数理解析研究所講究録
第 1769 巻 2011 年 140-143

On inequalities between subword histories

Szilárd Zsolt Fazekas,
Kyoto Sangyo University
email:szilard.fazekas@gmail.com

June 10, 2011

By taking out letters from a word we get a subword. Both continuous subwords (also called factors or simply subwords) and scattered subwords were extensively studied. In [4] the authors introduced Parikh matrices, structures that contain more information about the words than Parikh vectors, which tell us only the number of different letters building the word. In [5] the notion of subword histories appeared and has been developed into a powerful tool in the investigation of relations between certain scattered subwords of a given word. In particular, several characteristic equalities and inequalities regarding sums of subword occurrences were presented, perhaps most notably the Cauchy inequality for words [5]. The decidability of equalities between subword histories was settled with a positive answer. This paper tries to answer the question about the decidability of inequalities between subword histories and succeeds in giving partial results, that is, certain cases where the inequalities hold and an algorithm to decide whether a subword inequality belongs to one of these particular cases. By alphabet we mean a set $\Sigma = \{a_1, a_2, .., a_n\}$. A word over Σ is a finite sequence of elements of Σ. The set of all words over Σ is denoted by Σ^*. A word $u = a_1a_2...a_m$ is a scattered subword of $w = b_1b_2...b_n$ if there is an increasing vector of indices $I = (i_1, i_2, .., i_m)$ such that $a_j = b_{i_j}, 1 \leq j \leq m$. In this case we will call the vector I an *occurrence* of u in w. We say that two occurrences $I = (i_1, .., i_m), J = (j_1, .., j_m)$ are different if they differ in at least one position, that is $\exists k : 1 \leq k \leq m$ such that $i_k \neq j_k$. By writing $|w|_u$ we mean the number of different occurrences of u in w.

From now on we will use the term *subword inequality* (SI) rather than the longer *inequality between subword histories*, and we mean basically the same, except for the coefficients of the terms. A SI is of the form:

$$\sum_{i=1}^{m} \alpha_i |w|_{u_i} \leq \sum_{j=1}^{n} \beta_j |w|_{v_j}$$

where the α's and β's are positive integers, the *coefficients* of the terms. For the sake of simplicity we will write the above SI as $\sum_{i=1}^{m} \alpha_i u_i \leq \sum_{j=1}^{n} \beta_j v_j$.

We start out by characterizing some restricted forms of subword inequalities.

The results are then combined in Theorem 3, which is our main result. As we mentioned earlier, this result is a one-way implication saying that certain types of subword inequalities hold. Although the reverse is not proved, our conjecture is that it is true, i.e. only the described cases yield inequalities that hold for any word.

In [5] the authors give an example of a SI which is true for any word:

$$baab < bab + baaab$$

It turns out that this example encompasses the very essence of the problem. In fact, all SIs that are "extended" versions of the one above hold for any word. We will elaborate in this section on what extended in the previous sentence exactly means. First we examine the inequalities where both sides comprise exactly one term.

Theorem 1. *For any two words $u, v \in \Sigma^*$ with $u \neq v$ there exist $w_1, w_2 \in \Sigma^*$ such that:*

- $|w_1|_u < |w_1|_v$ *and*
- $|w_2|_u > |w_2|_v$

We saw that inequalities between monomial subword histories, i.e. of the form $u \leq v$, hold if and only if $u = v$. Let us continue with the case when the left hand side has one term and the right hand side has two.

Lemma 1. *A SI of the form $z \leq u+v$ holds if and only if for some $x_1, x_2 \in \Sigma^*$ and $a \in \Sigma$:*

- $z = x_1 a x_2$
- $u = x_1 x_2$
- $v = x_1 a^2 x_2$

The decomposition in Lemma 2 is not unique for a given left hand side term. For example, if the term *baabba* is on the left hand side, we can choose the triple (x_1, a, x_2) to be (ba, a, bba) or (baa, b, ba), respectively. The resulting SIs (with dots marking the decomposition):

- $ba.a.bba \leq ba.bba + ba.aa.bba$
- and $baa.b.ba \leq baa.ba + baa.bb.ba$ hold in both cases.

In the proof of the previous lemma we saw that whenever the terms are identical except for one block, the SI reduces to an inequality between binomial coefficients. Let's take, for instance,

$$b.a.b + b.aaa.b \leq bb + b.aa.b + b.aaaa.b$$

It becomes clear that this inequality holds when we express it in terms of binomial coefficients:

$$\binom{n}{1} + \binom{n}{3} \leq \binom{n}{0} + \binom{n}{2} + \binom{n}{4}$$

In general, using some basic properties of binomial coefficients, we can extend the previous lemma to multiple terms on both sides.

Lemma 2. *Let us consider a set of inequalities* $u_i \leq v_i + v_{i+1}, 1 \leq i \leq n$. *If all these inequalities hold and in addition to this,* $v_{i+1} \leq u_i + u_{i+1}$ *for all* $1 \leq i \leq n-1$, *then*

$$u_1 + u_2 + .. + u_n \leq v_1 + v_2 + .. + v_{n+1}$$

also holds.

In general for $ba^ib \leq ba^jb + ba^kb$, where $j < i < k$, the term with k a's will be equal to the one with i a's when the containing word will have $i+k$ a's so we have to set the coefficient of the shorter term in such a way that it compensates for the cases when the containing word has less than $i+k$ a's in the middle.

Lemma 3. *A SI of the form* $\alpha z \leq \beta_1 u + \beta_2 v$ *holds if and only if there exist* $x_1, x_2 \in \Sigma^*, a \in \Sigma$ *and* $0 \leq j < i < k$ *such that:*

- $z = x_1 a^i x_2$,
- $u = x_1 a^j x_2$,
- $v = x_1 a^k x_2$ *and*
- $\alpha\binom{n}{i} \leq \beta_1\binom{n}{j} + \beta_2\binom{n}{k}$ *holds for every* $n \geq 0$.

Now for SI's having arbitrary coefficients we can state our main result, which follows from Lemma 3 and Lemma 4.

Theorem 2. *A SI of the form* $\alpha_1 u_1 + .. + \alpha_n u_n \leq \beta_1 v_1 + .. + \beta_{n+1} v_{n+1}$ *holds if both* $\alpha_i u_i \leq \beta_i v_i + \beta_{i+1} v_{i+1}$ *and* $\beta_{i+1} v_{i+1} \leq \alpha_i u_i + \alpha_{i+1} u_{i+1}$ *hold for every* $i \leq n$ *and* $i \leq n-1$, *respectively.*

References

[1] C. Ding and A. Salomaa: *On some problems of Mateescu concerning subword occurrences*, Fundamenta Informaticae 73 (2006), 65-79

[2] S. Fossé and G. Richomme: *Some characterizations of Parikh matrix equivalent binary words*, Information Processing Letters 92 (2004), 77-82

[3] M. Lothaire (ed.): *Combinatorics on Words*, Addison Wesley, Reading, MA, 1983.

[4] A. Mateescu, A. Salomaa: *Matrix indicators for subword occurrences and ambiguity*, International Journal of Foundations of Computer Science 15 (2004), 277-292

[5] A. Mateescu, A. Salomaa and S. Yu: *Subword histories and Parikh matrices*, Journal of Computer and System Sciences 68 (2004), 1-21

[6] A. Salomaa: *Connections between subwords and certain matrix mappings*, Theoretical Computer Science 340 (2005), 188-203

[7] A. Salomaa: *Counting (scattered) subwords*, EATCS Bulletin 81 (2003), 165-179

[8] T.-F. Şerbănuţă: *Extending Parikh matrices*, Theoretical Computer Science 310 (2004), 233-246