,\} \text{ and } L_{r,s,b} = L_{r,s} \cup L_{r,s}b$$

for $s \geq 0$ and $p_r \in P$. Clearly, there are $\binom{n-|p_r|}{s} = \binom{n-l}{s} \in \Theta(n^s)$ many words of length n in the languages $L_{r,s}$. Considering the distance between $L_{r,-1}$ and $L_{r,-1,b}$ we obtain pref-$D(n, L_{r,-1}, L_{r,-1,b}) = 1$. For the distance between $L_{r,s}$ and $L_{r,s,b}$, all words from $L_{r,s}b$ contribute 1 while the words from $L_{r,s}$ contribute nothing. For $s \geq 1$, we obtain

$$\text{pref-}D(n, L_{r,s-1}, L_{r,s-1,b}) = \sum_{i=1}^{n} \binom{i-l}{s-1} = \binom{n-l+1}{s}$$
$$= \frac{(n-l+1) \cdot (n-l) \cdot (n-l-1) \cdots (n-l-s+2)}{s!}$$

which gives us a term of the form $\frac{n^s+y_{s-1}\cdot n^{s-1}+y_{s-2}n^{s-2}+y_{s-3}n^{s-3}+\cdots+y_0}{s!}$, where a rough and simple estimation yields $|y_i| \leq 3^s \cdot l^s$, $0 \leq i \leq s-1$.

We start the construction by using the union of auxiliary languages with $x_k \cdot k!$ many different prefixes, that is,

$$L_1 = \bigcup_{i=0}^{x_k\cdot k!-1} L_{i,k-1} \text{ and } L_2 = \bigcup_{i=0}^{x_k\cdot k!-1} L_{i,k-1,b}.$$

So, we start with a distance of the form

$$x_k n^k + z_{k-1}n^{k-1} + z_{k-2}n^{k-2} + z_{k-3}n^{k-3} + \cdots + z_0,$$

where x_k is already the correct coefficient and $|z_i| \leq x_k \cdot 3^k \cdot l^k$, $0 \leq i \leq k-1$.

Next we correct the remaining coefficients. Let $x_{max} = \max\{\, x_i \mid 0 \leq i \leq k \,\}$. Concluding inductively, we assume that currently

$$\text{pref-}D(n, L_1, L_2) = x_k n^k + x_{k-1}n^{k-1} + \cdots + x_{k-i+1}n^{k-i+1} + z_{k-i}n^{k-i} + \cdots + z_0,$$

where the coefficients $x_k, x_{k-1}, \ldots, x_{k-i+1}$ are already correct and, moreover, $|z_{k-i}|, |z_{k-i-1}|, \ldots, |z_0| \leq 3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i$.

In order to obtain the correct coefficient x_{k-i}, we set $d = z_{k-i} - x_{k-i}$ and distinguish the two cases, where d is negative or positive. Clearly, if $d = 0$ the coefficient x_{k-i} is already correct and nothing has to be done.

If $d < 0$, the distance has to be increased. To this end, the auxiliary languages $L_{j,k-i-1}$ and $L_{j,k-i-1,b}$ are used. We add their unions with $|d| \cdot (k-i)!$ many new different prefixes to L_1 and L_2, that is,

$$\bigcup_{j=0}^{|d|\cdot(k-i)!-1} L_{j,k-i-1} \text{ is added to } L_1 \text{ and } \bigcup_{j=0}^{|d|\cdot(k-i)!-1} L_{j,k-i-1,b} \text{ is added to } L_2.$$

Since all the prefixes p_j are new and $L_1 \subseteq L_2$, again all words from L_2 contribute 1 to the distance while the words in L_1 contribute nothing. In particular, we have added $|d|n^{k-i} + z'_{k-i-1}n^{k-i-1} + z'_{k-i-2}n^{k-i-2} + \cdots + z'_0$ words up to length n to L_2, where $|z'_{k-i-1}|, |z'_{k-i-2}|, \ldots, |z'_0| \leq |d| \cdot 3^{k-i} \cdot l^{k-i} \leq |d| \cdot 3^k \cdot l^k$. This implies

$$\text{pref-}D(n, L_1, L_2) = x_k n^k + x_{k-1}n^{k-1} + \cdots + x_{k-i}n^{k-i} + z_{k-i-1}n^{k-i-1} + \cdots + z_0,$$

where $x_k, x_{k-1}, \ldots, x_{k-i}$ are already correct and $|z_{k-i-1}|, |z_{k-i-2}|, \ldots, |z_0|$ are at most

$$\begin{aligned}
&3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i + |d| \cdot 3^k \cdot l^k \\
&= 3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i + (3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i + x_{max}) \cdot 3^k \cdot l^k \\
&= 3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i + 3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i \cdot 3^k \cdot l^k + x_{max} \cdot 3^k \cdot l^k \\
&\leq 3^i \cdot x_{max} \cdot (3^k \cdot l^k)^{i+1}.
\end{aligned}$$

This concludes the first case.

If $d > 0$, the distance has to be decreased. To this end, words from L_2 are added to L_1 so that they do not contribute to the distance anymore. Let $\tilde{p}$ be one of the $x_k \cdot k!$ prefixes used at the beginning of the induction to establish a polynomial distance of degree k. Moreover, we may assume that $\tilde{p}$ has not been used for the current purpose before.

Then, for $r, t \geq 0$ and $s \geq r$, another auxiliary language is defined as $\tilde{L}_{\tilde{p},r,s,t} = \{\, \tilde{p}ub^r v \mid uv \in \{a,b\}^*, |u| = t, |uv|_b = s - r \,\}$. Here, we set $\tilde{L}_{\tilde{p},r,r-1,t} = \{\, \tilde{p}b^r \,\}$. In these languages the position of the block b^r is fixed, so that the union $\bigcup_{j=0}^{d\cdot(k-i)!-1} \tilde{L}_{\tilde{p},i,k-1,j}$ contains $dn^{k-i} + z'_{k-i-1}n^{k-i-1} + z'_{k-i-2}n^{k-i-2} + \cdots + z'_0$ words up to length n, for $n \geq d\cdot(k-i)!+l+i$, where $|z'_{k-i-1}|, |z'_{k-i-2}|, \ldots, |z'_0| \leq d \cdot 3^{k-i} \cdot (l+i)^{k-i} \leq d \cdot 3^k \cdot l^k$. Now all these words are concatenated with a symbol b and are added to L_1. Since all words do belong to L_2 as well, we obtain

$$\text{pref-}D(n, L_1, L_2) = x_k n^k + x_{k-1}n^{k-1} + \cdots + x_{k-i}n^{k-i} + z_{k-i-1}n^{k-i-1} + \cdots + z_0,$$

where the coefficients $x_k, \ldots, x_{k-i}$ are already correct and analogously to the first case $|z_{k-i-1}|, |z_{k-i-2}|, \ldots, |z_0|$ are at most $3^i \cdot x_{max} \cdot (3^k \cdot l^k)^{i+1}$. This concludes the second case.

The construction is concluded by the observation that choosing $n_0 > l + k$ is sufficient for the auxiliary languages applied in the initial step and the correction steps in the first case. For the corrections in the second case

$$d \cdot (k-i)! + l + i \leq 3^{k+1} \cdot x_{max} \cdot (3^{k^2} \cdot l^{k^2}) \cdot k! \leq n_0$$

is sufficient.

Finally, it has to be shown that the prefix length l always can be chosen appropriately. In the first step, $x_k \cdot k!$ many prefixes are used. For the correction steps, no additional prefix is used in the second case, and $|d| \cdot (k-i)!$ prefixes in the first case. The latter is less than

$$(3^{i-1} \cdot x_{max} \cdot (3^k \cdot l^k)^i + x_{max}) \cdot (k-i)! \leq 3^k \cdot x_{max} \cdot 3^{k^2} \cdot l^{k^2} \cdot k!.$$

Therefore, altogether less than $3^k \cdot x_{max} \cdot 3^{k^2} \cdot l^{k^2} \cdot (k+1)!$ many prefixes are necessary. On the other hand, there are 2^l prefixes of length l. So it is sufficient to choose l large enough so that $2^l \geq 3^k \cdot x_{max} \cdot 3^{k^2} \cdot l^{k^2} \cdot (k+1)!$ which is always possible since k and x_{max} are constants and on the right-hand side there is only a polynomial in l. □

5 Decidability of the Order of the Distances

From a practical as well as from a theoretical point of view, it is interesting to decide the order of magnitude of the distance between regular languages. In the definition of the distances, the number of words in the symmetric difference of the languages plays a crucial role. Summing up the distance of each of these words gives the distance of two languages. So, the question arises of how many words up to a certain length are in a given language. The function that counts the number of words of a fixed length n is called the *density function* (see, for example, [12,13] and the references therein). The function that counts the number of words up to a given length n is called *census function*. Clearly, both are closely related. So, first we deduce some decidability results for census functions from results on density functions shown in [12]. From these we derive the orders of possible distances between regular languages and show that the orders are decidable.

More formally, let L be a language over some alphabet Σ. Then its *density function* $\varrho_L : \mathbb{N} \to \mathbb{N}$ is defined as $\varrho_L(n) = |L \cap \Sigma^n| = |\{\, w \in L \mid |w| = n \,\}|$ and its census function $\text{cens}_L : \mathbb{N} \to \mathbb{N}$ as $\text{cens}_L(n) = \sum_{i=0}^{n} \varrho_L(i) = |\{\, w \in L \mid |w| \leq n \,\}|$.

The regular languages often are given in terms of minimal deterministic finite automata (DFA). For simplicity, in the following we write cens_A for $\text{cens}_{L(A)}$, where A is a DFA.

Proposition 11. *Let A be a minimal DFA. Then it is decidable whether* cens_A *is ultimately constant.*

Proof. The function cens_A is ultimately constant if and only if A accepts a finite language. The finiteness of a regular language is decidable by checking whether each accepting path of A is acyclic. □

In [12] the following gaps for the density of regular languages have been shown: (i) For any $k \geq 0$, there is no regular language whose density is in $\omega(n^k) \cap o(n^{k+1})$, and (ii) there is no regular language whose density is in $\omega(n^\ell)$ for all $\ell \geq 0$, and in $2^{o(n)}$. So, there is no density function of order $\Theta(\sqrt{n})$, $\Theta(n \log(n))$, or $\Theta(2^{\sqrt{n}})$. But note, the density of, say, the regular language $R_k = \{\, w \in \{a, b\}^* \mid |w|_a = k + 1 \text{ and } |w| \text{ is even} \,\}$ is $\varrho_{R_k}(n) \in \Theta(n^{k+1})$ if n is even, and $\varrho_{R_k}(n) = 0$ if n is odd. So, it is neither in $O(n^k)$ nor in $\Omega(n^{k+1})$.

In the following, we say that the density is *polynomial* if the function mapping n to $\max\{\, \varrho(i) \mid 0 \leq i \leq n \,\}$ is of order $\Theta(n^k)$, for some $k \geq 1$. It is *exponential*, if it is neither constant nor polynomial. In the latter case it has to be of the form $2^{\Omega(n)}$.

Since the density function of every regular language is either bounded by a constant, polynomial, or exponential, the next corollary follows.

Corollary 12. *The census function of every regular language is either ultimately constant, polynomial, or exponential.*

Proof. By definition we obtain the census function $\text{cens}(n)$ by summing up the densities up to n. Summing up polynomials of degree $k \geq 0$ gives a polynomial at most of degree $k + 1$. Similarly, summing up exponential functions of the form $2^{\Omega(n)}$ gives again an exponential function of that form. □

Though not explicitly stated, from the results in [12] it follows that it is decidable whether the density function of a regular language has an upper bound that is constant, polynomial, or exponential.

Moreover, the results in [12] imply a decision procedure for the question whether the census function of a regular language is polynomial or exponential, and for the former cases, whether it is of a certain degree.

Theorem 13. *Let A be a DFA. Then it is decidable whether* cens_A *is exponential or a polynomial. If it is a polynomial, the degree can be computed.*

Proof. If $L(A)$ is a unary language, then cens_A is either ultimately constant or linear. By Theorem 11 we can decide whether it is ultimately constant. If not by the results in [12] it can be decided whether ϱ_A is exponential or polynomial, where in the latter case the degree of the polynomial is computable. From the orders of the density we can derive the order of cens_A. □

Now we turn to the classes of parameterized prefix distances between regular languages. As mentioned before, for their computation the words in their symmetric difference are central, since only these contribute to the distance.

Let L_1 and L_2 be two languages. By $L_1 \oplus L_2$ we denote their symmetric difference. Let us recall briefly the observation $1 \leq \text{pref-}d(w, L_1) \leq |w| + |s|$, for $w \notin L_1$ and s being a shortest word in L_1.

Theorem 14. *Let L_1 and L_2 be two regular languages. Then it is decidable whether the parameterized prefix distance* $\text{pref-}D(n, L_1, L_2)$ *is ultimately constant.*

Proof. The family of regular languages is effectively closed under symmetric difference. So, a representation, say a DFA A, accepting $L_1 \oplus L_2$ can effectively be constructed from DFA accepting L_1 and L_2. Clearly, if $L_1 \oplus L_2$ is finite, then $\text{pref-}D(n, L_1, L_2)$ is ultimately constant. Conversely, if $L_1 \oplus L_2$ is infinite, then $\text{pref-}D(n, L_1, L_2)$ cannot be bounded by a constant, since all the infinitely many words in the symmetric difference contribute at least 1 to the distance. Now the theorem follows from the decidability of finiteness of regular languages. □

Theorem 15. *Let L_1 and L_2 be two regular languages. Then it is decidable whether the parameterized prefix distance* $\text{pref-}D(n, L_1, L_2)$ *is exponential.*

Proof. As in the proof of Theorem 14 we may assume without loss of generality that a DFA A accepting $L_1 \oplus L_2$ can effectively be constructed from L_1 and L_2. Moreover, one can decide whether $\text{pref-}D(n, L_1, L_2)$ is ultimately constant. So, assume that it is not.

Any word $|w|$ in the symmetric difference contributes at least 1 and at most $|w| + |s|$ to the distance, where s is the shortest word in the language w does not belong to. Therefore, we know $\text{cens}_A(n) \leq \text{pref-}D(n, L_1, L_2) \leq (c + n) \cdot \text{cens}_A(n)$, where c is the maximum of the lengths of the shortest words in L_1 and L_2. Since cens_A can only be ultimately constant, polynomial, or exponential, $\text{pref-}D(n, L_1, L_2)$ is exponential if and only if cens_A is exponential. Now the theorem follows from the possibility to decide whether cens_A is exponential. □

Theorem 16. *Let L_1 and L_2 be two regular languages and $k \geq 1$ be a constant. Then it is decidable whether the parameterized prefix distance* $\text{pref-}D(n, L_1, L_2)$ *belongs to $\Omega(n^k) \cap O(n^{k+1})$.*

Proof. It is decidable whether $\text{pref-}D(n, L_1, L_2)$ is ultimately constant or exponential. If it is neither of these, both census functions $\text{cens}_{L_1 \setminus L_2}$ and $\text{cens}_{L_2 \setminus L_1}$

are ultimately constant or polynomial. Theorem 13 shows that the degree k of the polynomial can be computed. With the fact, that each word $|w|$ contributes at least 1 and at most $|w| + |s|$ to the distance, where s is the shortest word in the language to which w does not belong, we derive pref-$D(n, L_1, L_2) \in \Omega(n^k) \cap O(n^{k+1})$. □

References

1. Aho, A.V., Sethi, R., Ullman, J.D.: Compilers: Principles, Techniques, and Tools. Addison-Wesley (1986)
2. Àlvarez, C., Jenner, B.: A very hard log space counting class. In: Structure in Complexity Theory Conference, pp. 154–168. IEEE Computer Society (1990)
3. Berstel, J., Reutenauer, C.: Rational Series and Their Languages. EATCS Monographs on Theoretical Computer Science. Springer (1988)
4. Choffrut, C., Pighizzini, G.: Distances between languages and reflexivity of relations. Theoret. Comput. Sci. 286, 117–138 (2002)
5. Eilenberg, S.: Automata, Languages, and Machines. Academic Press (1974)
6. Kruskal, J.B.: An overview of sequence comparison. In: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 1–44. Addison-Wesley (1983)
7. Mohri, M.: Finite-state transducers in language and speech processing. Computational Linguistics 23, 269–311 (1997)
8. Mohri, M.: On the use of sequential transducers in natural language processing. In: Finite-State Language Processing, pp. 355–381. MIT Press (1997)
9. Nerode, A., Kohn, W.: Models for hybrid systems: Automata, topologies, controllability, observability. In: Grossman, R.L., Ravn, A.P., Rischel, H., Nerode, A. (eds.) HS 1991 and HS 1992. LNCS, vol. 736, pp. 317–356. Springer, Heidelberg (1993)
10. Salomaa, A., Soittola, M.: Automata-Theoretic Aspects of Formal Power Series. Texts and monographs in computer science. Springer (1978)
11. Schützenberger, M.P.: Finite counting automata. Inform. Control 5, 91–107 (1962)
12. Szilard, A., Yu, S., Zhang, K., Shallit, J.: Characterizing regular languages with polynomial densities. In: Havel, I.M., Koubek, V. (eds.) MFCS 1992. LNCS, vol. 629, pp. 494–503. Springer, Heidelberg (1992)
13. Yu, S.: Regular languages. In: Handbook of Formal Languages, vol. 1, ch. 2, pp. 41–110. Springer, Berlin (1997)