Absoluteness of subword inequality is undecidable

Shinnosuke Seki

August 9, 2011

Abstract

Mateescu, Salomaa, and Yu asked: is it decidable whether a given subword history assumes only non-negative values for all words over a given alphabet. In this paper, we solve this open problem by proving that this problem is undecidable even under stronger conditions than supposed originally.

1 Subword history and inequality

Let Σ be an alphabet, and by Σ^* we denote the set of all words over Σ including the empty word λ.

Parikh mappings (vectors), introduced in [10], provide us with numerical properties of a word and a set of words. Some specific ordering of the letters in $\Sigma = \{a_1, \ldots, a_n\}$ in mind, the Parikh mapping of a word w is $(|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_n})$, where $|w|_a$ denotes the number of occurrences of a letter $a \in \Sigma$ in a word $w \in \Sigma^*$ (for instance, $|aab|_a = 2$ and $|aab|_b = 1$). This idea can be generalized as counting in w the number of occurrences of another word u as a (continuous) subword or a scattered subword. The latter is of especial interest. In general, u is a *scattered subword* of w if there exist an integer $k \geq 1$ and words $x_1, \ldots, x_k, y_0, y_1, \ldots, y_k$, some of which are possibly empty, such that

$$u = x_1 \cdots x_k \text{ and } w = y_0 x_1 y_1 \cdots x_k y_k.$$

For various usages of terminologies, the reader is referred to [12]. Then we can generalize the notation $|w|_a$ as $|w|_u$ to denote the number of occurrences of u as a scattered subword of w. For instance, $|aab|_{ab} = 2$ because two occurrences of a precede that of b. It is a convention made in [7] to assume that $|w|_\lambda = 1$ for the empty word λ and any word $w \in \Sigma^*$.

The number of scattered subwords can provide more information about the word w itself than Parikh mapping. For $\Sigma = \{a, b\}$, the Parikh mapping $(3, 3)$ admits all 20 words in $aaa \sqcup\!\sqcup bbb$ like $ababba$ as w, where $\sqcup\!\sqcup$ is the shuffle operation. Adding a condition $|w|_{ab} = 8$ to this Parikh mapping reduces the candidate of w to $aababb$ [7]. More advanced logic can be implemented by adding and/or multiplying such conditions; $|w|_a \times |w|_b = 4$ implies that $w \in a \sqcup\!\sqcup bbbb \cup aa \sqcup\!\sqcup bb \cup aaaa \sqcup\!\sqcup b$. This idea led Mateescu, Salomaa, and Yu to propose the notion of subword history as follows.

Definition 1 ([7])**.** A *s*ubword history in Σ and its *value* in a word w are defined recursively as follows:

- Every word u in Σ^* 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 respective values

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

The notation $|w|_u$ is now further generalized as $|w|_{SH}$ for a subword history SH to denote the value of SH in w.

For a non-negative integer $e \geq 0$ and a subword history SH, we denote $\overbrace{SH \times SH \times \cdots \times SH}^{e \text{ times}}$ by $\prod^e SH$. For instance, $\prod^2 SH = SH \times SH$ and $\prod^3 SH = SH \times SH \times SH$. Let us set $\prod^0 SH$ be λ for any subword history SH. In light of the next proposition, this setting does not contradict the convention that $|w|_\lambda = 1$ for any word w.

Proposition 1. *Let SH be a subword history in Σ with value α, c be an integer, and e be a non-negative integer. Then $c(SH)$ and $\prod^e SH$ are subword histories with respective values $c\alpha$ and α^e.*

Two subwords SH_1 and SH_2 are *equivalent* if $|w|_{SH_1} = |w|_{SH_2}$ for every word $w \in \Sigma^*$. It is not difficult to observe that the subword histories $a \times b$ and $ab+ba$ assume the same value in any word (see [7]). These two subword histories are hence equivalent. A subword history is *linear* if it is obtained without using the operation ×. We say that a linear subword history is *of degree n* if its longest monomial is of length n. For instance, the degree of $abb + 2c + 3$ is 3 due to its first term. More generally, we can define the degree of a subword history as the minimum degree of equivalent linear subword histories.

Mateescu, Salomaa, and Yu proposed a method of constructing from a given subword history an equivalent *linear* subword history, and as its corollary, the problem of deciding the equivalence of two given subword histories turned out to be decidable [7]. In the paper, the authors called for a continuation of research on inequalities between subword histories. Specifically, they left the following problem open: for a given subword history SH, is it decidable whether $|w|_{SH} \geq 0$ holds for every word w in Σ^*. Let us call this problem `SubwordIneqAbsoluteness`. From the point of view of decidability, it is irrelevant whether this problem is formalized with $\geq$ or with $>$. Indeed, deciding whether $|w|_{SH} > 0$ holds for every word $w \in \Sigma^*$ is equivalent to deciding whether $|w|_{SH-\lambda} \geq 0$; note that $SH - \lambda$ is a valid subword history with value $|w|_{SH} - 1$.

2 Main results

In this section, we prove that `SubwordIneqAbsoluteness` is undecidable even under strong restrictions (Corollary 2). This is our main contribution in this paper.

First of all, we show that this problem is at least as hard as the problem of deciding for given 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. Let us call the latter problem `SubwordEqSolvability`. The reader can consult [11, 2], if needs arise, on undecidability, polynomial-time Karp reduction, and NP-hardness.

Lemma 1. `SubwordEqSolvability` *is polynomial-time Karp reducible to* `SubwordIneqAbsoluteness`.

Proof. Assume that as an instance of `SubwordEqSolvability` two subword histories SH_1 and SH_2 are given. Let $SH = SH_1 - SH_2$. Then, the answer to this instance is no if and only if $|w|_{SH \times SH} > 0$ for every word $w \in \Sigma^*$. Note that $SH \times SH$ is a valid subword history (Proposition 1), and its value in w is $(|w|_{SH})^2$. □

In order to prove the undecidability of `SubwordIneqAbsoluteness`, therefore, it suffices to prove that `SubwordEqSolvability` is undecidable.

Theorem 1. `SubwordEqSolvability` *is undecidable.*

Proof. This proof is based on the unsolvability of Diophantine equation proved by Matiyasevich in [8], the answer to the Hilbert's tenth problem. Let a Diophantine equation

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

be given, where $\ell \ge 1$, $c_1, \ldots, c_\ell$ are integer constants, $x_1, x_2, \ldots, x_m$ are positive integer variables, and $e_{i,1}, e_{i,2}, \ldots, e_{i,m}$ are non-negative integer exponents for $1 \le i \le \ell$. (It is well known that we can restrict the attention to positive integer variables, see [11].)

Let $\Sigma = \{a_1, \ldots, a_m\}$. Consider a word w in $a_1^{n_1} \sqcup\!\sqcup a_2^{n_2} \sqcup\!\sqcup \cdots \sqcup\!\sqcup a_m^{n_m}$ for some non-negative integers $n_1, \ldots, n_m$. Then for $1 \le j \le m$, we have

$$|w|_{a_j} = n_j.$$

Proposition 1 implies that $\prod^{e_{i,j}} a_j$ is a subword history for any $1 \le i \le \ell$, and its value in w is $n_j^{e_{i,j}}$. Using the proposition once again, we see that

$$c_i \left(\prod^{e_{i,1}} a_1 \times \prod^{e_{i,2}} a_2 \times \cdots \times \prod^{e_{i,m}} a_m \right)$$

is a subword history whose value in w is $c_i n_1^{e_{i,1}} n_2^{e_{i,2}} \cdots n_m^{e_{i,m}}$. Let us denote this subword history by SH_i, and let $SH = \sum_{1 \le i \le \ell} SH_i$, which is also a subword history. Now it should be clear that

$$|w|_{SH} = \sum_{1 \le i \le \ell} c_i n_1^{e_{i,1}} n_2^{e_{i,2}} \cdots n_m^{e_{i,m}}. \qquad (1)$$

This is the value we can obtain by substituting $(n_1, \ldots, n_m)$ into the given Diophantine equation. Therefore, if the Diophantine equation has a positive integer solution $(n_1, n_2, \ldots, n_m)$, then for such w, $|w|_{SH} = 0$. Conversely, assume that there exists a word $v \in \Sigma^*$ such that $|v|_{SH} = 0$. According to Definition 1 and Eq. (1),

$$|v|_{SH} = \sum_{1 \le i \le \ell} c_i |v|_{a_1}^{e_{i,1}} |v|_{a_2}^{e_{i,2}} \cdots |v|_{a_m}^{e_{i,m}}.$$

Since this value is 0, $(|v|_{a_1}, |v|_{a_2}, \ldots, |v|_{a_m})$ is a positive integer solution to the given Diophantine equation. Consequently, if `SubwordEqSolvability` were decidable, then we would be able to determine the solvability of the given Diophantine equation, a contradiction. □

Corollary 1. `SubwordIneqAbsoluteness` *is undecidable.*

As being mentioned previously, the equivalence of two subword histories is decidable, and note that this does not contradict Corollary 1.

It is worth observing that in the proof of Theorem 1, we reduce a given Diophantine equation into an element of a restricted class of subword histories, which we call the class of *letter-restricted subword histories*. The definition of letter-restricted subword history is obtained by restricting that monomials be letters in Σ or λ in Definition 1.

It is well known that the Diophantine equations remain unsolvable even if the number of variables involved is limited to be 9 [9]. In the proof of Theorem 1, the number of variables equals that of letters in Σ. Thus, over an alphabet of 9 letters, `SubwordEqSolvability` is undecidable, and hence, so is `SubwordIneqAbsoluteness`. Combining this with what was mentioned in the last paragraph, now we present our strongest result on the undecidability of these problems as of this moment.

Theorem 2. *If the Diophantine equations are unsolvable over n variables, then* `SubwordEqSolvability` *and* `SubwordIneqAbsoluteness` *are unsolvable even for the class of letter-restricted subword histories over an alphabet of n letters.*

Corollary 2. `SubwordEqSolvability` *and* `SubwordIneqAbsoluteness` *are undecidable even for the class of letter-restricted subword histories over a nonary alphabet.*

Corollary 2 does not mean that `SubwordEqSolvability` or `SubwordIneqAbsoluteness` is decidable over an alphabet of size at most 8. It is conjectured that Diophantine equations remain unsolvable even over three variables. If so, then Theorem 2 implies that these problems would be undecidable even for the class of letter-restricted subword histories over a ternary alphabet.

How small do we have to make the size of alphabet to make these problems decidable? We cannot help but leave this matter unsettled in this paper, but can provide a result to illustrate how hard `SubwordIneqAbsoluteness` is. Manders proved that it is NP-complete to decide the solvability of a given Diophantine equation of the form $c_1x^2 + c_2y + c_3 = 0$ [6]. Our construction of a subword history from a given Diophantine equation in the proof of Theorem 1 can be done in a polynomial time. In addition, the subword history thus constructed can be transformed in a polynomial time into a linear subword history $c_1 a + 2c_1 aa + c_2 b + c_3$ of degree 2 by the above-mentioned product elimination by Mateescu, Salomaa, and Yu, where the letters a and b correspond to the variables x and y, respectively. With Lemma 1, we can prove the following theorem, though it does not settle the question at the beginning of this paragraph.

Theorem 3. `SubwordEqSolvability` *and* `SubwordIneqAbsoluteness` *are NP-hard even for the class of letter-restricted subword histories of degree 2 over a binary alphabet.*

3 System of Diophantine equations

In this section, we glance at the polynomial-time Karp reduction from a given system of Diophantine equations to a subword inequality. The reduction should be in itself trivial from our proof of Theorem 1, but let us spend some space for this because of an implication it has on a significant problem called *preimage problem.*

A system of Diophantine equations is a finite collection $(Eq_1, Eq_2, \dots, Eq_k)$ of Diophantine equations. Using our method, the equations $Eq_1, \dots, Eq_k$ are transformed into the respective subword histories $SH_1, \dots, SH_k$. From them, we construct the following subword history:

$$SH := \prod_{1 \leq i \leq k} ((SH_i \times SH_i) + 1).$$

Then, for $w \in \Sigma^*$, $|w|_{SH} = 1$ if and only if for all $1 \leq i \leq k$, $|w|_{SH_i} = 0$. Since SH always assumes a positive integer value, deciding whether $|w|_{SH} = 1$ can be done both by equation and by inequality.

Given a subword history SH and a word $w \in \Sigma^*$, it is a pen-and-paper calculation to obtain the value of SH in w, and it remains the case no matter how many subword histories are given. From the subword histories $SH_1, \dots, SH_k$ and the values $n_1, \dots, n_k$ thus calculated from w, we can build the following system of subword equations:

$$\begin{cases} SH_1 & = & n_1 \\ & \vdots & \\ SH_k & = & n_k \end{cases}$$

and after that, we hide w. Can we find w, or more desirably, can we eliminate the candidates of w? In Section 1, an example was cited from [7] to see the uniqueness of the word $w \in \{a, b\}^*$ satisfying $|w|_a = |w|_b = 3$ and $|w|_{ab} = 8$. This is interpreted in the above-mentioned framework as finding w when $(3, 3, 8)$ is given (assume that we know to what subword history each coordinate is related in this vector). Problems of this type are collectively termed *preimage* *problems* (see, e.g., [1, 5] for a preimage problem in chemoinformatics). Preimage problems can be formalized not only for words but for various objects like graphs so long as some of their properties can be quantified. What we mentioned in the previous paragraph, however, demonstrates how computationally-hard the preimage problem is even for words. One reason for this hardness is that in counting occurrences of a subword, search range covers the whole of a given word (global scope) in our current formalization. As such, if we reformulated problems so as to confine the search range, then the reformulated preimage problem could be solved even efficiently. In [1], Akutsu and Fukagawa counted only the occurrences of words as a *continuous* subword, and showed that in this setting, the preimage problem can be solved in a polynomial time.

4 Concluding remarks, discussions, and future directions

In this paper, it was proved to be undecidable whether there exists a word in which an equation between given two subword histories holds. With the polynomial-time Karp reduction, this amounted to the answer to the open problem by Mateescu, Salomaa, and Yu posed in [7]. This problem was proved to remain undecidable even under conditions on the size of alphabet, on the class of subword histories considered, and on the length of monomials involved. As such, our main results are stronger than a solution to the original open problem.

Results in this paper are oriented toward unsolvability, and therefore, cannot be employed to make use of a number of known decidability results on the solvability of Diophantine equations (see [3]). This motivates us to make a research on the characterization of subword histories that is polynomial-time Karp reducible to a Diophantine equation whose solvability is decidable. It might be worth recalling that the Diophantine equations are reduced to the *very restricted* class $\mathcal{SH}$ of letter-restricted subword histories. Thus, for any class of subword histories that does not contain $\mathcal{SH}$ as a subset, it remains unknown whether `SubwordEqSolvability` or `SubwordIneqAbsoluteness` is decidable. The most significant difference between Diophantine equations and equations on subword histories is that the latter is defined on the group which is not Abelian. In this paper, this difference has been barely encountered because our attention was mainly on the class of letter-restricted subword histories, in which commutativity does not count so much. This observation gives us an impression that combinatorics on words will play an important role in working on the abovementioned problems (see [4] and the references therein).

Acknowledgements

We wish to express our sincere gratitude to Professor Dr. Arto Salomaa for reading earlier versions of this paper carefully and making quite a few valuable comments. He originated the idea of letter-restricted subword history. We are indebted to Dr. Szilard Zsolt Fazekas for his introducing the open problem to us and for helpful discussions with him. This paper also benefited from my discussions with Mr. Kei Taneishi on Diophantine equations, number theory, and chemoinformatics.

This research is financially supported by the Funding Program for Next Generation World-Leading Researchers (NEXT program) to Professor Dr. Yasushi Okuno.

References

[1] T. Akutsu and D. Fukagawa. Inferring a graph from path frequency. In A. Aposolico, M. Crochemore, and K. Park, editors, *CPM 2005*, volume 3537 of *LNCS*, pages 371–382. Springer, 2005.

[2] S. Arora and B. Barak. *Computational Complexity A Modern Approach.* Cambridge University Press, 2009.

[3] A. Baker. *A Concise Introduction to the Theory of Numbers.* Cambridge University Press, 1984.

[4] S. Z. Fazekas. On inequalities between subword histories. *Int. J. Found. Comput. S.*, 19(4):1039–1047, 2008.

[5] H. Fujiwara, J. Wang, L. Zhao, H. Nagamochi, and T. Akutsu. Enumerating treelike chemical graphs with given path frequency. *J. Chem. Inf. Model.*, 48:1345–1357, 2008.

[6] K. L. Manders. Computational complexity of decision problems in elementary number theory. In *Model Theory of Algebra and Arithmetic*, volume 834 of *LNM*, pages 211–227. Springer, 1980.

[7] A. Mateescu, A. Salomaa, and S. Yu. Subword histories and Parikh matrices. *J. Comput. Syst. Sci.*, 68:1–21, 2004.

[8] Y. Matiyasevich. Solution of the tenth problem of Hilbert. *Matematikai Lapok*, 21:83–87, 1970.

[9] Y. Matiyasevich. *Hilbert's Tenth Problem.* MIT Press, 1993.

[10] R. J. Parikh. On context-free languages. *Journal of the Association for Computing Machinery*, 13:570–581, 1966.

[11] G. Rozenberg and A. Salomaa. *Cornerstones of undecidability.* Prentice Hall, 1994.

[12] G. Rozenberg and A. Salomaa, editors. *Handbook of Formal Languages*, volume 1-3. Springer, 1997.