A note on the decidability of subword inequalities*

Szilárd Zsolt Fazekas[1] Robert Mercaş[2]

August 17, 2012

Abstract

Extending the general undecidability result concerning the absoluteness of inequalities between subword histories, in this paper we show that the question whether such inequalities hold for all words is undecidable already over a binary alphabet and bounded number of blocks, and even in very simple cases an answer requires an intractable computation.

1 Introduction

The notion of Parikh matrices was introduced in [3] as a generalization of Parikh vectors in order to provide a more informative description of words. Beside the classical multiplicity of each letter given by the Parikh vectors, the matrices also provide information regarding the order in which some of these letters appear in the words. A scattered subword consists of a concatenation of some of the letters of a word, in the order they appear in it. Over the course of developing the theory of Parikh matrices, in [4] the authors introduced the notion of subword history, which was used in investigating relations between different scattered subwords of a word. In the same paper, the authors settled the decidability of equalities between subword histories with a positive answer and called for a continuation with respect to inequalities between subword histories. Certain easily testable cases when inequalities hold and a characterization of valid small inequalities were given in [1]. The main question, "is it decidable whether the value of a given subword history is non-negative in all words over a given alphabet?" was proved undecidable in [6]. Seki proved that for a nine letter alphabet the problem can be reduced to that of the solvability of Diophantine equations and, thus, proved undecidable according to [5]. In this paper we show that using once more reductions to Diophantine equations the problem is undecidable even for a binary alphabet and even if the number of blocks is bounded.

Considering the reader to be familiar with general Combinatorics on Words concepts we end this section with a few definitions. For more details, see [2].

*The work of Szilárd Zsolt Fazekas was supported by the *Japanese Society for the Promotion of Science* under number *P10827* and the work of Robert Mercaş was supported by the *Alexander von Humboldt Foundation*

[1]College of Nyíregyháza, Mathematics and Informatics Institute, Sóstói út 31/b, Nyíregyháza 4400, Hungary, szilard.fazekas@gmail.com

[2]Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, PSF 4120,D-39016 Magdeburg, Germany, robertmercas@gmail.com

A factor (continuous subword) z of a word w is called a *block* of w if $z = a^k$, for some $a \in \Sigma, k > 0$, and there are no words u, v such that $w = uazv$ or $w = uzav$. In other words blocks are maximal factors that are powers of one letter. If $w = a_1^{k_1} a_2^{k_2} ... a_n^{k_n}$ with $a_i \neq a_{i+1}$ for $1 \leq i \leq n-1$, we will call $red(w) = a_1 a_2 ... a_n$ the reduced form of w, and $pow(w) = (k_1, k_2, .., k_n)$ the power vector of w.

A word $u = a_1 a_2 ... a_m$ is a scattered subword of $w = b_1 b_2 ... b_n$ 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. For the phrase inequality between subword histories or, in other words, subword inequality we will use the shorthand SI in the paper.

Consider an alphabet Σ and a word $w \in \Sigma^*$. A subword history in Σ and its value in w are defined recursively as follows.

- Every $u \in \Sigma^*$ is a subword history in Σ, referred to as monomial, and its value in w equals $|w|_u$.
- Assume that SH_1 and SH_2 are subword histories with values α_1 and α_2, respectively. Then

$$-(SH_1),\ (SH_1) + (SH_2) \text{ and } (SH_1) \times (SH_2)$$

  are subword histories with values

$$-\alpha_1,\ \alpha_1 + \alpha_2 \text{ and } \alpha_1 \times \alpha_2,$$

  respectively.

In line with previous notations, the value assumed by a subword history SH in a word w will be denoted by $|w|_{SH}$. Two subword histories are termed equivalent if they assume the same value in any w. A subword history is linear if it is obtained without using the operation $\times$. As we will investigate inequalities between subword histories, the following result from [4] will be useful.

Theorem 1. *Every subword history is equivalent to a linear subword history. Moreover, given a subword history, an equivalent linear subword history can be effectively constructed.*

We write $SH_1 \leq SH_2$ if, for all words w, the value of SH_1 in w is at most that of SH_2 in w. Finally, we recall a result by Seki, which serves as the starting point of our investigation. It shows that solving subword inequalities is reducible to solving subword equalities.

Lemma 1. *[6] The problem of "deciding for two subword histories SH_1 and SH_2 whether there exists a word $w \in \Sigma^*$ such that $|w|_{SH_1} = |w|_{SH_2}$ holds" is polynomial-time Karp reducible to "for a given subword history SH, is it decidable whether $|w|_{SH} \geq 0$ holds for every word w in Σ^*".*

2 Undecidability for binary alphabets

Theorem 2. *For a Diophantine equation of the form:*

$$\sum_{i=1}^{n} c_i x_1^{e_{i,1}} x_2^{e_{i,2}} \cdots x_m^{e_{i,m}} = 0$$

it is possible to construct a subword equation over a binary alphabet, which has a solution if and only if the Diophantine equation does.

Proof. Let us construct a system of subword inequalities having terms

$$u_j = a_1^{n_1} a_2^{n_2} \dots a_{j-1}^{n_{j-1}} a_j a_{j+1}^{n_{j+1}} \dots a_m^{n_m}$$

where $a_i \in \{a, b\}$, $\forall i \in \{1, .., n\}$, and let us define the "solution word" w as

$$w = a_1^{n_1} a_2^{n_2} \dots a_m^{n_m}.$$

Clearly, $|w|_{u_j} = n_j$, as the words have the same number of blocks and all blocks of u_j but the jth one are equal to their counterparts in w.

Let $u_j^{\times n}$ denote the non-linear subword history we get when multiplying the monomial u_j with itself n times. For example, $u_j^{\times 2} = u_j \times u_j$. Now, according to the definition of subword histories, if the value of u_j in w is n_j, then the value of $u_j^{\times e_{i,j}}$ in w is $n_j^{e_{i,j}}$, therefore the value assumed by

$$c_i \left(u_1^{\times e_{i,1}} \times u_2^{\times e_{i,2}} \times \dots \times u_m^{\times e_{i,m}} \right)$$

in the word w is exactly $c_i n_1^{e_{i,1}} n_2^{e_{i,2}} \cdots n_m^{e_{i,m}}$.

We have seen that for a Diophantine equation we can construct a subword equation, which will have a solution if the Diophantine equation has. Conversely, if the subword equation has a solution, i.e., there exists a word w such that

$$\sum_{i=1}^{n} c_i |w|_{u_1} \cdots |w|_{u_n} = 0$$

then, by assigning the values $x_i = |w|_{u_i}$, we get a solution for the Diophantine equation. □

Corollary 1. *Given a subword inequality over a binary alphabet, it is undecidable whether it holds for all words or not.*

From Theorem 1 we know that we can construct linear subword histories equivalent to the sides of the SI, hence in the remainder of the paper we will deal only with inequalities where both sides consist of linear subword histories. This will be especially significant when discussing 2-block inequalities.
By modifying the reduction from [6, Theorem 1] of solvability of Diophantine equations to subword equations we showed that the problem in general is undecidable even for binary alphabets. However, note that the linearized version of the subword equations used in the reduction to simulate their Diophantine counterpart are complex. The question offers itself, are there some restricted forms of inequalities which are decidable? In what follows, when we talk about subword inequalities we use 'term' and 'monomial' interchangeably. Some simple cases were shown to be solvable (see [1]), namely equations where the left side consists of one term and the right side of two.

There are various kinds of restrictions one can consider:

1. fix the maximum number of blocks a term can have,

2. prescribe some relation between the terms, e.g. that they all have the same reduced form, or that they have the same length, etc.,

3. require that the number of terms on one side or as a whole is bounded by a constant (see [1]).

Another path to follow in investigating these inequalities is to consider some proper subset of Σ^* as the 'solution space', that is, for a given SI check whether it holds for all words in some language $\mathcal{L} \subset \Sigma^*$.

Here we initiate the discussion over the first case. By a k-block subword history we mean a subword history (SH) in which every term has at most k blocks. For example, in a 2-block SH over a binary alphabet the reduced form of every term is one of the words a, b, ab, ba. Please recall that we are talking about linear subword equations. Hence, to have a precise definition, we interpret the restriction as follows: a (general, possibly non-linear) SI of the form $SH_1 \leq SH_2$ is a k-block SI, if there exist k-block linear subword histories SH_1' and SH_2' such that SH_1 is equivalent to SH_1' and SH_2 is equivalent to SH_2'. Losing the linearity requirement would give us a problem which is already settled [6, Theorems 2 and 3].

We answer the following question: "for a given k, is it decidable whether a k-block SI holds for all words over the alphabet". First we look at the simple case of $k = 1$, which is reducible to solving univariate polynomial equations over the natural numbers. Then for $k \geq 2$ we show that the reduction from Diophantine equations is possible even if we consider terms which have no more blocks than the number of variables in the equations. From that, one can derive the undecidability of inequalities with terms having at least 9 blocks and get that even with at most two blocks per term the question is NP-hard. Note that in this modified version of the question we did not mention the size of the alphabet. This is because the case $k = 1$ is equivalent to studying unary SIs, whereas in the other cases a binary alphabet will suffice for our reductions and having a larger alphabet does not offer any improvements.

Over a unary alphabet, the number of occurrences of u in w is simply $|w|_u = \binom{|w|}{|u|}$. This allows us to write any unary SI in the form:

$$\sum_{i=1}^{m} c_i a^i \geq 0,$$

where the coefficients c_i are non-negative integers and m is the length of the longest term in the SI. Since for solving a one variable Diophantine equation it is enough to check if the solution is in the set of divisors of the free term, or when this one does not exist, of the coefficient of the second smallest degree, the following result follows:

Proposition 1. *Given a 1-block subword inequality, it is decidable in linear time whether it holds for all words. Moreover, if the answer is negative, a counterexample can be found within the same time frame.*

A first observation concerning the earlier reduction of Diophantine equations to SIs is that the solution word w which we construct has as many blocks as the

number of variables. It immediately follows that for all words u having more blocks than w we have $|w|_u = 0$. Therefore, these long us can be discarded if we can show that should the subword equation have a solution, it will always have a solution word which has as many blocks as the number of variables in the Diophantine equation.

Lemma 2. *For a Diophantine equation of the form:*

$$\sum_{i=1}^{n} c_i x_1^{e_{i,1}} x_2^{e_{i,2}} \cdots x_m^{e_{i,m}} = 0 \tag{1}$$

it is possible to construct an m-block linear subword equation $\mathcal{Q}_m = 0$ over a binary alphabet, which has solutions if and only if the Diophantine equation does. Moreover, if (1) has a solution, there exists a word w with $|red(w)| = m$ such that $|w|_{\mathcal{Q}_m} = 0$.

Proof. As in Theorem 2 before, let us construct a system of subword inequalities having terms

$$u_j = a_1^{n_1} a_2^{n_2} \dots a_{j-1}^{n_{j-1}} a_j a_{j+1}^{n_{j+1}} \dots a_m^{n_m} \in \{a, b\}^*$$

where $a_i \neq a_{i+1}$, $\forall i \in \overline{1, .., n-1}$, and let us define the "solution word" w as

$$w = a_1^{n_1} a_2^{n_2} \dots a_m^{n_m}.$$

Now we can provide a subword equation which has only terms with at most m blocks, but it is not linear:

$$\mathcal{Q}' = \sum_{i=1}^{n} c_i \left(u_1^{\times e_{i,1}} \times u_2^{\times e_{i,2}} \times \ldots \times u_m^{\times e_{i,m}} \right)$$

Let $\mathcal{Q}$ be the linear subword history equivalent to $\mathcal{Q}'$. To get the linear m-block subword equation let us take all terms from $\mathcal{Q}$, that have at most m blocks, and denote the sum of them by $\mathcal{Q}_m$. Furthermore, let the set of terms appearing in it be $\{\mathcal{Q}_m\}$. Now let us get to the proof of the first part of the lemma. (IF) We know that $|w|_u = 0$ for all $u \in \Sigma^*$ such that $red(u) > red(w)$. Clearly, $\mathcal{Q}_m$ assumes the same value in w as $\mathcal{Q}$, because for all words $u \in \{\mathcal{Q}\} \setminus \{\mathcal{Q}_m\}$ we have $|w|_u = 0$ as $|red(u)| > m = |red(w)|$, hence $|w|_{\mathcal{Q}'} = |w|_{\mathcal{Q}} = |w|_{\mathcal{Q}_m}$. This, in turn, means that if the Diophantine equation has a solution, the value assumed by $\mathcal{Q}_m$ in w is 0.
(ONLY IF) To prove this direction, first we have to show that if there exists any word w' such that $|w'|_{\mathcal{Q}} = 0$, there exists a word w with $red(w) \leq m$ for which $|w|_{\mathcal{Q}} = 0$. If $red(w') \leq m$ we are done. In this case $\mathcal{Q}_m = 0$ is exactly the linear m-block subword equation we were looking for. Now suppose $red(w') > m$. From the assumption $|w'|_{\mathcal{Q}'} = |w'|_{\mathcal{Q}} = 0$ we get that by the assignments $x_j = |w'|_{u_j}$ our Diophantine equation has a solution. Similarly to our initial construction of w let us define it as follows

$$w = a_1^{|w'|_{u_1}} a_2^{|w'|_{u_2}} \dots a_m^{|w'|_{u_m}},$$

and consider the words $v_j = a_1^{|w'|_{u_1}} a_2^{|w'|_{u_2}} \dots a_{j-1}^{|w'|_{u_{j-1}}} a_j a_{j+1}^{|w'|_{u_{j+1}}} \dots a_m^{|w'|_{u_m}}$. Reiterating the previous argument

$$\mathcal{R}' = \sum_{i=1}^{n} c_i \left(v_1^{\times e_{i,1}} \times v_2^{\times e_{i,2}} \times \ldots \times v_m^{\times e_{i,m}} \right) = 0$$

will have a solution if and only if (1) has one. Define $\mathcal{R}$ as the linear equivalent of $\mathcal{R}'$ and $\mathcal{R}_m$ as before. Then,

$$\mathcal{R}_m = \{u \in \{\mathcal{R}\} \mid |red(u)| \leq m\},$$

and $\mathcal{R}_m = 0$ is the subword equation with the required properties.

The last statement of the theorem follows instantly because the solution word w was constructed so that it has exactly m blocks. □

The next theorem follows directly from Lemma 2.

Theorem 3. *For $k \geq 9$, it is undecidable whether a given k-block subword inequality holds for all words.*

Note that this result, although not groundbreaking, brings a bit more insight on the undecidability question of subword inequalities. In particular, we show that the question is undecidable already for SIs with more than 8 blocks.

References

[1] Szilárd Zsolt Fazekas. On inequalities between subword histories. *International Journal of Foundations of Computer Science*, 19(4):1039–1047, 2008.

[2] M. Lothaire. *Combinatorics on Words*. Cambridge University Press, 1997.

[3] Alexandru Mateescu, Arto Salomaa, Kai Salomaa, and Sheng Yu. A sharpening of the parikh mapping. *ITA*, 35(6):551–564, 2001.

[4] Alexandru Mateescu, Arto Salomaa, and Sheng Yu. Subword histories and Parikh matrices. *Journal of Computer and System Sciences*, 68(1):1–21, 2004.

[5] Yuri Matiyasevich. *Hilbert's Tenth Problem*. MIT Press, 1993.

[6] Shinnosuke Seki. Absoluteness of subword inequality is undecidable. *Theoretical Computer Science*, 418:116–120, 2012.