UPPER TRIANGULAR MATRICES AND SUBWORD OCCURRENCES

Arto Salomaa

Received April 7, 2009

Abstract. Many recent investigations have dealt with the quantity $|w|_u$, the number of occurrences of a word u as a scattered subword of a word w. The quantity gives important numerical information about the word w. Sufficiently many values $|w|_u$, for different u's, characterize the word w completely. Certain upper triangular matrices, often referred to as *Parikh matrices* have turned out to be very useful for computing numbers $|w|_u$. In this paper we discuss some properties of Parikh matrices, as well as some criteria concerning matrix equivalence of words. Special emphasis is on the so-called *Cauchy inequality* for words.

1 Matrices characterizing words It is desirable to express properties of words as numbers. The main goal of such an *arithmetization* is to reach a situation where noncommutativity is eliminated. The theory of formal power series, [6], contains numerous such constructions. In what follows we assume that the reader is familiar with the basics of formal languages. Whenever necessary, [11] may be consulted. As customary, we use small letters from the beginning of the English alphabet a, b, c, d, possibly with indices, to denote letters of our formal alphabet Σ. Words are usually denoted by small letters from the end of the English alphabet.

The most direct numerical fact about a word w is its *length* $|w|$. The *Parikh vector*, [10, 11], $\Psi(w) = (i_1, \ldots, i_n)$ indicates the number of occurrences of the letter a_j, $1 \leq j \leq n$, in w, provided w is over the alphabet $\Sigma = \{a_1, \ldots, a_n\}$. To get more information about a word, one has to focus the attention to subwords and to the number of occurrences of a specific subword in the given word. In this article, u being a *subword* of w means that w, as a sequence of letters, contains u as a subsequence. More formally, there exist words $x_1, \ldots, x_k$ and $y_0, \ldots, y_k$, some of them possibly empty, such that

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

We also consider *factors* u of a word w: u is a factor of w if there are words x and y such that $w = xuy$. Throughout this article, we understand subwords and factors in the way mentioned. (In classical language theory, [11], our subwords are usually called "scattered subwords", whereas our factors are called "subwords".)

The notation used throughout the article is $|w|_u$, the number of occurrences of the word u as a subword of the word w. This number can be defined formally as follows. Occurrences can be viewed as vectors. If $|u| = t$, each occurrence of u in w can be identified as the t-tuple $(i_1, \ldots, i_t)$ of increasing positive integers, where for $1 \leq j \leq t$, the jth letter of u is the i_jth letter of w. For instance, the 5 occurrences of $u = abc$ in $w = abcbcacab$ are

$$(1,2,3),\ (1,2,5),\ (1,2,7),\ (1,4,5),\ (1,4,7).$$

2000 *Mathematics Subject Classification.* 68R15, 68Q45.

Key words and phrases. subword, scattered subword, Parikh matrix, matrix equivalence, Cauchy inequality.

(We will return to this example below.) Clearly, $|w|_u = 0$ if $|w| < |u|$. We also make the *convention* that, for any w and the empty word λ, $|w|_\lambda = 1$.

In [4] the number $|w|_u$ is denoted as a "binomial coefficient" $|w|_u = \binom{w}{u}$. If w and u are words over a one-letter alphabet, $w = a^i,\ u = a^j$, then $|w|_u$ equals the ordinary binomial coefficient: $|w|_u = \binom{i}{j}$. Our convention concerning the empty word reduces to the fact that $\binom{i}{0} = 1$.

Assume that Σ is an alphabet containing the letters a and b. Then, for any word w,

$$(|w|_a) \times (|w|_b) = |w|_{ab} + |w|_{ba}.$$

This simple equation is one of the few general facts about occurrences of subwords. A slight variation immediately leads to difficulties. No explicit characterization is known for the relation between $(|w|_u, |w|_v)$ and $(|w|_{uv}, |w|_{vu})$, where u, v, w are arbitrary words.

We are now ready to describe a method using matrices for the computation of numbers $|w|_u$. Our matrix mappings use upper triangular square matrices, with nonnegative integer entries, 1's on the main diagonal and 0's below it. (A somewhat different construction was introduced in [2].) The set of all such triangular matrices is denoted by $\mathcal{M}$, and the subset of all matrices of dimension $k \geq 1$ is denoted by $\mathcal{M}_k$. Every such matrix has an inverse but the inverse may contain negative entries. Given an ordered alphabet $\Sigma = \{a_1, \ldots, a_n\},\ n \geq 2$, we now consider a subset of $\mathcal{M}_{n+1}$. Matrices in this subset are referred to as *Parikh matrices*, in analogy to the one-dimensional case. The original $(n+1)$-dimensional Parikh matrix, [7], tells us the values $|w|_u$, where u is a factor of the ordered product $a_1 \ldots a_n$ of the letters of the alphabet. When considering *generalized Parikh matrices* introduced first in [18], the values $|w|_x$ can be obtained as entries, where x belongs to any previously chosen finite language. The price one pays is in the dimension of the matrix. Denote the entries of the matrices M by m_{ij}. Before the basic definition of a generalized Parikh matrix we still recall the definition of the "Kronecker delta". For letters c and d,

$$\delta_{c,d} = \begin{cases} 1 & \text{if } c = d, \\ 0 & \text{if } c \neq d. \end{cases}$$

Definition 1. *Let $u = b_1 \ldots b_k$ be a word, where each b_i, $1 \leq i \leq k$, is a letter of the alphabet $\Sigma = \{a_1, \ldots, a_n\}$. The* Parikh matrix mapping with respect to *u, denoted Ψ_u, is the morphism:*

$$\Psi_u : \Sigma^* \to \mathcal{M}_{k+1},$$

defined, for $a \in \Sigma$, by the condition: if $\Psi_u(a) = M_u(a) = (m_{ij})_{1 \leq i,j \leq (k+1)}$, then for each $1 \leq i \leq (k+1)$, $m_{ii} = 1$, and for each $1 \leq i \leq k$, $m_{i(i+1)} = \delta_{a,b_i}$, all other elements of the matrix $M_u(a)$ being 0. Matrices of the form $\Psi_u(w),\ w \in \Sigma^*$, are referred to as generalized Parikh matrices. *For $u = a_1 \ldots a_n$ they are referred to as* Parikh matrices.

Thus, the generalized Parikh matrix $M_u(w)$ associated to a word w is obtained by multiplying the matrices $M_u(a)$ associated to the letters a of w, in the order in which the letters appear in w. The above definition implies that if a letter a does not occur in u, then the matrix $M_u(a)$ is the identity matrix.

What is the information content of the matrix $M_u(w)$? It is expressed in the following theorem, due originally to [18]. For $1 \leq i \leq j \leq k$, denote $u_{i,j} = b_i \ldots b_j$. The theorem is easy to establish inductively.

Theorem 1. *For all i and j, $1 \leq i \leq j \leq k$, we have $m_{i(1+j)} = |w|_{u_{i,j}}$.*

Returning to our previous example we have, for $w = abcbcacab$ and $u = abc$,

$$M_u(w) = \begin{pmatrix} 1 & 3 & 5 & 5 \\ 0 & 1 & 3 & 5 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Observe that the ordering of the letters is the natural alphabetic one and that the matrix is actually a Parikh matrix (rather than a generalized one).

There can be no dependency in the general case between the nontrivial entries of a Parikh matrix, [15]. It is not easy to find words w and w' such that $|w|_u = |w'|_u$ holds for all proper factors u of the word $abcde$ but $|w|_{abcde} \neq |w'|_{abcde}$. Here we are dealing with an alphabet of five letters. However, for alphabets of any size, corresponding words w and w' can be found. The independence result deals with Parikh matrices but cannot be extended to generalized Parikh matrices. For instance, the entry $(1, 3)$ in the matrix $M_{aab}(w)$ can always be computed from the entry $(1, 2)$.

There is no essential difference between Parikh matrices and generalized ones. It was shown in [19] that, for any generalized Parikh matrix $M_u(w),\ |u| = k$, one can construct a word w' over an ordered alphabet Σ_k with k letters such that $M_u(w) = M_k(w')$, where the latter matrix is the Parikh matrix of w' for Σ_k.

2 Cauchy inequality Parikh matrices can be used to prove various facts, in particular inequalities, concerning the numbers of different subword occurrences. The study of *subword histories*, [8, 14, 16], constitutes a general approach. In this section we will focus the attention on a particular inequality, the *Cauchy inequality*,

$$|w|_y|w|_{xyz} \leq |w|_{xy}|w|_{yz},$$

valid for all words w, x, y, z, [8]. It can be claimed to be a really fundamental property of words, because of its generality and because it reduces to equality in a great variety of cases. The choice for the name of the inequality is motivated by the resemblance to the well-known algebraic Cauchy inequality for real numbers and also by the methods used in the proof. The reader is referred to [8] for further details, as well as a combinatorial proof of the inequality. Below we give a simpler proof based on generalized Parikh matrices. The basic ideas in this proof are due to [18, 19].

We begin with a simple example. Consider the words

$$w = a^{i_1}b^{j_1}c^{k_1},\ x = a^{i_2},\ y = b^{j_2},\ z = c^{k_2}.$$

Clearly, $|w|_y = \binom{j_1}{j_2}$. Straightforward calculations show that

$$|w|_y|w|_{xyz} = \binom{i_1}{i_2}\binom{j_1}{j_2}^2\binom{k_1}{k_2} = |w|_{xy}|w|_{yz}.$$

In general, if

$$w = x_1y_1z_1,\ |w|_x = |x_1|_x = m,\ |w|_y = |y_1|_y = n,\ |w|_z = |z_1|_z = p,$$

then both sides of the Cauchy inequality equal mn^2p and, thus, the inequality is not proper.

Consider, next, words over a one-letter alphabet. If the words w, x, y, z are of lengths n, i, j, k, respectively, then the inequality assumes the form

$$\binom{n}{j}\binom{n}{i+j+k} \leq \binom{n}{i+j}\binom{n}{j+k},$$

which is easily verified to be true. Here we have an equality exactly in case $i = 0$ or $k = 0$. Assume that

$$y = a^i b^j a^k,\ x = a^{i_1},\ z = a^{k_1}$$

and $w = a^{i+i_1+i'} b^{j+j'} a^{k+k_1+k'}$. Then again it is easy to verify that the inequality is not proper. The reader might want to consider more sophisticated examples. For instance, if $w = a^3b^3a^3b^3a^3$, and $y = aba,\ x = z = a$, then $|w|_y|w|_{xyz} = 29160$, whereas $|w|_{xy}|w|_{yz} = 35721$.

Thus, the general result is:

Theorem 2. *For arbitrary words $w,\ x,\ y,\ z$, the inequality*

$$|w|_{xyz}|w|_y \leq |w|_{xy}|w|_{yz}$$

holds true.

Lemma 1. *The value of any 2-dimensional minor of the matrix $M_u(w)$ is a nonnegative integer.*

Proof. The assertion holds if w is a letter. In this case there is no minor, where the upper right and lower left entries are both nonzero. Consequently, 0 and 1 are the only possible values for the minor.

Assume inductively that the assertion holds for the word w', and consider the word $w = w'a$, where a is a letter. Let D be the 2-dimensional minor of $M_u(w)$ determined by the four entries

$$m_{i_\mu,j_\nu},\ 1 \leq \mu,\nu \leq 2,\ 1 \leq i_1 < i_2 \leq k+1,\ 1 \leq j_1 < j_2 \leq k+1.$$

Consider the second column (corresponding to j_2) in D. Its entries are

$$\text{either } m'_{i_\mu,j_2} \text{ or } m'_{i_\mu,j_2} + m'_{i_\mu,j_2-1},\ \mu = 1,2,$$

depending whether a is not or is the j_2th letter in u. The same conclusion holds for the first column (corresponding to j_1) in D. This means that D is the sum of at most four determinants, each of which is either a minor of $M_u(w')$ or consists of two identical columns. The assertion now follows by the inductive hypothesis.

We are now in the position to prove Theorem 2. Consider arbitrary w, x, y, z and denote $u = xyz$. Then

$$\begin{vmatrix} |w|_{xy} & |w|_{xyz} \\ |w|_y & |w|_{yz} \end{vmatrix}$$

appears as a minor in the Parikh matrix $M_u(w)$, by Theorem 1. Hence Theorem 2 follows by Lemma 1. □

We mention finally a "dual" of the Cauchy inequality, [8], interesting on its own right.

Lemma 2. *For all words x, y, z, w,*

$$|xyz|_w|y|_w \leq |xy|_w|yz|_w.$$

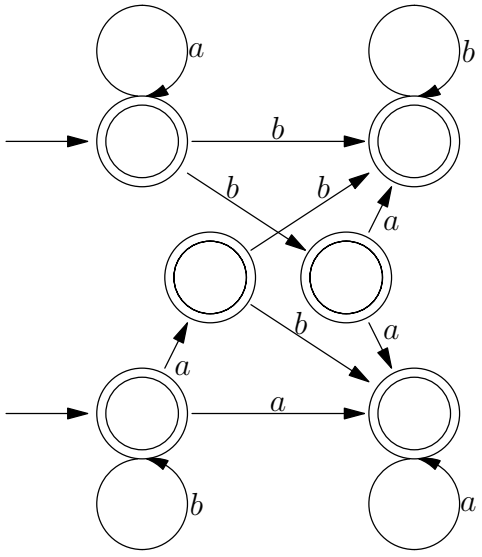
3 M-ambiguity and M-equivalence From now on we consider exclusively Parikh matrices, rather than the generalized ones. (Recall also the remark made at the end of Section 1.) Thus, from now on we consider the alphabet $\Sigma_k = \{a_1, \ldots, a_k\}$, and the word defining the matrices is $u = a_1 \ldots a_k$. (In examples we use letters from the beginning of the English alphabet.) We denote the Parikh matrix mapping by Ψ_k.

Definition 2. *Two words $w_1, w_2 \in \Sigma_k^*$ are termed* M-equivalent*, in symbols $w_1 \equiv_M w_2$, if $\Psi_k(w_1) = \Psi_k(w_2)$. A word $w \in \Sigma_k^*$ is termed* M-unambiguous *if there is no word $w' \neq w$ such that $w \equiv_M w'$. Otherwise, w is termed* M-ambiguous*. If $w \in \Sigma_k^*$ is M-unambiguous (resp. M-ambiguous), then also the Parikh matrix $\Psi_k(w)$ is called unambiguous (resp. ambiguous).*

There is an extensive literature concerning M-equivalence, see [1, 3, 12, 13] and their references. Recently also a unique word has been associated to each Parikh matrix, [17]. The set of M-unambiguous words is known if the alphabet consists of two or three letters. For three letters, the situation is rather complicated, [19], but for two letters the following simple result holds, [5, 9, 12].

Theorem 3. *A word $w \in \{a, b\}^*$ is M-ambiguous if and only if w has the factors ab and ba in non-overlapping positions.*

We present here a nondeterministic automaton accepting the set of M-unambiguous words. The two initial states are indicated by incoming arrows, and all states are final. (In fact, it makes no difference if we consider all states to be initial as well.)



If a word $y \in \Sigma_k^*$ is M-ambiguous, so is every word xyz where $x, z \in \Sigma_k^*$. However, M-unambiguous words may possess M-ambiguous subwords. For instance, the unambiguous word *abcba* has the ambiguous subword *abba*. Below we list some short M-ambiguous words. Consider the alphabet $\Sigma_k = \{a_1, \ldots, a_k\}$. The words

$$a_i a_{i+j} \text{ and } a_{i+j} a_i, \ 1 \leq i \leq k-2, \ 2 \leq j \leq k-i,$$

are ambiguous. So are each of the words

$$\begin{array}{ll} a_ia_{i+1}a_{i+1}a_i, & a_{i+1}a_ia_ia_{i+1}, \\ a_ia_{i+1}a_i^ja_{i+1}a_i, & a_{i+1}a_i^{j+2}a_{i+1}, \\ a_{i+1}a_ia_{i+1}^ja_ia_{i+1}, & a_ia_{i+1}^{j+2}a_i, \end{array}$$

where $1 \leq i \leq k-1$ and $j \geq 0$.

Various methods have been presented for establishing the M-equivalence of two words. We have to worry only about occurrences of factors of the word $a_1 \dots a_k$ as subwords. It follows from the considerations above that changing a factor a_ia_{i+j}, $j \geq 2$, to $a_{i+j}a_i$, or vice versa, yields an M-equivalent word. The factors a_ia_{i+1} and $a_{i+1}a_i$ can be swapped if two such swappings take place in opposite directions, and there are no "harmful" letters between the two swapping positions. Explicitly, this can be stated as the following lemma, [5, 9, 1].

Lemma 3. *Assume that $1 \leq i \leq k-1$ and x, y, z are arbitrary words such that $|y|_{a_{i-1}} = |y|_{a_{i+2}} = 0$. Then*

$$xa_ia_{i+1}ya_{i+1}a_iz \equiv_M xa_{i+1}a_iya_ia_{i+1}z.$$

The two methods presented above (swapping a factor a_ia_{i+j}, $j \geq 2$, to $a_{i+j}a_i$, or vice versa, and Lemma 3) are sufficient for showing the M-equivalence in many cases. Our following theorem exhibits such a construction.

Theorem 4. *Denote $x = a_1 \dots a_k$, and let y be the mirror image of x. Then $xy \equiv_M yx$. Moreover, if w and w' are words composed of the factors x and y such that x (resp. y) appears as a factor equally many times in w and w', then $w \equiv_M w'$.*

Proof. Clearly, the second assertion follows from the first: we just perform sufficiently many swappings between xy and yx. The first assertion, $xy \equiv_M yx$ is established by an induction on k. The assertion clearly holds for $k = 2$. Assume that it holds for the value k, and consider an alphabet with $k+1$ letters. We obtain

$$a_1 \dots a_ka_{k+1}a_{k+1}a_k \dots a_1 \equiv_M a_1 \dots a_{k+1}a_ka_ka_{k+1} \dots a_1.$$

We now swap a_{k+1} with the neighboring letters, yielding the M-equivalent word

$$a_{k+1}a_1 \dots a_ka_k \dots a_1a_{k+1}.$$

By the induction hypothesis, we see that the first assertion holds. □

If the alphabet is not binary, then the two methods presented are definitely not sufficient for showing the M-equivalence of arbitrary words. The illustration considered in Section 1 provides a simple counterexample. It is easy to see that

$$abcbcacab \equiv_M bcacababc.$$

Moreover, the only other words M-equivalent to these two words are obtained by changing the order of letters in the *caca*-factor. Consequently, the two methods are insufficient for showing the M-equivalence of these two words.

In fact, we have the following equivalence criterion corresponding to Lemma 3.

Lemma 4. *Assume that $1 \leq i \leq k-2$ and x, y_1, y_2, z are words such that none of the letters $a_{i-1}, a_{i+2}, a_{i+3}$ appears in y_1, and none of the letters a_{i-1}, a_i, a_{i+3} appears in y_2. Then*

$$\begin{aligned} &xa_ia_{i+1}a_{i+2}y_1a_{i+1}a_{i+2}a_iy_2a_{i+2}a_ia_{i+1}z \equiv_M \\ &\quad xa_{i+1}a_{i+2}a_iy_1a_{i+2}a_ia_{i+1}y_2a_ia_{i+1}a_{i+2}z. \end{aligned}$$

We omit the proof of this lemma. Matters of "circular variance" causing the equivalence above have been investigated more generally in [17]. We hope to return to the resulting equivalence criteria in another context.

4 Conclusion We have shown the importance of matrix constructions for the analysis of numerical properties of words. This new research area contains numerous open problems. Some of them have been hinted at above. Parikh matrices can be extended to concern languages, [3, 8, 9]. How much can be said about a language on the basis of the set of the Parikh matrices associated to its words?

References

[1] Atanasiu, A., Atanasiu, R. and Petre, I., Parikh matrices and amiable words. *Theoretical Computer Science* 390 (2008) 102–109.

[2] Černý, A., On fairness of D0L systems. *Discrete Applied Mathematics* 155 (2007) 1769–1773.

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

[4] Eilenberg, S., *Automata, Languages and Machines, Vol. B* Academic Press, New York (1976).

[5] Fossé, S. and Richomme, G., Some characterizations of Parikh matrix equivalent binary words. *Inform. Proc. Lett.* 92 (2004) 77–82.

[6] Kuich, W. and Salomaa, A. *Semirings, Automata, Languages*. Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[7] Mateescu, A., Salomaa, A., Salomaa, K. and Yu, S., A sharpening of the Parikh mapping. *Theoret. Informatics Appl.* 35 (2001) 551–564.

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

[9] Mateescu, A. and Salomaa, A., Matrix indicators for subword occurrences and ambiguity. *Int. J. Found. Comput. Sci.* 15 (2004) 277–292.

[10] Parikh, R.J., On context-free languages. *J. Assoc. Comput. Mach.*, 13 (1966) 570–581.

[11] Rozenberg, G. and Salomaa, A. (eds.), *Handbook of Formal Languages 1–3*. Springer-Verlag, Berlin, Heidelberg, New York (1997).

[12] Salomaa, A., On the injectivity of Parikh matrix mappings. *Fundamenta Informaticae* 64 (2005) 391–404.

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

[14] Salomaa, A., On languages defined by numerical parameters. In K.G. Subramanian, K. Rangarajan, M. Mukund (eds.):*Formal Models, Languages and Applications*, World Scientific Publishing Company (2006), 320–336 (Chapter 22).

[15] Salomaa, A., Independence of certain quantities indicating subword occurrences. *Theoretical Computer Science* 362 (2006) 222–231.

[16] Salomaa, A. and Yu, S., Subword conditions and subword histories. *Information and Computation* 204 (2006) 1741–1755.

[17] Salomaa, A. and Yu, S., Subword occurrences, Parikh matrices and Lyndon images. Submitted for publication (2008).

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

[19] Şerbănuţă, V.G. and Şerbănuţă, T.F., Injectivity of the Parikh matrix mappings revisited. *Fundamenta Informaticae* 73 (2006) 265–283.

Turku Centre for Computer Science
Joukahaisenkatu 3–5 B, 20520 Turku, Finland
E-mail : asalomaa@utu.fi