Journal of Universal Computer Science, vol. 8, no. 2 (2002), 141-152
submitted: 15/9/01, accepted: 29/1/02, appeared: 28/2/02 © J.UCS

Additive Distances and Quasi-Distances Between Words[1]

Cristian S. Calude
(Computer Science Department, The University of Auckland, Private Bag 92109
Auckland, New Zealand
Email: cristian@cs.auckland.ac.nz)

Kai Salomaa[2]
(Department of Computing and Information Science, Queen's University
Kingston, Ontario, Canada K7L 3N6
Email: ksalomaa@cs.queensu.ca)

Sheng Yu[3]
(Department of Computer Science, The University of Western Ontario
London, Ontario, Canada N6A 5B7
Email: syu@csd.uwo.ca)

Abstract: We study additive distances and quasi-distances between words. We show that every additive distance is finite. We then prove that every additive quasi-distance is regularity-preserving, that is, the neighborhood of any radius of a regular language with respect to an additive quasi-distance is regular. Finally, similar results will be proven for context-free, computable and computably enumerable languages.

Key Words: some words
Category: F.1.1, F.4.3

1 Introduction

Let Σ be a finite alphabet. By the *neighborhood* of a word $w \in \Sigma^*$ of radius α with respect to a distance measure δ, we mean the set of all words u that have the distance measure $\delta(u, w)$ at most α. We denote this neighborhood by $E(\{w\}, \delta, \alpha)$. Naturally, the neighborhood of a language L of a radius α with respect to δ, denoted $E(L, \delta, \alpha)$, is the union of $E(\{w\}, \delta, \alpha)$ for all words $w \in L$. A distance δ is said to be finite if $E(\{w\}, \delta, \alpha)$ is finite for all $w \in \Sigma^*$ and $\alpha \geq 0$. Informally, δ is said to be additive if its measurement distributes over concatenation. We say that δ regularity-preserving (context-free-preserving, computable-preserving, computably enumerable-preserving) if $E(R, \delta, \alpha)$ is regular (context-free, computable, computably enumerable) for every regular (context-free, computable, computably enumerable) language R and radius $\alpha \geq 0$.

In this paper, we prove that every additive distance is finite. We prove that every additive distance (or quasi-distance) is regularity-preserving. We also show

[1] C. S. Calude, K. Salomaa, S. Yu (eds.). *Advances and Trends in Automata and Formal Languages. A Collection of Papers in Honour of the 60th Birthday of Helmut Jürgensen.*
[2] Work supported by the Natural Sciences and Engineering Research Council of Canada Grant OGP0147224
[3] Work supported by the Natural Sciences and Engineering Research Council of Canada Grant OGP0041630

that additive neighborhoods of any radius of context-free, computable and computably enumerable languages are, respectively, context-free, computable and computably enumerable. Examples of various additive and non-additive distance measures are also given in the paper.

The paper is organized as follows: In the next section we introduce the basic notation. In Section 3, we define distances and quasi-distances. Our results concerning finite, additive, and regularity-preserving distance measures are presented in Section 4. Additive neighborhoods of context-free and computable and computably enumerable languages are studied in Sections 5 and 6.

A preliminary but slightly different version of this paper has appeared in [Calude, Salomaa, Yu, 01].

2 Preliminaries

We assume that the reader is familiar with the basics of formal languages and finite automata in particular, cf. [Hopcroft and Ullman, 79, Salomaa, 73, Yu, 97]. Here we introduce the notation we will use in the later sections.

The symbol Σ denotes a finite alphabet and Σ^* the set of finite words over Σ. The empty word is denoted by λ and the length of a word $w \in \Sigma^*$ by $|w|$. The shuffle of words $u, v \in \Sigma^*$,

$$\omega(u, v) \subseteq \Sigma^*$$

is the set of all words $x_1 y_1 x_2 \ldots x_m y_m$ such that $u = x_1 \cdots x_m$, $v = y_1 \cdots y_m$, $x_i, y_i \in \Sigma^*$, $i = 1, \ldots, m$, $m > 0$. The catenation of languages $S, T \subseteq \Sigma^*$ is denoted by ST.

A deterministic finite automaton (DFA) is a five-tuple

$$A = (Q, \Sigma, \gamma, s, F)$$

where Q is the finite set of states, Σ is the finite alphabet, $s \in Q$ is the initial state, $F \subseteq Q$ is the set of final states, and $\gamma : Q \times \Sigma \to Q$ is the state-transition function. If A is defined as above except that γ is a function $Q \times \Sigma \to \mathcal{P}(Q)$ then we say that A is a nondeterministic finite automaton (NFA). (Here $\mathcal{P}(Q)$ is the set of subsets of Q.)

The state-transition relation γ of an NFA is extended in the natural way to a function $\hat{\gamma} : Q \times \Sigma^* \to \mathcal{P}(Q)$. We denote also $\hat{\gamma}$ simply by γ and the language accepted by A is $L(A) = \{w \in \Sigma^* \mid \gamma(s, w) \cap F \neq \emptyset\}$.

A context-free grammar is a four-tuple

$$G = (N, T, S, P)$$

where N is the finite nonterminal alphabet, T is the finite terminal alphabet, $N \cap T = \emptyset$, $S \in N$ is the initial nonterminal and P is the finite set of productions of the form $A \to w$, $A \in N$, $w \in (N \cup T)^*$.

The single step derivation relation of G, $\Rightarrow_G$, is defined by setting $u \Rightarrow_G v$ if we can write $u = u_1 A u_2$, $v = u_1 w u_2$, where $A \to w \in P$, $(u_1, u_2 \in (N \cup T)^*)$. We denote by $\Rightarrow_G^*$ the reflexive and transitive closure of $\Rightarrow_G$ and the language generated by the grammar G is $L(G) = \{w \in T^* \mid S \Rightarrow_G^* w\}$. A language is said to be context-free if it is generated by a context-free grammar.

A context-free grammar $G = (N, T, S, P)$ is said to be in Chomsky normal form if all productions of P are of the following forms $A \to BC$, $A \to a$, $A, B, C \in N$, $a \in T$. It is well known that any context-free language can be generated by a context-free grammar in Chomsky normal form.

A context-free grammar $G = (N, T, S, P)$ is said to be right-linear if the productions of P are of the forms $A \to bB$, $A \to b$ where $A, B \in N$, $b \in T$. Right-linear grammars generate exactly the regular languages.

3 Distances and Quasi-Distances

We want to measure the distance between distinct words of Σ^*. Let S be a set. We say that a function $\delta : S \times S \to [0, \infty)$ is a *distance* if it satisfies the following three conditions:

(D1) $\delta(x, y) = 0$ iff $x = y$, for all $x, y \in S$,
(D2) $\delta(x, y) = \delta(y, x)$, for all $x, y \in S$,
(D3) $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$, for all $x, y, z \in S$.

Condition (D3) is called the triangle-inequality. A function $\delta : S \times S \to [0, \infty)$ that satisfies (D2) and (D3) and the weaker condition

(D1') $\delta(x, x) = 0$, for all $x \in S$,

is called a *quasi-distance* on S. A quasi-distance allows the possibility that $\delta(x, y) = 0$, for $x \neq y$.

Note that if δ is a quasi-distance on S we can define an equivalence relation $\sim_\delta$ on S by setting $x \sim_\delta y$ iff $\delta(x, y) = 0$. Then the mapping δ' defined by $\delta'([x]_{\sim_\delta}, [y]_{\sim_\delta}) = \delta(x, y)$ is a distance on $S/\sim_\delta$. (Since δ satisfies the condition (D3) it follows that the value of $\delta'([x]_{\sim_\delta}, [y]_{\sim_\delta})$ does not depend on the representatives x and y.)

Let δ be a (quasi-) distance on S, $K \subseteq S$ and $\alpha \geq 0$. The *neighborhood of K of radius α* (with respect to δ) is

$$E(K, \delta, \alpha) = \{x \in S \mid (\exists y \in K)\ \delta(x, y) \leq \alpha\}.$$

A natural distance between words of the same length is the so called *Hamming distance.* Since we need to compare also words of different lengths, there is more than one natural way to extend Hamming distance.

Let $\#$ be a symbol not appearing in Σ and put $\Gamma = \Sigma \cup \{\#\}$. For $a, b \in \Gamma$ define

$$\Delta(a, b) = \begin{cases} 1, & \text{if } a \neq b, \\ 0, & \text{if } a = b. \end{cases}$$

Define $\Delta_n : \Gamma^n \times \Gamma^n \to \mathbb{N}$ by setting

$$\Delta_n(x_1 \cdots x_n, y_1 \cdots y_n) = \sum_{i=1}^{n} \Delta(x_i, y_i).$$

The *prefix-Hamming distance* δ_{pH} on Σ^* is defined as follows. Let $u, v \in \Sigma^*$. Then

$$\delta_{\mathrm{pH}}(u, v) = \begin{cases} \Delta_{|v|}(u\#^k, v), & \text{if } k = |v| - |u| \geq 0, \\ \Delta_{|u|}(u, v\#^k), & \text{if } k = |u| - |v| > 0. \end{cases}$$

The prefix-Hamming distance counts the number of distinct symbols in the first $\min\{|u|,|v|\}$ positions of the words u and v and adds to the result the length of the remaining suffix. It is easy to verify that δ_{pH} satisfies the triangle-inequality and, thus, it is a distance. On the other hand, this distance is not very useful from a practical point of view because inserting or deleting one letter can change the distance of given words by an arbitrary amount (depending on the length of the words).

A better extension is the function which considers all possible ways to pad both words and then takes the minimum of the obtained distances. Let $u, v \in \Sigma^*$. Then we define

$$\delta_{\mathrm{H}}(u,v) = \min\{\Delta_k(x,y) \mid k \geq \max\{|u|,|v|\}, \\ x \in \omega(u, \#^{k-|u|}), y \in \omega(v, \#^{k-|v|})\}. \tag{1}$$

Notice that for all $u, v \in \Sigma^*$, $\delta_H(u,v) \leq \max\{|u|,|v|\}$, and $\delta_{\mathrm{H}}(u,v) = \min\{\Delta_{|uv|}(x,y) \mid x \in \omega(u, \#^{|v|}),\ y \in \omega(v, \#^{|u|})\}$.

In general, $\delta_{\mathrm{H}}(u,v) \neq \Delta_{\max\{|u|,|v|\}}(x,y)$, for every $x \in \omega(u, \#^{\max\{|u|,|v|\}-|u|})$, $y \in \omega(v, \#^{\max\{|u|,|v|\}-|v|})$. For example, take $u = abab, v = baba$ and observe that $\omega(u, \#^0) = \{u\}, \omega(v, \#^0) = \{v\}$, $\Delta_4(u,v) = 4 > \delta_{\mathrm{H}}(u,v) = \Delta_5(u\#, \#v) = 2$.

It is convenient to look at (1) as a process. Consider changing a word into another word by means of the following three types of *edit steps* ([Manber, 89]): a) *insert*—insert a character into a word, b) *delete*—delete a character from a word, c) *replace*—replace one character with a different character. Edit steps can be applied in any order. For example, to change the word *abab* into *baba* we can use rule c) (replace) four times and we get $bbab, baab, babb, baba$. We can be more efficient by deleting the first character of *abab* to get *bab*, then insert *a* at the end, so with only two edit steps we obtain *baba*. As we have seen below, $\delta_{\mathrm{H}}(abab, baba) = 2$; it can be obtained by first constructing the extended words $abab\#$ and $\#baba$ and then computing their Δ_5 distance. In fact, we have:

Lemma 1. *For all words u, v, $\delta_{\mathrm{H}}(u,v)$ coincides with the minimal number of edit steps necessary to change u into v.*[4]

Corollary 2. *The function δ_{H} satisfies (D1)–(D3).*

The function δ_{H} is a distance by Corollary 2; as it extends Hamming's distance it is appropriate to call it the *shuffle-Hamming distance*.

An immediate property of the shuffle-Hamming distance follows: insertions and deletions of the special symbol # do not count.

Lemma 3. *For all $u, v \in \Sigma^*$, and $i \geq 0$, $\delta_{\mathrm{H}}(u,v) = \delta_{\mathrm{H}}(\bar{u}, \bar{v})$, for all $\bar{u} \in \omega(u, \#^i), \bar{v} \in \omega(v, \#^i)$.*

Other possible distances can be obtained by varying edit steps (e.g., allowing adjacent characters in one word to be interchanged while copied to the other word) or by assigning cost functions to edit steps (e.g., capturing the idea that the cost of replacing a character is less than the combined costs of deletion and insertion). See [Calude and Calude, 83] for more examples of discrete distances.

[4] This number is called the *Levenstein distance* or *edit-distance*; see [Stephen, 94], pp 40–41 and [Cormen et al., 90], pp 325–326; it has been introduced by by Levenstein [Levenstein, 65] and, independently, by Ulam [Ulam, 86].

4 Neighborhoods of Regular Languages

Let L be a regular language over Σ. We are interested in the following question: Which conditions the distance δ should satisfy in order to guarantee that all the languages $E(L, \delta, \alpha)$, $\alpha \geq 0$, are regular? We say that a distance δ is *regularity-preserving* if $E(L, \delta, \alpha)$ is a regular language for all regular languages L and $\alpha \geq 0$.

It is fairly straightforward to construct examples of distances on Σ^* that are not regularity-preserving. Here is such an example.

Example 1. Given a language $L \subseteq \Sigma^*$ we can define

$$\delta_L(u, v) = \begin{cases} 0, & \text{if } u = v, \\ 1/2, & \text{if } u \neq v, u, v \in L \\ 1, & \text{if } u \neq v, \text{ and at least one of } u, v \text{ is not in } L. \end{cases}$$

It is easy to verify that δ_L is a distance. Then by choosing L to be some non-regular language and $w \in L$, we have $E(\{w\}, \delta_L, 1/2) = L$. □

Clearly we need to impose some additional conditions on the distance δ. Note that the distance in Example 1 has the property that for $n \geq 0$ and $\alpha \geq 1/2$, the inequality $\delta(u, a^n b^n) \leq \alpha$ has infinitely many solutions. Hence, the following finiteness requirement seems to be a suitable candidate to guarantee that a distance is regularity-preserving.

We say that a (quasi-) distance δ on Σ^* is *finite* if for all $w \in \Sigma^*$ and $\alpha \geq 0$, the set $E(\{w\}, \delta, \alpha)$ is finite.

Both the shuffle-Hamming distance and the prefix-Hamming distance considered above are clearly finite. The following example shows that finiteness of a distance δ is, unfortunately, not sufficient to guarantee that δ is regularity-preserving.

Example 2. Let $\Sigma = \{a, b, c\}$. By slightly modifying the prefix-Hamming distance δ_{pH} we construct a finite distance δ on Σ^* that is not regularity-preserving.

For $u, v \in \Sigma^*$ we define

$$\delta(u, v) = \begin{cases} 3/2, & \text{if } u = a^n b a^n, v = a^n c a^n, n \geq 0, \text{ or vice versa,} \\ \delta_{\text{pH}}(u, v), & \text{otherwise.} \end{cases}$$

Clearly δ satisfies the conditions (D1) and (D2), so in order to show that it is a distance it is sufficient to verify the triangle-inequality. Assuming that (D3) does not hold, we must have $x, y, z \in \Sigma^*$ such that

$$\delta(x, z) > \delta(x, y) + \delta(y, z). \tag{2}$$

Since for all $u, v \in \Sigma^*$, $\delta(u, v) \geq \delta_{\text{pH}}(u, v)$ and δ_{pH} is a distance, it follows that if (2) holds, then necessarily $\delta(x, z) \neq \delta_{\text{pH}}(x, z)$, that is, $x = a^n b a^n$, $z = a^n c a^n$, $n \geq 0$, or vice versa. Thus $\delta(x, z) = 3/2$, and (2) implies that $\delta(x, y) = 0$ or $\delta(y, z) = 0$. Both possibilities directly yield a contradiction.

Also, δ is finite since for any $\alpha \geq 2$ and $w \in \Sigma^*$ we have $E(\{w\}, \delta, \alpha) = E(\{w\}, \delta_{\text{pH}}, \alpha)$.

To see that δ is not regularity-preserving choose $L = a^*ba^*$. Then

$$E(L, \delta, 3/2) - E(L, \delta, 1) = \{a^n ca^n \mid n \geq 0\},$$

which implies that at least one of the languages $E(L, \delta, 3/2)$ and $E(L, \delta, 1)$ is not regular. □

The above example shows that we need to look for stronger restrictions for regularity-preserving distances. Since elements of Σ^* have a unique decomposition into subwords (of given length) it is perhaps reasonable to assume that the distances should "respect" such decompositions. Thus we say that a (quasi-) distance δ on Σ^* is *additive* if always when $w = w_1w_2$ ($w_1, w_2 \in \Sigma^*$) we have for all $\alpha \geq 0$,

$$E(\{w\}, \delta, \alpha) = \bigcup_{\beta_1+\beta_2=\alpha} E(\{w_1\}, \delta, \beta_1)E(\{w_2\}, \delta, \beta_2). \tag{3}$$

First we observe that an additive distance is always finite. Note that an additive quasi-distance δ need not be finite. If, for some $b \in \Sigma$, $\delta(b, \lambda) = 0$, then any δ-neighborhood is necessarily infinite.

Lemma 4. *Every additive distance is finite.*

Proof. Let δ be an additive distance on Σ^*. By (3), for any $w = b_1 \cdots b_k$, $b_i \in \Sigma$, $i = 1, \ldots, k$, $E(\{w\}, \delta, \alpha)$ is contained in the catenation of the languages $E(\{b_1\}, \delta, \alpha)$, …, $E(\{b_k\}, \delta, \alpha)$. Thus, it is sufficient to show that $E(\{b\}, \delta, \alpha)$ is finite for $b \in \Sigma$ and $\alpha \geq 0$.

Let $u = c_1 \cdots c_m$, $c_i \in \Sigma$, be an arbitrary word of Σ^*. The additivity condition implies that $u \in E(\{b\}, \delta, \alpha)$ iff there exists $i \in \{1, \ldots, m\}$ such that

$$\delta(b, c_i) + \sum_{j\in\{1,\ldots,m\},\ j\neq i} \delta(\lambda, c_j) \leq \alpha. \tag{4}$$

There exist only a finite number of words $u = c_1 \cdots c_m$ that satisfy the above inequality. □

Both the prefix-Hamming distance and the shuffle-Hamming distance are additive.

Proposition 5. *The distances δ_{pH} and δ_{H} defined on an alphabet Σ are additive.*

Proof. We show that δ_{H} is additive as the proof for the distance δ_{pH} is simpler.

Let $w = w_1w_2$ be an arbitrary decomposition of a word $w \in \Sigma^*$. We show that for every $u \in \Sigma^*$,

$$u \in E(\{w_1w_2\}, \delta_{\mathrm{H}}, \alpha) \text{ iff } u \in \bigcup_{\beta_1+\beta_2=\alpha} E(\{w_1\}, \delta_{\mathrm{H}}, \beta_1)E(\{w_2\}, \delta_{\mathrm{H}}, \beta_2).$$

Assume $\delta_{\mathrm{H}}(u, w_1w_2) \leq \alpha$. As edit steps (in the process of changing a word into another word) can be applied in any order, we can start the process of changing u into w_1w_2 in such a way to obtain first w_1 from a prefix u_1 of u, and then w_2 (from the remaining suffix u_2 of u). Consequently, $\delta_{\mathrm{H}}(u_1, w_1) + \delta_{\mathrm{H}}(u_2, w_2) = \delta_{\mathrm{H}}(u, w_1w_2) \leq \alpha$. Conversely, if $u_i \in E(\{w_i\}, \delta_{\mathrm{H}}, \beta_i)$, $i = 1, 2$, $\beta_1 + \beta_2 \leq \alpha$, then we have $\delta_{\mathrm{H}}(u_1u_2, w_1w_2) \leq \delta_{\mathrm{H}}(u_1, w_1) + \delta_{\mathrm{H}}(u_2, w_2) \leq \alpha$. □

From Example 2 we know that a finite distance need not preserve regularity. Below we show that, on the other hand, additivity is a sufficient condition to guarantee that even a quasi-distance preserves regularity. Note that, as observed above, an additive quasi-distance need not be finite. First we prove the following lemma.

Lemma 6. *Assume that δ is an additive quasi-distance on Σ^*.*

(i) *For each $b \in \Sigma$ and $\alpha \geq 0$, $E(b, \delta, \alpha)$ is regular.*
(ii) *Let $b \in \Sigma$ and $\alpha \geq 0$ be fixed. There exists an integer k and numbers $0 = \alpha_1 < \ldots < \alpha_k = \alpha$ such that*

$$E(b, \delta, \alpha_i), \quad i = 1, \ldots, k,$$

are all the distinct neighborhoods of b having radius at most α.

Proof. (i) Let $u = c_1 \cdots c_m$, $m \geq 0$, $c_i \in \Sigma$, $i = 1, \ldots, m$. As in the proof of Lemma 4 it follows that $u \in E(b, \delta, \alpha)$ iff the inequality (4) holds. (Note that, in contrast to Lemma 4, δ is now only a quasi-distance, so this does not imply the finiteness of the neighborhood.)

Denote

$$\Theta = \{d \in \Sigma \mid \delta(d, \lambda) = 0\}.$$

Let Ψ be the set of finite multisets of elements of Σ,

$$\{c_i, c_{j_1}, \ldots, c_{j_r}\}$$

such that $\delta(\lambda, c_{j_l}) \neq 0$, $l = 1, \ldots, r$ and

$$\delta(b, c_i) + \sum_{l=1}^{r} \delta(\lambda, c_{j_l}) \leq \alpha.$$

Then $u = c_1 \cdots c_m$ satisfies the inequality (4) iff u is the shuffle of a sequence obtained by listing the elements of a multiset belonging to Ψ (in arbitrary order) and a word in Θ^*. The shuffle of a finite language and a regular language is always regular.

(ii) In the construction above the elements of the multisets belonging to Ψ completely determine the neighborhoods of radius at most α around b. Thus as the radii α_s, $s = 1, \ldots, k$, we can simply take all the (distinct) sums $\delta(b, c_i) + \sum_{l=1}^{r} \delta(\lambda, c_{j_l})$ where the multiset $\{c_i, c_{j_1}, \ldots, c_{j_r}\}$ belongs to Ψ. (Note that Ψ is a finite collection of multisets.) □

The above construction implies that Lemma 6 (ii) can be written in the following stronger form:

Corollary 7. *Assume that δ is an additive quasi-distance on Σ^* and let $b \in \Sigma$ and $\alpha \geq 0$ be fixed. Then we can write*

$$E(b, \delta, \alpha) = R_1 \cup \ldots \cup R_k,$$

where $0 = \alpha_1 < \ldots < \alpha_k = \alpha$ and $R_i = \{w \in \Sigma^* \mid \delta(b, w) = \alpha_i\}$, $i = 1, \ldots, k$, is regular.

Proof. Without loss of generality we can assume that the numbers α_i in Lemma 6 (ii) are chosen so that there exists $w_i \in \Sigma^*$ with $\delta(b, w_i) = \alpha_i$, $i = 1, \ldots, k$. Let R_i, $i = 1, \ldots, k$, be as above. By Lemma 6 (ii), $R_i = E(b, \delta, \alpha_i) - E(b, \delta, \alpha_{i-1})$, $i = 2, \ldots, k$, and $R_1 = E(b, \delta, 0)$. By Lemma 6 (i), these sets are regular. □

Now we are ready to prove the main result of this section.

Theorem 8. *Assume that δ is an additive quasi-distance on Σ^* and let $L \subseteq \Sigma^*$ be regular. Then $E(L, \delta, \alpha)$ is regular for all $\alpha \geq 0$.*

Proof. Let $\alpha \geq 0$ be fixed and let $A = (Q, \Sigma, \gamma, s, F)$ be a DFA such that $L = L(A)$. Without loss of generality we can assume that the initial state s is not reachable from any other state.

By Corollary 7, for each $b \in \Sigma$ we can write

$$E(b, \delta, \alpha) = R_1^b \cup \ldots \cup R_{k(b)}^b,$$

where

$$R_j^b = \{w \in \Sigma^* \mid \delta(w, b) = \alpha_j^b\}, \quad 0 \leq \alpha_j^b \leq \alpha,$$

is regular, $j = 1, \ldots, k(b)$. Denote $D' = \{\alpha_j^b \mid b \in \Sigma, 1 \leq j \leq k(b)\}$ and

$$D = \{\beta \leq \alpha \mid \beta = \beta_1 + \ldots + \beta_r, \beta_i \in D', 1 \leq i \leq r\}.$$

We construct an NFA $B = (Q_B, \Sigma, \gamma_B, s_B, F_B)$ such that

$$L(B) = E(L(A), \delta, \alpha).$$

Choose $Q_B = Q \times D$, $s_B = (s, 0)$ and

$$F_B = \begin{cases} F \times D \cup \{s_B\} & \text{if } \lambda \in E(L(A), \delta, \alpha) \\ F \times D, & \text{otherwise.} \end{cases}$$

The transition relation γ_B is defined as follows. Let $q \in Q$, $\beta \in D$ and $b \in \Sigma$. Then

$$(q', \beta + \alpha_j^b) \in \gamma_B((q, \beta), b) \tag{5}$$

for every $q' \in \gamma(q, R_j^b)$, $1 \leq j \leq k(b)$, such that $\beta + \alpha_j^b \leq \alpha$. (Here $\gamma(q, R_j^b) = \{\gamma(q, v) \mid v \in R_j^b\}$.) Since R_j^b is regular, the set $\gamma(q, R_j^b)$ ($\subseteq Q$) can even be effectively determined.

Let $w = b_1 \cdots b_m$, $m \geq 1$, $b_i \in \Sigma$, $i = 1, \ldots, m$. Since δ is additive

$$\begin{aligned} & w \in E(L(A), \delta, \alpha) \text{ iff } (\exists u \in L(A)) \text{ such that} \\ & u \in \bigcup_{\beta_1 + \ldots + \beta_m = \alpha} E(b_1, \delta, \beta_1) \cdots E(b_m, \delta, \beta_m). \end{aligned} \tag{6}$$

In the transitions (5), on input b the first component of the states of B simulates the computation of A on an arbitrary (nondeterministically chosen) word of $v \in R_j^b$, and in the second component we correspondingly increment the distance by $\alpha_j^b = \delta(b, v)$. By observation (6), some sequence of the nondeterministic choices on input $w = b_1 \cdots b_m$ leads to an accepting state of F_B iff w is in $E(L(A), \delta, \alpha)$. By the choice of the set F_B, the NFA B accepts λ if and only if $\lambda \in E(L(A), \delta, \alpha)$. □

5 Additive Neighborhoods of Context-Free Languages

We show that also the family of context-free languages is closed under additive quasi-distances, that is, for any context-free language L, additive quasi-distance δ and $\alpha \geq 0$, the neighborhood $E(L, \delta, \alpha)$ is context-free. In this case a construction (following the idea of the proof of Theorem 8) of a pushdown automaton that nondeterministically simulates the possible computations on all inputs in the neighborhood would not work due to the fact that the different computations could use the stack in very different ways. Our proof uses a grammatical approach where a context-free grammar distributes the distance bound into different parts of a generated word. Additivity is obviously necessary for the construction to work and the construction relies essentially on Lemma 6 (ii).

Theorem 9. *Let δ be an additive quasi-distance on Σ^*. For every context-free language $L \subseteq \Sigma^*$ and $\alpha \geq 0$, the neighborhood $E(L, \delta, \alpha)$ is context-free.*

Proof. Let $G = (N, \Sigma, S, P)$ be a Chomsky normal form grammar generating the language L. Denote

$$L'_{b,\beta} = E(b, \delta, \beta), \;\; b \in \Sigma, \beta > 0.$$

By Lemma 6 (ii), there exist an integer k_b and values $0 = \alpha_{b,1} < \ldots < \alpha_{b,k_b} = \alpha$ such that for any $\beta \leq \alpha$, $L'_{b,\beta} = L'_{b,\alpha_{b,i}}$ for some $1 \leq i \leq k_b$. Without loss of generality we can assume that the sequence of sets $L'_{b,\alpha_{b,i}}$, $1 \leq i \leq k_b$, is strictly increasing for a fixed $b \in \Sigma$ (just eliminate from the $\alpha_{b,j}$ sequence possible unnecessary values that do not strictly increase the neighborhood). Now we define

$$L_{b,\alpha_{b,1}} = L'_{b,\alpha_{b,1}} \text{ and } L_{b,\alpha_{b,i}} = L'_{b,\alpha_{b,i}} - L'_{b,\alpha_{b,i-1}}, \; i = 2, \ldots, k_b. \qquad (7)$$

By Lemma 6 (i) the languages $L_{b,\alpha_{b,i}}$, $b \in \Sigma$, $1 \leq i \leq k_b$ are regular and let

$$G_{b,i} = (N_{b,i}, \Sigma, S_{b,i}, P_{b,i})$$

be a right-linear grammar generating $L_{b,\alpha_{b,i}}$. Again without loss of generality we can assume that the non-terminal alphabets of G and grammars $G_{b,i}$, $b \in \Sigma$, $1 \leq i \leq k_b$ are all pairwise distinct.

Denote

$$I = \{\alpha_{b,i} \mid b \in \Sigma, 1 \leq i \leq k_b\}$$

and let

$$K = \{x \in [0, \alpha] \mid x = i_1 + \cdots + i_l, \; i_j \in I, j = 1, \ldots, l, (l \in \mathbb{N})\}.$$

Since I is finite it follows that also K is finite.

We define a context-free grammar $G' = (N', \Sigma, S', P')$ where

$$N' = N \times K \cup \bigcup_{b \in \Sigma, 1 \leq i \leq k_b} N_{b,i},$$

$S' = (S, \alpha)$, (note that $\alpha = \alpha_{b,k_b}$, for all $b \in \Sigma$ and hence $(S, \alpha) \in N \times K$), and

$$P' = P_1 \cup P_2 \cup P_3,$$

where

$$P_1 = \{(A, \beta) \to (B, \beta_1)(C, \beta_2) \mid A \to BC \in P,\ \beta, \beta_1, \beta_2 \in K, \beta_1 + \beta_2 \leq \beta\},$$

$$P_2 = \{(A, \beta) \to S_{b,i} \mid A \to b \in P,\ \alpha_{b,i} \leq \beta \in K\},$$

and

$$P_3 = \bigcup_{b \in \Sigma, 1 \leq i \leq k_b} P_{b,i}.$$

We claim that $L(G) = E(L, \delta, \alpha)$. Let $w \in \Sigma^*$ be arbitrary. Since δ is additive, we have

$$\begin{aligned} w \in E(L, \delta, \alpha) \text{ iff } & w = u_1 \cdots u_n,\ \ u_i \in E(b_i, \delta, \beta_i), \\ & b_i \in \Sigma, u_i \in \Sigma^*, i = 1, \ldots, n, \\ & \text{where } b_1 \cdots b_n \in L,\ \beta_1 + \ldots + \beta_n = \alpha. \end{aligned} \tag{8}$$

For any $b_i \in \Sigma$, the languages $L'_{b_i, \alpha_{b_i,1}}, \ldots, L'_{b_i, \alpha_{b_i, k_{b_i}}}$ are exactly all the distinct neighborhoods of b_i of radius at most α. In (8), the radii β_i ($1 \leq i \leq n$), do not necessarily belong to the finite set $\{\alpha_{b_i,1}, \ldots, \alpha_{b_i,k_{b_i}}\}$, but there exists $j \in \{1, \ldots, k_{b_i}\}$ such that $u_i \in L_{b_i, \alpha_{b_i,j}}$ since

$$\bigcup_{j=1,\ldots,k_{b_i}} L_{b_i, \alpha_{b_i,j}} = E(b_i, \delta, \alpha).$$

If j is chosen as above, from (7) it follows that $\delta(u_i, b_i) = \alpha_{b_i,j}$ and thus $u_i \in E(b_i, \delta, \beta_i)$ implies

$$\alpha_{b_i,j} \leq \beta_i. \tag{9}$$

Now the grammar G' can generate the string w as follows. The productions of P_1 simulate the derivation of $b_1 \cdots b_n$ according to G and nondeterministically distribute the correct "weights" $\alpha_{b_i,j}$, $j \in \{1, \ldots, k_{b_i}\}$, to the nodes that correspond to the terminal symbols b_i (in the simulated derivation of G). The fact that $\beta_1 + \ldots + \beta_n = \alpha$ and equation (9) guarantee that this is always possible. Using a rule of P_2, in the node "corresponding to b_i" the grammar G' continues with the initial nonterminal of the grammar $G_{b_i,j}$ and using the rules of P_3 it can generate the word $u_i \in L_{b_i, \alpha_{b_i,j}}$.

For the converse inclusion, using the definition of the productions of G', we see that if G' generates a string w, then necessarily the right side of (8) holds for w with $\beta_1 + \ldots + \beta_n = \alpha$ replaced by $\beta_1 + \ldots + \beta_n \leq \alpha$. Note that the productions of P_1 and P_2 guarantee that the cumulative "weight", that is, the sum of the second components of the nonterminals in a sentential form of G' is nonincreasing at any derivation step. Since δ is additive, this means that necessarily $w \in E(b_1 \cdots b_n, \delta, \beta_1 + \ldots + \beta_n) \subseteq E(b_1 \cdots b_n, \delta, \alpha)$. □

As can be expected, in Theorem 9 the additivity assumption is necessary and context-free languages are not closed under extensions generated by general distances: for example, choose a non-context-free language in Example 1.

Furthermore, by modifying the distance used in Example 2 we see that context-free languages are not closed even under finite distances. Let $\Sigma = \{a, b, c\}$ and define

$$\delta'(u,v) = \begin{cases} 3/2, & \text{if } u = a^n ba^n ba^n, v = a^n ca^n ca^n, n \geq 0, \text{ or vice versa,} \\ \delta_{\mathrm{pH}}(u,v), & \text{otherwise} \end{cases}$$

where $u, v \in \Sigma^*$ and δ_{pH} is the prefix-Hamming distance. Exactly as in Example 2 it is verified that δ' is a finite distance, and by choosing $L = a^*ba^*ba^*$ we have

$$E(L, \delta', 3/2) \cap a^*ca^*ca^* = \{a^n ca^n ca^n \mid n \geq 0\},$$

which is not context-free. The above shows actually that even for a regular language the neighborhood with respect to a finite distance is not necessarily context-free.

6 Neighborhoods of Computable and Computably Enumerable Languages

In this section we show that the neighborhoods of computable (computably enumerable) languages are computable (computably enumerable) not only in case of additive quasi-distances, but also for finite ones.

Theorem 10. *Let δ be quasi-distance on Σ^* that is additive or finite. For every computable (computably enumerable) language $L \subseteq \Sigma^*$ and $\alpha \geq 0$, the neighborhood $E(L, \delta, \alpha)$ is computable (computably enumerable).*

Proof. Let $\alpha > 0$ be fixed. Assume first that L is computable. Clearly, $u \in E(L, \delta, \alpha)$ iff $L \cap E(\{u\}, \delta, \alpha) \neq \emptyset$. If δ is a finite quasi-distance, then the set $L \cap E(\{u\}, \delta, \alpha)$ is finite, so $E(L, \delta, \alpha)$ is computable.

Assume now that δ is an additive quasi-distance. We first prove that Lemma 6 (ii) holds true not only for elements in the alphabet Σ, but for arbitrary words $w \in \Sigma^*$. Indeed, in view of Lemma 6, for any given $b \in \Sigma$, we have the sequence $0 = \alpha_{b,1} < \ldots < \alpha_{b,k_b} = \alpha$ such that $E(b, \delta, \alpha_{b,j})$ are all the distinct neightborhoods of b of radius at most α. We can choose each $\alpha_{b,j}$ value to be maximal, that is, if $\alpha_{b,j} < x < \alpha_{b,j+1}$, then $E(b, \delta\alpha_{b,j}) = E(b, \delta, x)$.

Now consider an arbitrary $w \in \Sigma^*$, let $w = a_1 \cdots a_n$ ($a_i \in \Sigma$). Since δ is additive, for any $\gamma < \alpha$ we have

$$E(w, \delta, \gamma) = \bigcup_{\beta_1 + \ldots + \beta_n = \gamma} E(a_1, \delta, \beta_1) \cdots E(a_n, \delta, \beta_n). \tag{10}$$

In the right side each neighborhood $E(a_i, \delta, \beta_i)$ is equal to $E(a_i, \delta, \alpha_{a_i,j})$ with the property that $\alpha_{a_i,j} \leq \beta_i < \alpha_{a_i,j+1}$. This means that there exist only a finite number (depending on w) of distinct right sides of (10), that is, a finite number of distinct neighborhoods $E(w, \delta, \gamma)$ where $\gamma < \alpha$. Hence $L \cap E(\{u\}, \delta, \alpha)$ is computable, so $E(L, \delta, \alpha)$ is computable.

Finally, a simple dovetailing argument shows that the neighborhood $E(L, \delta, \alpha)$ is computably enumerable provided δ is a finite quasi-distance or an additive quasi-distance (in the last case we use Theorem 8 to note that for every word w, $E(\{w\}, \delta, \alpha)$ is regular).

References

[Calude and Calude, 83] C.S. Calude, E. Calude: On some discrete metrics, *Bull. Math. Soc. Sci. Math. R. S. Roumanie (N. S.)* 27 (75) (1983), 213–216.

[Calude, Salomaa, Yu, 01] C.S. Calude, K. Salomaa, S. Yu: Metric lexical analysis, in O. Boldt, H. Jürgensen (eds.) *Automata Implementation*, Lectures Notes in Computer Science 2214, Springer-Verlag, Heidelberg, 2001, 48–59.

[Cormen et al., 90] T.H. Cormen, C.E. Leiserson, R.L. Rivest: *Introduction to Algorithms,* MIT Press, Cambridge, MA, 1990.

[Hopcroft and Ullman, 79] J.E. Hopcroft, J.D. Ullman: *Introduction to Automata Theory, Languages, and Computation,* Addison-Wesley, Reading, MA, 1979.

[Levenstein, 65] V.I. Levenstein: Binary codes capable of correcting deletions, insertions, and reversals, *Doklady Akademii nauk SSSR* 163, 4 (1965), 845–848. (Russian). Also in *Cybernetics and Control Theory* 10, 8 (1966), 707–710.

[Manber, 89] U. Manber: *Introduction to Algorithms—A Creative Approach*, Addison-Wesley, Reading, MA, 1989.

[Salomaa, 73] A. Salomaa: *Formal Languages,* Academic Press, New York, 1973.

[Stephen, 94] G.A. Stephen: *String Searching Algorithms*, World Scientific, Singapore, 1994.

[Yu, 97] S. Yu: Regular languages. In: *Handbook of Formal Languages, Vol. I.* (G. Rozenberg, A. Salomaa, eds.) pp. 41–110, Springer-Verlag, Berlin, 1997.

[Ulam, 86] S. Ulam: Some ideas and prospects in biomathematics, *The Annual Review of Biophysics and Bioengineering*, 1 (1972), 277–292. Reprinted in S. Ulam: *Science, Computers, and People* (M. C. Reynolds, G.- C. Rota, eds.), Birkhäuser, Boston, 1986, 115-136.