GENERALIZED DAVENPORT-SCHINZEL SEQUENCES

by

Radek Adamec

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Mathematics and Statistics

SIMON FRASER UNIVERSITY
June 1992

APPROVAL

Name: Radek Adamec

Degree: Master of Science

Title of thesis: Generalized Davenport-Schinzel Sequences

Examining Committee: A. H. Lachlan
Chair

T. C. Brown, Senior Supervisor

A. Freedman

B. Thomson

J. Nešetřil, External Examiner
Department of Applied Mathematics
Charles University, Prague, Czechoslovakia

Date Approved: June 19, 1992

Abstract

A sequence is called Davenport-Schinzel if it contains no subsequence of the type *ababa* and avoids immediate repetitions of symbols. It was proven by S. Hart and M. Sharir that the maximum length of a Davenport-Schinzel sequence on n symbols is $n\alpha(n)$, where α is the inverse Ackermann function. Here, we consider sequences avoiding an arbitrary finite forbidden subsequence. A complete characterization of forbidden subsequences on two letters with linear upper bounds is given.

Contents

Chapter 1

Introduction

In 1965 Davenport and Schinzel posed the problem of determining the maximum length of a sequence on n letters with no immediate repetition of the same letter, not containing any subsequence of the type *ababa* (i.e. the occurrences of two letters can give no configuration of the type $a \ldots b \ldots a \ldots b \ldots a$).

Originally this problem arose as a combinatorial problem connected with differential equations [1], [2]. In later studies, connections with path compression algorithms in combinatorics were discovered [3], as well as further applications in combinatorial geometry [5]. We will give a geometrical motivation for the problem in Chapter 2 using line segments in the plane.

In 1986 the maximum length of Davenport-Schinzel sequences on n letters was proven nonlinear in n, and a tight "almost linear" upper bound was found [3]. The upper bound is $n\alpha(n)$ where $\alpha(n)$ is the inverse Ackermann function. Later P. Komjáth found a simplified construction for the lower bound [7], which we present in Chapter 4. Considering more general subwords of the type *abab*... of length $s+2$ he proved that the lower bound is $n\alpha^s(n)$.

Here, we consider a natural generalization of the original problem. We study sequences not containing a given forbidden subword (i.e. subsequence) on generally more than two letters. Results regarding this, including those previously obtained in [9], form Chapters 3 to 5.

Chapter 3 deals with general forbidden words. In order to extend the original problem to forbidden words on more than two letters we have to replace the condition of no immediate repetitions by a new notion of k-regularity. A word is k-regular if every two occurrences of the same letter are at least k positions apart. Thus a word without immediate repetitions is 2-regular. A k-regular word avoiding a forbidden subword always has bounded length provided that k is sufficiently large. Excluding the trivial case of small k we show that the order of magnitude of the maximum length of a k-regular word avoiding a forbidden subword f only depends on f. In Theorem 3 we find a sufficient condition for the linearity of the maximum length function. Finally we show in Corollary 2 that every forbidden word f can be reduced to a word red f not having more than two occurrences of the same letter in a row, such that the maximum length function for a word avoiding red f has the same order of magnitude as that for a word avoiding f.

In Chapter 4 we show that the maximum length function for the reduced forbidden word *abbaab* is linear. Combining this result with Komjáth's construction from [7] we give a complete characterization of two-letter forbidden words yielding linear upper bounds.

In Chapter 5 we find a linear upper bound for *abcd*... *mab* and prove a lemma describing the behaviour of words avoiding *abcabc*.

In Chapter 6 we study connections between forbidden subwords and forbidden submatrices in (0,1)-matrices. We present some results from [8] here.

Chapter 7 contains the results of computer searches for various forbidden words as well as some conjectures based on them.

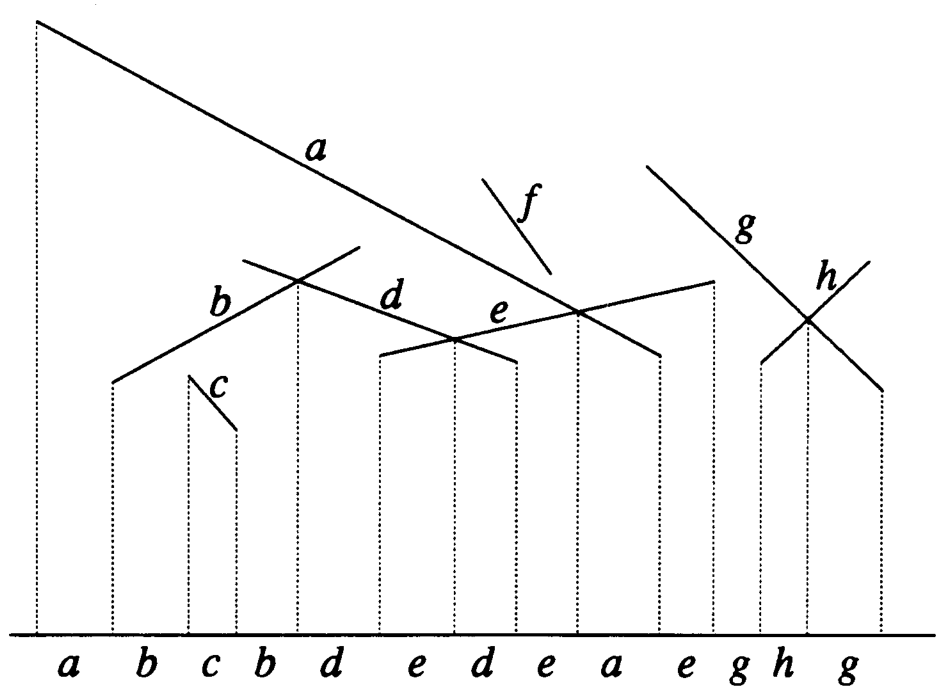
Chapter 2

Geometrical Motivation

We will now give a geometrical motivation for the original problem by Davenport and Schinzel. Consider a set S of n open segments in the plane such that they only have a finite number of points of intersection. We colour the segments with n distinct colours $a_1, a_2, \dots a_n$. This way some points on the plane are coloured uniquely, while the points of intersection have several colours assigned to them. Choose a set of coordinates such that all the segments lie in the halfplane $y \geq 0$. Pick a point $P = (x, 0)$ on the x-axis. If there exists a point $A = (x, y)$ coloured uniquely with a colour c, such that the vertical segment AP does not contain any other colour, we say that the colour c is visible from the point P. We colour each point on the x-axis with the colour visible from it if there is such a colour. We leave the point uncoloured otherwise. This way we obtain a sequence of segments on the x-axis, possibly degenerated to a point, coloured with the colours $a_1, a_2, \dots a_n$. The corresponding sequence on n letters $a_1, a_2, \dots a_n$ will be called the sequence generated by S (see the picture on the next page).

Observation 1 *The sequence generated by S is Davenport Schinzel (i.e. avoids ababa and immediate repetitions).*

Proof: Let w be the sequence generated by S. There are no immediate repetitions because the only way that a sequence can be broken up is by a segment of another colour. Since the segments in S are open, we cannot have a segment interrupted by the endpoint of another segment. Furthermore, an *ababa* configuration requires that two segments intersect in two points. This would mean having an infinite number of points of intersection and a whole segment without unique colouring, which is impossible.



It follows that the maximum length of a sequence generated by n segments is at most that of a Davenport-Schinzel sequence on n letters, which is $O(n\alpha(n))$. A natural question arises whether each Davenport-Schinzel sequence can be realized by segments. This is not generally known. There, however, exists a realization by segments of the Davenport-Schinzel sequences constructed by Hart and Sharir in [3] to prove that the upper bound $O(n\alpha(n))$ is tight. Thus for every n_0 there exists $n \geq n_0$ and a set of n segments generating a sequence of length $\Omega(n\alpha(n))$. This does not mean that the maximum length of a sequence generated by segments is $\Omega(n\alpha(n))$ in the sense we will use Ω for. What it actually proves is that the maximum length is not $o(n\alpha(n))$. (See definition 2 for asymptotic notation.) The construction of segment sets yielding superlinear sequences is due to Wiernik and can be found in [5]. It involves the same double induction that Hart and Sharir use in [3], and we use in Chapter 4 to prove the superlinearity of Davenport-Schinzel sequences.

Chapter 3

General Forbidden Words

Definition 1 *Let A be an infinite alphabet.*

- *By A^* we denote the free monoid over A. The elements of the sets A^* and A will be called words and letters respectively. Furthermore 1_A is the empty word and A^+ is the set of all non-empty words.*

- *Two words $u, v \in A^*$ are isomorphic, $u \cong v$, if $v = \alpha(u)$ for some $\alpha \in \mathrm{Aut}\, A^*$.*

- *Every $w \in A^*$ can be considered a mapping $w : [1, n] \to A$ for some n. Thus we define $\mathrm{rank}\, w = |\mathrm{Im}\, w|$ i.e. the number of letters in w, and $|w| = n$ is the length of w.*

- *We call $w \in A^*$ a word on n letters if $\mathrm{rank}\, w \leq n$.*

- *A word $u \in A^*$ is a factor of $v \in A^*$ if $v = xuy$ for some $x, y \in A^*$.*

- *A word $u \in A^*$ is a subword of $v \in A^*$ if $v = y_0x_1y_1x_2y_2 \ldots x_my_m$ where $u = x_1x_2 \ldots x_m$ and $x_i, y_j \in A^*$.*

- *A word $u \in A^*$ avoids $v \in A^*$ if there exists no subword w of u isomorphic to v.*

- *A word w is regular if $\mathrm{rank}\, w = |w|$ i.e. there are no repetitions of letters in w.*

- *A word w is k-regular if every factor of w of length at most k is regular.*

There exists ambiguity in the asymptotic notation as used by various authors. We will stick to the following definition from "Introduction to Algorithms" by Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest.

Definition 2 *Let $g(n)$ be a nonnegative function on the set of natural numbers.*

- $O(g(n)) = \{f(n)$: *there exist positive constants* c *and* n_0 *such that* $0 \leq f(n) \leq cg(n)$ *for all* $n \geq n_0\}$

- $o(g(n)) = \{f(n)$: *for any positive constant* $c > 0$, *there exists a constant* $n_0 > 0$ *such that* $0 \leq f(n) < cg(n)$ *for all* $n \geq n_0\}$

- $\Omega(g(n)) = \{f(n)$: *there exist positive constants* c *and* n_0 *such that* $0 \leq cg(n) \leq f(n)$ *for all* $n \geq n_0\}$

- $\omega(g(n)) = \{f(n)$: *for any positive constant* $c > 0$, *there exists a constant* $n_0 > 0$ *such that* $0 \leq cg(n) < f(n)$ *for all* $n \geq n_0\}$

- $\Theta(g(n)) = \{f(n)$: *there exist positive constants* c_1, c_2, *and* n_0 *such that* $0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$ *for all* $n \geq n_0\}$

In accordance with generally accepted convention we write $f(n) = O(g(n))$, *or* $f(n)$ *is* $O(g(n))$ *etc. rather than* $f(n) \in O(g(n))$.
We say that $f(n)$ *and* $g(n)$ *are of the same order if* $f(n) = \Theta(g(n))$.

We concentrate on the following problem. A forbidden word $f \in A^+$ is given. Denote by $s(n)$ the maximum length of a k-regular word w on n letters avoiding f. What is the asymptotic behaviour of $s(n)$ as a function in one variable n?

It is easy to observe that $s(n)$ is always a non-decreasing function. If $k < \operatorname{rank} f$ we have $s(n) = +\infty$ for almost all n. (This is because the infinite k-regular sequence $a_1 a_2 \ldots a_k a_1 a_2 \ldots a_k \ldots$ on k letters avoids f.) The following Theorem 1 shows that if $k \geq \operatorname{rank} f$ then $s(n) < +\infty$ for all n.

Theorem 1 *Let* $f \in A^+$, $k \geq \operatorname{rank} f = r$, *and* $s(n)$ *be the maximum length of a k-regular word on* n *letters avoiding* f. *Then* $s(n)$ *is* $O(n^r)$.

Proof: We will prove that

$$s(n) < r|f|\binom{n}{r}.$$

Suppose w is a k-regular word on n letters avoiding f such that

$$|w| \geq r|f|\binom{n}{r}.$$

We can without loss of generality assume that equality holds because if we cut off the end of w, the word is still k-regular and avoids f. We write w as

$$w = w_1 w_2 \dots w_l$$

where $l = |f|\binom{n}{r}$, and $|w_i| = r$ for all i. Since $r \leq k$ and w is k-regular, each factor w_i is regular. Hence $\operatorname{rank} w_i = r$ for all i, and $\operatorname{Im} w_i$ is an r-subset of $\operatorname{Im} w$. As $\operatorname{rank} w \leq n$, there are at most $\binom{n}{r}$ different r-subsets of $\operatorname{Im} w$, and by the pigeon-hole principle there exist

$$\frac{l}{\binom{n}{r}} = |f|$$

factors $w_{i_1}, w_{i_2}, \dots w_{i_{|f|}}$ such that

$$\operatorname{Im} w_{i_1} = \operatorname{Im} w_{i_2} = \dots \operatorname{Im} w_{i_{|f|}}.$$

Taking any bijection from $\operatorname{Im} f$ to $\operatorname{Im} w_{i_1}$ we can find a subword isomorphic to f in w by picking one letter in each factor w_{i_j}. This is a contradiction, so $s(n)$ is bounded by the function $r|f|\binom{n}{r}$, which is a polynomial in one variable n of degree r. Note that we are interested in the maximum length $s(n)$ as a function of n for a fixed forbidden word f yielding constants r and $|f|$.

Remark 1 The function $s(n)$ depends on f and k. To simplify the notation we suppose f and k known by context. We say that $s(n)$ is the maximum length function for (f, k).

Observation 2 *Let $f, g \in A^+$ be two forbidden words, f a subword of g and $k \geq$* rank g ($\geq$ rank f). *Let $s(n)$ and $t(n)$ be the maximum length functions for (f, k) and (g, k) respectively. Then $s(n) \leq t(n)$ for every n.*

Proof: Let $w \in A^*$ be a k-regular word on n letters avoiding f such that $|w| = s(n)$. Since w avoids f it must avoid g, too, and $|w| \leq t(n)$.

Algorithm A(k) Let $w \in A^*$ and $k \geq 1$. We define an algorithm $A(k)$ which, applied to w, finds a k-regular subword of w:

Let $w = a_1 a_2 \dots a_m$, $a_1, a_2 \dots, a_m \in A$ and $w_0 = 1_A$. For $i = 1, 2, \dots, m$ define

$$w_i = \begin{cases} w_{i-1}a_i & \text{if } w_{i-1}a_i \text{ is k-regular} \\ w_{i-1} & \text{otherwise.} \end{cases}$$

The k-regular subword w_m of w is the output of $A(k)$.

Theorem 2 *A forbidden word $f \in A^+$ and integers $k, l \geq$* rank f *are given. Let $s(n)$ be the maximum length function for (f, k) and $t(n)$ the maximum length function for (f, l). Then $s(n)$ and $t(n)$ are of the same order.*

Proof: Without loss of generality $k < l$. Obviously $t(n) \leq s(n)$ for every n because each l-regular word is k-regular as well. Now let w_0 be a k-regular word on n letters avoiding f such that $|w_0| = s(n)$. For $i = 1, 2, \dots, l-k$ we define w_i by applying $A(k+i)$ to w_{i-1}. In the end we get an l-regular subword w_{l-k} avoiding f. Then $|w_{l-k}| \leq t(n)$. For $i = 1, 2, \dots, l-k$ the word w_{i-1} can be written in the form

$$w_{i-1} = a_1 v_1 a_2 v_2 \dots a_p v_p$$

where $a_j \in A$, $v_j \in A^*$, $w_i = a_1 a_2 \dots a_p$ and v_j are the sections left out by the algorithm $A(k+i)$. The algorithm $A(k+i)$ reads letters from w_{i-1} one by one from the left to the right, and includes some of them in w_i. At the moment that the first letter of the factor v_j is to be read, the output produced so far is $a_1 a_2 \dots a_j$. The first letter of v_j is not accepted by $A(k+i)$, which means that it occurs among the last $k+i-1$ letters of $a_1 a_2 \dots a_j$. Every following letter of v_j is rejected through the same criterion, and therefore all letters from v_j are included among the last $k+i-1$ letters of $a_1a_2\dots a_j$. It follows immediately that

$$\operatorname{rank} v_j \leq k+i-1\,.$$

Addition of one letter can only increase the rank of a word by at most one, so

$$\operatorname{rank} a_jv_j \leq k+i$$

and since w_{i-1} avoids f we have

$$|a_jv_j| \leq s(k+i)\,.$$

Therefore

$$|w_{i-1}| \leq |w_i|s(k+i)$$

and finally

$$s(n) = |w_0| \leq |w_{l-k}|c \leq ct(n)$$

where $c = s(k+1)s(k+2)\dots s(l)$ is a constant independent of n.

Remark 2 Let $s(n)$ be the maximum length function for (f,k) where $k \geq \operatorname{rank} f$. It follows from Theorem 2 that $\Theta(s(n))$ is independent of k. As we are interested in the asymptotic behaviour of $s(n)$ we do not have to care about k (providing $k \geq \operatorname{rank} f$). So from now on saying $s(n)$ is the maximum length function for $f \in A^+$ we will mean $s(n)$ is the maximum length function for $(f, \operatorname{rank} f)$.

Observation 3 *The maximum length function $s(n)$ is $\Theta(1)$ if and only if the forbidden word f is regular.*

Proof: Let f be regular, $r = \operatorname{rank} w$, and w avoid f. If $|w| \geq r$ then the initial factor v of w of length r contains each letter at most once because w is k-regular. This would mean $v \cong f$, which cannot happen, so the length of w must be less than r. On the other hand if f is not regular then the regular word $w = a_1a_2\dots a_n$ on n letters always avoids f, and $s(n) \geq n$.

From now on we will not consider the trivial case of a regular forbidden word. Thus we can always assume that $s(n) \geq n$ holds for all n, and consequently $s(n) = \Omega(n)$.

Definition 3 *By A^f we denote the set of all non-regular words from A^+.*

Lemma 1 *Let $f \in A^f$ be a forbidden word, and $k \geq \operatorname{rank} f$. There exists a constant d satisfying the following property:*

> *For each k-regular word $w = uav$ avoiding f, where $a \in A$ and $u, v \in A^*$, there exists a k-regular subword w' of uv avoiding f such that $|w'| \geq |w| - d$.*

Proof: Denote $l = |f|$ and

$$d = \binom{3k-4}{k} k(l-1) + 2 .$$

First suppose $|u| \geq k-1$ and $|v| \geq d + 2k - 3$.We can divide w into disjoint factors:

$$w = u_1 u_2 a v_1 v_2 v_3$$

where $|u_2| = k-1$,$|v_1| = d + k - 2$ and $|v_2| = k - 1$. Furthermore we can express

$$v_1 = x_1 x_2 \dots x_{\frac{d+k-2}{k}}$$

such that $|x_j| = \operatorname{rank} x_j = k$. Suppose $\operatorname{rank} v_1 \leq 3k - 4$. By the pigeon-hole principle there are at least

$$\left\lceil \frac{d+k-2}{k} \Big/ \binom{3k-4}{k} \right\rceil = l$$

disjoint factors $x_{j_1}, x_{j_2}, \dots, x_{j_l}$ among all of the factors x_j, such that

$$\operatorname{Im} x_{j_1} = \operatorname{Im} x_{j_2} = \dots = \operatorname{Im} x_{j_l} .$$

Then we can find a subword of w isomorphic to f taking one letter in each factor x_{j_i}, which is a contradiction. Therefore $\operatorname{rank} v_1 \geq 3k - 3$. There must be at least $k-1$ different letters $a_1, a_2, \dots, a_{k-1}$ occurring in v_1 none of which occurs in $u_2 v_2$ because $\operatorname{rank} u_2 v_2 \leq 2k - 2$. We get

$$v_1 = y_0 a_1 y_1 a_2 y_2 \dots a_{k-1} y_{k-1} .$$

We form a new word

$$w' = u_1 u_2 a_1 a_2 \dots a_{k-1} v_2 v_3 .$$

It is obviously a k-regular subword of uv avoiding f, and $|w'| = |w| - d$. If $|u| < k-1$ or $|v| < d + 2k - 3$ the proof is analogical or even easier.

Theorem 3 *Let $f \in A^f$, $k \geq \text{rank}\, f$. Let c be a constant such that in each k-regular w avoiding f there is a letter occurring at most c times. Then $s(n)$ is $\Theta(n)$.*

Proof: Denote $l = |f|$. We are going to prove that the length of a k-regular word w on n letters avoiding f is at most cdn where d is the constant (dependent on l) from Lemma 1. For $n = 0$ the statement is obvious. Now let w be a word on n letters and $a \in A$ a letter occurring at most c times in w. Using Lemma 1 we can find a k-regular subword w' of w with no occurrences of a such that $|w'| \geq |w| - cd$. Since w' is a word on $n-1$ letters we can use induction to get $|w'| \leq cd(n-1)$ and we are done.

Theorem 4 *Let $f, g \in A^f$ be two forbidden words such that*

$$f = av$$

$$g = a^2 v$$

where $a \in A$ and $v \in A^*$. Then the maximum length functions $s(n)$ and $t(n)$ for f and g respectively are of the same order.

Proof: Denote $k = \text{rank}\, f = \text{rank}\, g$ and $l = |g|$. Since f is a subword of g we have $s(n) \leq t(n)$ by Observation 2. Now let w be a k-regular word on n letters avoiding g such that $|w| = t(n)$. We form a subword w' of w by leaving out the first occurrence of each letter from w. Using Lemma 1 we can obtain a k-regular subword w'' of w' such that $|w''| \geq |w| - nd$. Furthermore w'' avoids f. Hence $|w''| \leq s(n)$ and

$$t(n) \leq s(n) + nd \leq s(n)(1 + d)$$

which proves Theorem 4.

Corollary 1 *It follows from Theorem 3 that for forbidden words*

$$f = avb$$

$$g = a^k v b^l$$

where $f \in A^f$, $a, b \in A$, $v \in A^*$ and $k, l \geq 1$ the maximum length functions $s(n)$ and $t(n)$ are of the same order.

Theorem 5 *Let $f, g \in A^f$ be two forbidden words such that*

$$f = a_1^{\alpha_1} a_2^{\alpha_2} \dots a_m^{\alpha_m}$$

$$g = a_1^{2\alpha_1 - 1} a_2^{2\alpha_2 - 1} \dots a_m^{2\alpha_m - 1}$$

where $a_j \in A$, $a_j \neq a_{j+1}$ and $\alpha_j \geq 1$. Then the maximum length functions $s(n)$ and $t(n)$ for f and g respectively are of the same order.

Proof: Denote $k = \operatorname{rank} f = \operatorname{rank} g$. As f is a subword of g we have $s(n) \leq t(n)$ by Observation 2. Now let w be a k-regular word on n letters avoiding g such that $|w| = t(n)$. We form a subword w' of w by leaving out every other occurrence, starting with the second one, of every letter in w. Obviously w' avoids f and $|w'| \geq \frac{1}{2}|w|$, yet it may not be k-regular. We apply $A(k)$ to w' and get a k-regular subword w'' of w'. The word w'' divides w' the following way:

$$w'' = a_1 a_2 \dots a_p$$

$$w' = a_1 x'_1 y'_1 a_2 x'_2 y'_2 \dots a_p x'_p y'_p$$

where $a_j \in A$ and $x'_j, y'_j \in A^*$. For each $j = 1, 2, \dots, p$ the factor $x'_j y'_j$ is the factor omitted by $A(k)$. The factors x'_j and y'_j are chosen such that $|x'_j|$ is divisible by k, and $|y'_j| < k$. It is also possible to express w in the form

$$w = a_1 x_1 y_1 a_2 x_2 y_2 \dots a_p x_p y_p$$

such that x'_j is a subword of x_j and similarly y'_j is a subword of y_j for all $j = 1, 2, \dots, p$. Now consider any $j, 1 \leq j \leq p$. The words x'_j and x_j can be divided as follows:

$$x'_j = z'_1 z'_2 \dots z'_q$$

$$x_j = z_1 z_2 \dots z_q$$

where $|z'_i| = k$ and z'_i is a subword of z_i for every $i, 1 \leq i \leq q$. The algorithm $A(k)$ read the letters from w' one by one from the left to the right, and included some of them in w''. At the moment that the first letter of the factor $x'_j y'_j$ was to be read, the output produced so far was $a_1 a_2 \dots a_j$. The first letter of $x'_j y'_j$ was not accepted by

$A(k)$, which meant that it occurred among the last $k-1$ letters of $a_1a_2\dots a_j$. Every other letter from $x'_jy'_j$ was rejected through the same criterion, and therefore all the letters from $x'_jy'_j$ were among the last $k-1$ letters of $a_1a_2\dots a_j$. It follows that

$$\operatorname{rank} x'_jy'_j \le k-1\,,$$

so

$$\operatorname{rank} z'_i \le k-1$$

for every i. There must be a letter $a \in A$ occurring twice in z'_i. It follows from the construction of w' that a occurs at least three times in z_i. Since z_i is k-regular we have $|z_i| \ge 2k+1$ for every i and $|x_j| \ge q(2k+1)$. Realizing that $|x'_j| = qk$ we finally get

$$|x'_j| \le \frac{k}{2k+1}|x_j|\,.$$

This holds for every $j = 1, 2, \dots, p$. Summing up:

$$|x'_1x'_2\dots x'_p| \le \frac{k}{2k+1}|x_1x_2\dots x_p| \le \frac{k}{2k+1}|w|$$

and

$$|a_1y'_1a_2y'_2\dots a_py'_p| = |w'| - |x'_1x'_2\dots x'_p| \ge \frac{1}{2}|w| - \frac{k}{2k+1}|w| = \frac{1}{4k+2}|w|\,.$$

At the same time $|a_jy'_j| \le k$ for every j, i.e.

$$|a_j| \ge \frac{1}{k}|a_jy'_j|$$

and finally

$$|w''| = |a_1a_2\dots a_p| \ge \frac{1}{k}|a_1y'_1a_2y'_2\dots a_py'_p| \ge \frac{1}{k(4k+2)}|w|\,.$$

The word w'' is a k-regular word on n letters avoiding f, hence

$$s(n) \ge |w''| \ge \frac{1}{k(4k+2)}t(n)$$

which proves Theorem 5.

Definition 4 *For every positive integer n we put*

$$\operatorname{red} n = \begin{cases} 2 & \text{if } n \geq 2 \\ 1 & \text{if } n = 1. \end{cases}$$

Furthermore let

$$u = a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_m^{\alpha_m}$$

where $a_j \in A$, $a_j \neq a_{j+1}$ and $\alpha_j \geq 1$ for all j. We define the reduced subword of u as

$$\operatorname{red} u = a_1 a_2^{\operatorname{red} \alpha_2} a_3^{\operatorname{red} \alpha_3} \ldots a_{m-1}^{\operatorname{red} \alpha_{m-1}} a_m \,.$$

This definition enables us to formulate the following corollary:

Corollary 2 *Let $f, g \in A^f$ be two forbidden words such that $f = \operatorname{red} g$. Then the maximum length functions for f and g are of the same order.*

Proof: Let $g = a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_m^{\alpha_m}$ where $a_j \in A$, $a_j \neq a_{j+1}$ and $\alpha_j \geq 1$. Denote $f_0 = a_1^{\operatorname{red} \alpha_1} a_2^{\operatorname{red} \alpha_2} \ldots a_m^{\operatorname{red} \alpha_m}$. By Corollary 1 the maximum length functions for f and f_0 are of the same order. For $i \geq 1$ we define f_i as follows. If

$$f_{i-1} = a_1^{\beta_1} a_2^{\beta_2} \ldots a_m^{\beta_m}$$

then

$$f_i = a_1^{2\beta_1 - 1} a_2^{2\beta_2 - 1} \ldots a_m^{2\beta_m - 1} \,.$$

By Theorem 5 the maximum length functions for f_{i-1} and f_i are of the same order. Obviously there exists j such that g is a subword of f_j. At the same time f is a subword of g. Let $s(n)$, $t(n)$ and $r(n)$ be the maximum length functions for f, g and f_j respectively. We can use Observation 2 taking $k = \operatorname{rank} f_j = \operatorname{rank} g = \operatorname{rank} f$ to obtain $s(n) \leq t(n) \leq r(n)$ for every n. Since $s(n)$ and $r(n)$ are of the same order we have $s(n) = \Theta(t(n))$.

Chapter 4

Forbidden Words on Two Letters

Theorem 6 *Let $f = abba$, $a, b \in A$, and $a \neq b$. Then $s(n)$ is $\Theta(n)$.*

Proof: Let w be a 2-regular word avoiding f. Suppose that each letter occurs at least 6 times in w. The factor beginning with the second and ending with the last but one occurrence of a letter x will be called the body of x. We will show that the bodies of two different letters cannot intersect. More precisely, the body of a letter x cannot contain any occurrence of y from the body of y. Assume the opposite. Then there exists the following configuration in w.

$$x \ldots \underbrace{x \ldots y \ldots x} \ldots x$$

where $\underbrace{x \ldots x}$ is the body of x. Since the occurrence of y comes from the body of y, there must be another occurrence to the left of it. If the other occurrence were to the left of the first x, we would have a subword $yxxy$ in w. If it were right of the first x, there would be an $xyyx$ in w. Either way there is a subword isomorphic to f in w, which is a contradiction. Hence the bodies of two different letters are disjoint. The body of a letter x contains at least four occurrences of x and three gaps between neighbouring occurrences of x. If the number of letters in w is n, then there are at least $3n$ gaps between two neighbouring occurrences from the body of a letter. Each of these gaps must be filled by another letter to insure 2-regularity. An occurrence filling such a gap cannot come from the body of any letter, so it must be the first or the last occurrence. There are, however, only $2n$ first and last occurrences of letters in w, which is not enough to fill all the gaps. It follows from this contradiction that there exists a letter occurring at most five times in w, and by Theorem 3 the function $s(n)$ is $\Theta(n)$.

Corollary 3 *Let $f = a^i b^j a^k \in A^f$, where $a, b \in A$, $a \neq b$, and $i, j, k \geq 0$. Then $s(n)$ is $\Theta(n)$.*

Proof: The reduced subword $\operatorname{red} f$ is a subword of *abba*, and by Theorem 6 its maximum length function is $O(n)$. Since $f \in A^f$, we also have $s(n) = \Omega(n)$ for both f and $\operatorname{red} f$, and consequentely $s(n) = \Theta(n)$ for f.

In the proof of Theorem 6 we used the fact that the factors determined by two different letters cannot intersect "too much". We will use a similar technique for $f = abbaab$. The proof is, however, much more complicated.

Theorem 7 *Let $f = abbaab$, $a \neq b$, $a, b \in A$ be the forbidden word, $s(n)$ the maximum length function for f. Then $s(n)$ is $\Theta(n)$.*

Proof: Let $w \in A^*$ and c occur at least three times in w. The factor beginning with the second occurrence of c ending with the last but one occurrence of c will be called the body of c. The minimal factor containing all occurrences of c is the hull of c. First we prove the following lemma:

Lemma: Let w be a 2-regular word avoiding f, in which each letter occurs at least 12 times. We leave out the first and the last occurrence of every letter to get w'. If $c, d \in \operatorname{Im} w = \operatorname{Im} w'$ then there are the following four possibilities for their occurrences in w'. (We assume that c occurs before d in w'.)

(*i*) $\underbrace{c\ldots c}\ \underbrace{d\ldots d}$

(*ii*) $\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$

(*iii*) $\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$

(*iv*) $\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}$ or

$\underbrace{c\ldots c}\ d\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}$

where $\underbrace{c\ldots c}$ stands for a factor containing at least one c and no d, similarly for d.

Proof of lemma: We will discuss several possible cases of w.

(1) The body of c contains at least two occurrences of d.

(a) The body of c contains all occurrences of d. If there were three occurrences of c in the hull of d there would be a subword isomorphic to f in w. If there are two occurrences of c in the hull of d, the situation in w looks like this:

$$c\ \underbrace{c\ldots c}\ d\ c\ \underbrace{d\ldots d}\ c\ d\ \underbrace{c\ldots c}\ c$$

otherwise w would not avoid f. Then the situation in w' looks like this:

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$$

which is (*ii*). If there is one occurrence of c, or none, in the hull of d, we have the following possibilities for w:

$$c\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}\ c$$

or

$$c\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}\ c.$$

Then for w' we get

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$$

or

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$$

or

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ \underbrace{c\ldots c}$$

which is (*iii*), (*ii*) or (*ii*).

(b) There is d outside the body of c. Again there are two possibilities for w:

$$\underbrace{c\ldots c}\ d\ c\ \underbrace{d\ldots d}\ c\ d\ c$$

or

$$c\ \underbrace{c\ldots c}\ d\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}\ d\ c$$

(or the symmetrical ones). In w' we have

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c$$

or

$$\underbrace{c\ldots c}\ \underbrace{d\ldots d}\ c\ \underbrace{d\ldots d}$$

which is (ii) or (iv). (We get the same for the symmetrical cases.)

(2) There is at most one occurrence of d in the body of c. Then there is at most one occurrence of d in the hull of c in w', so we have the following diagrams for w':

$$\underbrace{d\ldots d}\ \underbrace{c\ldots c}\ d\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}$$

where one of the outside factors may be empty, or

$$\underbrace{d\ldots d}\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}$$

with one of the outside factors possibly empty. This means we have (iii) or (ii) in case of non empty outside factors and (iv) or (i) for empty outside factors. This completes the proof of the lemma.

Now we can go on with the proof of Theorem 7. Let w be a 2-regular word avoiding f. Suppose for contradiction that each letter of $\operatorname{Im} w$ occurs at least 12 times in w. We construct w' as in the lemma. Hence $\operatorname{Im} w' = \operatorname{Im} w$ and the lemma holds. Define a partial order $\trianglelefteq$ on $\operatorname{Im} w$ as follows:

$c \trianglelefteq d$ iff every occurrence of c is in the hull of d in w' or $c = d$.

Let $c, d, e \in \operatorname{Im} w'$, $e \trianglelefteq c$, $e \trianglelefteq d$ and c and d be incomparable. There are two possibilities for their occurrences in w':

$$\underbrace{c\ldots c}\qquad \overbrace{d\ \underbrace{c\ldots c}}^{\text{all } e \text{ must be here}}\qquad \underbrace{d\ldots d}$$

and

$$\underbrace{c\ldots c}\qquad \overbrace{\underbrace{d\ldots d}\ c}^{\text{all } e \text{ must be here}}\qquad \underbrace{d\ldots d}\ .$$

(Without loss of generality c occurs before d.) In the first case the ordered pair (e,d) will be called bad, in the second case the pair (e,c) is bad. We leave out all bad pairs from $\trianglelefteq$ and get a new relation $\trianglelefteq'$. We define the partial order $\preceq$ as the transitive hull of $\trianglelefteq'$. Obviously $\preceq$ is a suborder of $\trianglelefteq$. From now on by a successor we will always mean the successor in $\preceq$, the same for predecessors. Now we formulate some easy observations:

(1) Every element from $\operatorname{Im} w'$ has at most one immediate successor so the Hasse diagram of $\preceq$ is acyclic.
Proof of **(1)***:* If an element from $\operatorname{Im} w'$ had two immediate successors, it would form a bad pair with one of them. If there were a cycle in the Hasse diagram of $\preceq$, then the two neighbours of a minimal element of the cycle would have to be its immediate successors.

(2) If $c \preceq d$, $c \neq d$ then by the lemma above we have either

$$\underbrace{d\ldots d}\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}$$

or

$$\underbrace{d\ldots d}\ \underbrace{c\ldots c}\ d\ \underbrace{c\ldots c}\ \underbrace{d\ldots d}$$

in w'.

Now we form the subword w'' of w' by leaving out the first and the last occurrence of each letter from w'. Again $\operatorname{Im} w'' = \operatorname{Im} w'$.

(3) If c and d are incomparable then they cannot intersect in w'', i.e. we have either $\underbrace{c\ldots c}\ \underbrace{d\ldots d}$ or $\underbrace{d\ldots d}\ \underbrace{c\ldots c}$ in w''.

Let $a_1, a_2, \ldots, a_m$ be a linear extension of $\preceq$. For $i = 1, 2, \ldots, m$ let w_i'' be the subword of w'' containing all occurrences of $a_1, a_2, \ldots, a_i$ and no occurrence of $a_{i+1}, a_2, \ldots, a_m$. Of course $w''_0 = 1_A$ and $w''_m = w''$. For $i = 1, 2, \ldots, m$ we can express uniquely:

$$w''_i = u''_0 a_i u''_1 a_i u''_2 a_i \ldots u''_{p-1} a_i u''_p ,$$

$$w''_{i-1} = u''_0 u''_1 u''_2 \ldots u''_{p-1} u''_p .$$

(4) Factors $u''_1, u''_2, \ldots, u''_{p-1}$ contain only predecessors of a_i.
Proof of **(4)***:* This follows from **(3)**; w''_{i-1} contains only predecessors of a_i and letters incomparable with a_i.

Analogically we can uniquely express

$$w' = u'_{-1} a_i u'_0 a_i u'_1 a_i \ldots u'_{p-1} a_i u'_p a_i u'_{p+1} .$$

(5) For $j = 1, 2, \ldots, p-1$ the factor u''_j is a subword of u'_j; furthermore w''_i is a subword of w'.

(6) Let $a_k \preceq a_i$. The letter a_k only occurs either in a unique factor u'_j for some $j, 0 \leq j \leq p$, or in two neighbouring factors u'_{j-1} and u'_j for some $j, 1 \leq j \leq p$.
Proof of **(6)***:* This follows from **(2)**.

(7) If $u'_j, 0 \leq j \leq p$ contains a predecessor a_k of a_i then it contains an immediate predecessor a_l of a_i.
Proof of **(7)***:* Let a_l be an immediate predecessor of a_i such that $a_k \preceq a_l \preceq a_i$. Considering **(2)** for a_k and a_l we observe that a_i must occur before a_k in w' so there is a_l in $u'_{-1} u'_0 \ldots u'_j$. Similarly a_l occurs in $u'_j u'_{j+1} \ldots u'_{p+1}$ and because of **(6)** we have a_l in u'_j.

(8) If there is a predecessor a_k of a_i in both u'_{j-1} and u'_j for some $j, 1 \leq j \leq p$ then there is an immediate predecessor a_l of a_i occurring in both u'_{j-1} and u'_j.

Proof of **(8)***:* As in the proof of **(7)** we find a_l such that $a_k \preceq a_l \preceq a_i$. By **(2)** a_l occurs in both $u'_{-1}u'_0 \ldots u'_{j-1}$ and $u'_j u'_{j+1} \ldots u'_{p+1}$, and by **(6)** a_l occurs in both u'_{j-1} and u'_j.

(9) Any letter a_k occurring in w''_{i-1} can only occur in a unique factor u''_j for some $j, 0 \leq j \leq p$ or in two neighbouring factors u''_{j-1} and u''_j for some $j, 1 \leq j \leq p$.
Proof of **(9)***:* If there is a letter a_k occurring in two non neighbouring factors u''_j and u''_l then we get a subword $a_k a_i a_i a_k$ in w''_i where by **(3)** a_k and a_i are not incomparable so $a_k \preceq a_i$. By **(5)** we get $a_k a_i a_i a_k$ in w', which is a contradiction with **(2)**.

Denote by p_i the number of immediate predecessors of a_i. We are going to enumerate the number r_i of immediate repetitions in w_i by induction. Obviously $r_0 = 0$. Now let $1 \leq i \leq m$. If $|u''_j| \geq 0$ for some $j, 1 \leq j \leq p-1$, then by **(4)** and **(5)** u'_j contains a predecessor of a_i and by **(7)** u'_j contains an immediate predecessor of a_i. It follows from **(6)** that there are at most $2p_i$ factors among $u'_1, u'_2, \ldots, u'_{p-1}$ containing an immediate predecessor of a_i so there are at most $2p_i$ factors of positive length among $u''_1, u''_2, \ldots, u''_{p-1}$. In other words there are at least $p - 1 - 2p_i \geq 7 - 2p_i$ immediate repetitions of a_i in w''_i. If there is an immediate repetition R of a_k in w''_{i-1} separated by some occurrences of a_i in w''_i then by **(9)** there exists $j, 1 \leq j \leq p$ such that a_k occurs only in u''_{j-1} and u''_j, and the immediate repetition R is separated by only one occurrence of a_i, namely the j^{th} one. Using **(4)** we get $a_k \preceq a_i$ and by **(5)** a_k occurs in both u'_{j-1} and u'_j. Now we can use **(8)** to get an immediate predecessor a_l of a_i in both u'_{j-1} and u'_j. We will map R to a_l. Two different immediate repetitions in w''_{i-1} separated by a_i in w''_i must be separated by two different occurrences of a_i so they are mapped to two different immediate predecessors of a_i. (This follows from **(6)**.) Consequently there are at most p_i immediate repetitions in w''_{i-1} separated by a_i. Summing up,

$$r_i \geq r_{i-1} + 7 - 3p_i \,,$$

$$r_m \geq 7m - 3\sum_{i=1}^{m} p_i \,.$$

The sum of the p_i is the number of edges in the acyclic Hasse diagram of $\preceq$ on m vertices, which is at most $m-1$, therefore

$$r_m \geq 4m+3\,.$$

There are at least $4m+3$ immediate repetitions in w''_{i-1}. These repetitions must all be separated in the 2-regular word w by the $4m$ omitted occurrences of the m letters of $\operatorname{Im} w$. This is a contradiction so there is a letter occurring at most 11 times in w. Using Theorem 3 we get Theorem 7.

Corollary 4 *Let $f \in A^f$ be a forbidden word on two letters avoiding ababa and let $s(n)$ be the maximum length function for f. Then $s(n)$ is $\Theta(n)$.*

Proof: A word f on two letters avoids $ababa$ if and only if $f \cong a^i b^j a^k b^l$ where $a, b \in A$ and $i, j, k, l \geq 0$. Then $\operatorname{red} f$ is a subword of $abbaab$ and by Theorem 7 its maximum length function $s(n)$ is $O(n)$. Taking $f \in A^f$ we get $s(n) = \Theta(n)$.

Now we will prove the superlinearity of $s(n)$ for $f = ababa$ using Komjáth's construction from [7]. Komjáth uses double induction to prove the following statement for all values of k and m.

Statement S(k,m) There exists a number $n = F_k(m)$ and a word w formed from mn different letters (i.e. $\operatorname{rank} w = mn$) decomposed into square and round blocks, such that the following hold.

- (i) The word w avoids $ababa$.
- (ii) There are at most $3n-2$ blocks.
- (iii) Every letter occurs at least k times.
- (iv) There are n square blocks, each of length m, together containing every letter exactly once.
- (v) Square blocks are separated from each other and from both ends of the word by round blocks.
- (vi) If a and b are different letters, only one round block can contain

both of them.

(vii) For every letter, its unique occurrence in a square block is either its first or its last occurrence in the word.

A block means a regular factor. By saying that a word is decomposed into square and round blocks we simply mean that the word is decomposed into blocks, some of which we decide to call square; the remaining blocks are round. The unique occurrence of a letter in a square block is called its square occurrence. The square occurrence of each letter is either of the first or of the last type according to whether it is the first or the last occurrence of the letter in the word.

Proof of **S(1,m)***:* We claim $n = 3$. If $m > 1$ then $3m > 4$ and we can construct the following word on $3m$ letters $a_1, a_2, \dots, a_{3m}$.

$$w = (a_1)[a_1a_2 \dots a_m](a_2)[a_{m+1}a_{m+2} \dots a_{2m}](a_3)[a_{2m+1}a_{2m+2} \dots a_{3m}](a_4)\,.$$

For $m = 1$ we put

$$w = (a_1)[a_2](a_1)[a_3](a_3)[a_1](a_2)\,.$$

Square brackets denote square blocks, round brackets round blocks. The blocks are really regular factors, and the rank of w is as required. We check the conditions.

(i) This holds for $m > 1$ because no letter occurs more than twice. The case of $m = 1$ is easy to check.

(ii) There are exactly $7 = 3n - 2$ blocks.

(iii) Every letter occurs at least once.

(iv) There are three square blocks, each of length m, together containing each letter exactly once.

(v) Square blocks are separated as required.

(vi) This holds because the length of each round block is 1.

(vii) Holds for $m > 1$ because each letter occurs at most twice and therefore every occurrence of a letter is either first or last. For $m = 1$ we check the three letters in square blocks.

Deduction of **S(k,1)** *from* **S(k-1,2)**: The word w witnessing $S(k-1,2)$ has two-letter square blocks. We replace each square block $B = ab$ of w by

- $[\,a\,](\,a\,)[\,b\,](\,b\,)$ if a, b are both of the first type
- $(\,a\,)[\,a\,](\,b\,)[\,b\,]$ if a, b are both of the last type
- $(\,b\,)[\,a\,](\,a\,)[\,b\,]$ if a is of the first type and b is of the last type
- $(\,a\,)[\,b\,](\,b\,)[\,a\,]$ if a is of the last type and b is of the first type

We obtain a new word w' witnessing $S(k,1)$ for $n = F_k(1) = 2F_{k-1}(2)$. The blocks of w' stay regular; $\operatorname{rank} w' = \operatorname{rank} w = n$. We have to check the conditions.

(i) Suppose w' does not avoid f. Then the subword isomorphic to f must contain a letter duplicated in the transition from w to w'. However, in the first two cases only one occurrence of the duplicated letter can be used. If these were the only types of duplication used in the forbidden subword, we would have the same forbidden subword in w as well. In the third case we cannot have *ababa* or *babab* because a is of the first, and b is of the last type, and their occurrences in w' have the form

$$\underbrace{b\ldots b}\; a\; a\; b\; \underbrace{a\ldots a}\ .$$

We cannot have either of *ababa* and *babab* in the fourth case for much the same reasons; the occurrences of a and b are now

$$\underbrace{a\ldots a}\; b\; b\; a\; \underbrace{b\ldots b}\ .$$

If only a (or b) from one of the last two cases is used in the forbidden subword then, as in the first two cases, only one occurrence of a (or b) can be used, and again we have a forbidden subword in w. Therefore condition (i) holds.

(ii) The number of blocks of w is at most $3F_{k-1}(2) - 2$; the number of square blocks is $F_{k-1}(2)$. We have replaced each square block by four blocks. The number of blocks of w' is then at most $3F_{k-1}(2) - 2 + 3F_{k-1}(2) = 3[2F_{k-1}(2)] - 2 = 3n - 2$.

(iii) Every letter occurs at least k times in w and exactly once in some square block of w. Duplicating all the letters in square blocks we increase the number of occurrences of each letter by one, and consequently each letter occurs at least k times in w'.

(iv) In w there are $F_{k-1}(2)$ square blocks, each with two letters. Every square block of w yields two square blocks of w' of length 1. This way we get exactly $2F_{k-1}(2) = n$ square one-letter blocks in w'. They contain each letter exactly once because the blocks of w did.

(v) Two square blocks of w' obtained by the breakup of a square block of w are separated by a duplicated letter forming a round block. All other pairs of square blocks of w' are separated by the original round blocks of w, and so are the ends of w'.

(vi) New round blocks only contain one letter, so condition (vi) holds in w' if and only if it holds in w.

(vii) In all four cases we preserved the type of the square occurrence of a letter by putting the new round occurrence to the right for letters of the first type, and to the left for letters of the last type.

Proof of **S(k,m+1)** *assuming* **S(k,m)** *and* **S(k-1,t)** *for every* t*:* By the inductive hypothesis there exist numbers $n = F_k(m)$ and $N = F_{k-1}(n)$ satisfying $S(k,m)$ and

$S(k-1,n)$. We pick words $v_1, v_2, \ldots, v_N$ witnessing $S(k,m)$ on disjoint sets of letters. Consider the word

$$v = v_1 v_2 \ldots v_N \,.$$

We know that $\operatorname{rank} v_i = mn$, so $\operatorname{rank} v = mnN$. Let u be a word disjoint from v witnessing $S(k-1,n)$; of course $\operatorname{rank} u = nF_{k-1}(n) = nN$. The word u has exactly N square blocks. It can be written as

$$u = r_0 s_1 r_1 s_2 r_2 \ldots s_N r_N$$

where $s_1, s_2, \ldots, s_N$ are the square blocks, and each r_i contains at least one round block. Similarly, each of $v_1, v_2, \ldots, v_N$ has exactly n square blocks. We can write

$$v_i = r_{i,0} s_{i,1} r_{i,1} s_{i,2} \ldots s_{i,n} r_{i,n} \,.$$

We are going to build a word w witnessing $S(k, m+1)$. We set

$$w = r_0 t_1 r_1 t_2 r_2 \ldots t_N r_N$$

where each t_i is a combination of $s_i = a_{i,1} a_{i,2} \ldots a_{i,n}$ and v_i obtained as follows.

$$t_i = r_{i,0} \begin{pmatrix} s_{i,1} \\ a_{i,1} a_{i,1} \end{pmatrix} r_{i,1} \begin{pmatrix} s_{i,2} \\ a_{i,2} a_{i,2} \end{pmatrix} r_{i,2} \ldots \begin{pmatrix} s_{i,n} \\ a_{i,n} a_{i,n} \end{pmatrix} r_{i,n}$$

where

$$\begin{pmatrix} s_{i,j} \\ a_{i,j} a_{i,j} \end{pmatrix} = \begin{cases} a_{i,j} a_{i,j} s_{i,j} & \text{if } a_{i,j} \text{ is of the last type} \\ s_{i,j} a_{i,j} a_{i,j} & \text{if } a_{i,j} \text{ is of the first type.} \end{cases}$$

Putting $F_k(m+1) = nN$ we have to show that w really witnesses $S(k, m+1)$. The square blocks of w are the factors $s_{i,j}$ expanded by one occurrence of $a_{i,j}$ on the right or on the left. The round blocks of w are the factors r_i together with factors $r_{i,j}$ possibly expanded by one occurrence of $a_{i,j}$ on the left and $a_{i,j+1}$ on the right. Each block has been augmented by at most two different letters from a disjoint word, so the regularity of blocks has been preserved. Furthermore

$$\operatorname{rank} w = \operatorname{rank} v + \operatorname{rank} u = mnN + nN = (m+1)nN = (m+1)F_k(m+1)$$

as required.

(i) Suppose that w contains a subword *ababa*. One of the letters a and b must come from u, the other one from some v_i. If a comes from v_i and b from u, only one of the duplicated occurrences of b can be used. This is impossible, because they are the only occurrences of b between the elements of v_i. If a comes from u, and b from v_i then the middle a must come from the duplicated square occurrence. This cannot happen because in this case the middle a is the first or the last occurrence of a, or its neighbour.

(ii) The round blocks of w include the round blocks of u and the possibly augmented round blocks of v. The square blocks of w are the augmented square blocks of v. There are at most $3N-2$ blocks in u, $3n-2$ blocks in each v_i, $N(3n-2)$ blocks in v, and exactly N square blocks in u. The number of the blocks of w is then at least

$$3N - 2 + N(3n-2) - N = 3nN - 2 = 3F_k(m+1) - 2.$$

(iii) The number of occurrences of letters from v has not changed. The number of occurrences of each letter from u has increased by one because we doubled its unique square occurrence. Hence the number of occurrences of letters from v stays at least k, while the number of occurrences of letters from u goes up from at least $k-1$ to at least k.

(iv) The number of square blocks of w is the same as that of v. There are exactly n square blocks in every v_i, each of length m. Hence we get $Nn = F_k(m+1)$ square blocks of length m in v and $F_k(m+1)$ augmented square blocks of length $m+1$ in w. Each letter from v has exactly one square occurrence in w in the same block as in v and in the original word v_i. Round occurrences of letters from u stay round in w as well. One copy of the duplicated square occurrence of a letter was included in a square block of w, the other one in a round block. So the letters from u have unique square occurrences, too.

(v) The square blocks of w are separated by the same round blocks as they were in v. The augmentation of blocks and addition of new round blocks does not change this.

(vi) Letters from v stay in the same round blocks in w as they were in v. So (vi) holds for pairs from v. A pair containing a letter a from u and a letter b from v can only lie together in one round block of w, namely the unique block of v the letter a has been added to. Hence the only way that two round blocks of w could contain the same pair of elements is that the pair of elements lies in a round block from u and has been added to a round block r from v. This is, however, impossible because the letter added on the left was of the first type, and the letter added on the right of a round block was of the last type. Any round block other than r lies either right of the occurrence of the last type or left of the occurrence of the first type, and cannot contain both letters.

(vii) The square occurrences of letters from v have not changed; they are still of the first or of the last type. The square occurrence of the first type of a letter from u has been replaced by a square occurrence in w followed by a new round occurrence. Hence the square occurrence is still the first occurrence of the letter. The case of the square occurrence of the last type is symmetric.

Komjáth goes on to prove that $s(n)$ is $\Omega(n\alpha(n))$ for $f = ababa$. Since we are mainly interested in giving a characterization of forbidden words with maximum length functions in $\Theta(n)$, it suffices to show that $s(n)$ is not $O(n)$. We use the above construction to prove the following Lemma 2.

Definition 5 *A word w with* $\operatorname{rank} w = n$ *is block-regular if it can be decomposed into at most $3n - 2$ regular factors.*

Lemma 2 *Let $S(n)$ be the maximum length of a block-regular word w avoiding $f = ababa$ such that* $\operatorname{rank} w = n$. *Then $S(n)$ is not $O(n)$.*

Proof: Suppose there exists a constant k such that $S(n) < kn$ for all n. Let w be a word witnessing $S(k,1)$, and $n = F_k(1)$. Then w avoids *ababa*, and $\operatorname{rank} w = n$. The number of blocks is at most $3n-2$, so w is block-regular. Since every letter occurs at least k times, $|w| \geq kn$. This is a contradiction with $|w| \leq S(n)$.

Theorem 8 *The maximum length function $s(n)$ for $f = ababa$ is not $O(n)$.*

Proof: Suppose there exists a constant k such that $s(n) \leq kn$ for all n. By Lemma 2 there exists a number n and a block-regular word w with $\operatorname{rank} w = n$ such that $|w| \geq (k+3)n$. We form a word w' by applying $A(2)$ to w, in other words by removing immediate repetitions. We remove at most one letter from each block, so $|w'| \geq |w| - (3n-2) \geq (k+3)n - 3n + 2 = kn + 2$. This is a contradiction, because w' is 2-regular, avoids f, and therefore satisfies $|w'| \leq s(n)$.

We are now ready to state the main theorem.

Theorem 9 (main) *Let $f \in A^f$ be a forbidden word on two letters. The maximum length function is $\Theta(n)$ if and only if f avoids ababa.*

Proof: It follows from Observation 2 that if f does not avoid *ababa* then the maximum length function for f is at least as large as that for *ababa*, which is superlinear by Theorem 8. Combining this with Corollary 4 we get Theorem 9.

Chapter 5

Forbidden Words on Three Letters

Theorem 10 *Let $s(n)$ be the maximum length function for $f = abwab$, where $a, b \in A$, $w \in A^*$, and abw is a regular factor. Then $s(n)$ is $\Theta(n)$.*

Proof: Let $k = \operatorname{rank} f$, and v be a k-regular word avoiding f. Denote by x the letter that occurs last in v, that is the first occurrence of x in v is preceeded by an occurrence of every other letter. Suppose x occurs twice. Consider the first two occurrences of x. We can write $v = pxqxr$, where $p, q, r \in A^*$, p contains all letters from $\operatorname{Im} v$ except x, and q contains no x. Since v is k-regular, the two occurrences of x must be separated by a regular factor of length $k-1$. We can write this factor as uy, where $u \in A^*$ and $y \in A$. As y is different from x, it must appear in p. We have a subword $yxuyx$ in w. Since there is no x in u, the factor xuy is regular, and so is yxu. At the same time $\operatorname{rank} yxuyx = \operatorname{rank} yxu = k$, and therefore $yxuyx$ is isomorphic to f. This is a contradiction, so x only occurs once in w, and by Theorem 3 $s(n)$ is $\Theta(n)$.

Corollary 5 *The maximum length function is $\Theta(n)$ for $f = abcab$, where $a, b, c \in A$.*

Proof: Set $w = c$ in Theorem 10.

We know from Theorem 1 that $s(n) = O(n^3)$. Using Corollary 5 we can prove a stronger result.

Theorem 11 *Let $s(n)$ be the maximum length function for $f = abcabc$. Then $s(n) = O(n^2)$.*

Proof: We are going to prove the following claim, from which Theorem 11 follows immediately.

> Let $s(n)$ and $t(n)$ be the maximum length functions for *abcabc* and *abcab* respectively. Let K be a constant such that $t(n) \leq Kn$ for each n. Then $s(n) \leq Ln^2$ for each n, where $L = K + 1$.

The existence of K is guaranteed by Corollary 5, so this is enough. To prove the claim by induction consider a word w on n letters avoiding *abcabc*. If $n = 1$ then $|w| = 1$ and the proof is trivial. If $n > 1$, consider the letter x that occurs last in w. Then w can be written as uxv, where u contains all letters but x, and v avoids *abcab*. Indeed, if there were a word isomorphic to *abcab* in v, say $yztyz$, we could find $tyztyz$ in w because the factor ux contains all letters including t. Therefore $|v| \leq Kn$. At the same time u is a word on $n - 1$ letters avoiding *abcabc*, and by the induction hypothesis $u \leq L(n-1)^2$. Then

$$|w| \leq L(n-1)^2 + Kn + 1 = L(n-1)^2 + (L-1)n + 1 = Ln^2 - Ln + L - n + 1 = Ln^2 - (L+1)(n-1) \leq Ln^2$$

This holds for every w, so $s(n) \leq Ln^2$.

There is a reason to believe that the maximum length function s(n) is $\Theta(n)$ for the forbidden word $f = abcabc$ (see Chapter 7). Though the pattern discovered by the computer seems to be surprisingly regular and simple, we have not been able to find a proof of linearity. However, the following Lemma 3 might be helpful in finding the proof because it gives a similar insight into the behaviour of letters in a word w avoiding *abcabc*, as the lemma in the proof of Theorem 7 did for *abbaab*. In Theorem 7 we used the fact that two different letters could not intersect "too much". Here we are only able to prove that among every three letters there are two that do not intersect "too much". We cannot say anything about the way an arbitrary pair of letters intersect. In fact, the length of an alternating subsequence $ababa\ldots$ in a 3-regular word avoiding $abcabc$ can get arbitrarily large, as we can see in the following example:

$$w = abc_1abc_2\ldots abc_n$$

Obviously, the length of the alternating subsequence $ababa\ldots$ goes to infinity as n goes to infinity.

Definition 6 *Let $w \in A^*$. We define a partial order $\ll$ on $\operatorname{Im} w$ as follows:*

> *$a \ll b$ if all occurrences of a lie between two neighbouring occurrences of b or $a = b$*

We say that a and b are separated if all occurrences of a come before the first occurrence of b, or the other way round.

Lemma 3 *Let $f = abcabc$, a, b and c be three different letters from A and let w be a word avoiding f. Let $x, y, z \in \operatorname{Im} w$ be three different letters. Then there exists a pair of letters from among x, y and z that are separated or comparable in $\ll$.*

Proof: Consider two letters from among x, y and z, say x and z, that are neither separated nor comparable. There are two possibilities for their occurrences in w (without loss of generality x appears first in w):

$$w = \underbrace{x\ldots x}\,\underbrace{z\ldots z}\,\cdots\,\underbrace{z\ldots z} \qquad (5.1)$$

with at least four underbraced factors, and

$$w = \underbrace{x\ldots x}\,\underbrace{z\ldots z}\,\cdots\,\underbrace{x\ldots x} \qquad (5.2)$$

with at least five underbraced factors. First suppose w given by Equation 5.1. We will discuss the following possible cases of the first occurrence of y in w.

(1) The first y appears left of or in the first underbrace. There is no y in the second overbrace in the following expression else we would have a subword $yxzyxz$ in w.

$$\overbrace{\underbrace{x\ldots x}}\overbrace{\underbrace{z\ldots z}\,\cdots\,\underbrace{x\ldots x}}\overbrace{\underbrace{z\ldots z}}.$$

If there is no y in the first overbrace then x and y are separated or $x \ll y$. If there is a y in the first overbrace then there can be no y in the third overbrace else there would be a subword $xyzxyz$ in w. Hence z and y are separated or $z \ll y$.

(2) The first y appears left of or in the second underbrace and **(1)** does not hold. There can be no y in the second overbrace in the following expression else we would have $xyzxyz$ in w.

$$\underbrace{x\ldots x}\ \overbrace{\underbrace{z\ldots z}}\ \overbrace{\underbrace{x\ldots x}\ \cdots\ \underbrace{z\ldots z}}$$

If there is no y in the first overbrace then z and y are separated or $z \ll y$. If there is a y in the first overbrace then there can be no y right of the last overbrace else we would have a subword $xzyxzy$ in w. Hence $y \ll x$.

(3) The first y appears between the second and third underbrace. There can be no y in or right of the fourth underbrace or we would have a subword $xzyxzy$ in w. Hence all y appear in the overbrace and $y \ll z$.

$$\underbrace{x\ldots x}\ \underbrace{z\ldots z}\ \overbrace{\underbrace{x\ldots x}}\ \underbrace{z\ldots z}\ \cdots$$

(4) The first y is in or right of the third underbrace. There is a subword xzx left of the first y. If there are both x and z between the first and the last occurrence of y then we get either $xzyxzy$ or $zxyzxy$ in w. Hence at least one of x and z is either separated from or greater than y in $\ll$.

Let now w be given by Equation 5.2.

(1) The first y appears left of or in the first underbrace. There is no y in the first overbrace in the following expression else we would have a subword $yxzyxz$ in w. There is no y in the second overbrace either, or we would have $yzxyzx$.

$$\underbrace{x\ldots x}\ \overbrace{\underbrace{z\ldots z}\ \cdots\ \underbrace{x\ldots x}}\ \overbrace{\underbrace{z\ldots z}}\ \underbrace{x\ldots x}$$

Hence z and y are separated or $z \ll y$.

(**2**) The first y appears left of or in the second underbrace and (**1**) does not hold. We can consider the following expression and realize that the proof of case (**2**) above works without any changes.

$$\underbrace{x\ldots x}\,\overbrace{\underbrace{z\ldots z}}\,\overbrace{\underbrace{x\ldots x}\cdots\underbrace{z\ldots z}}\,\underbrace{x\ldots x}$$

The remaining two cases (**3**) and (**4**) are exactly the same as they were in the case of w given by Equation 5.1.

Chapter 6

Matrices

For a word $w = a_1 a_2 \ldots a_l$, $a_i \in A$, of length l we define a matrix $W = (w_{i,j})$ of the type $l \times l$ as follows.

$$w_{i,j} = \begin{cases} 1 & \text{if } a_i = a_j \\ 0 & \text{otherwise.} \end{cases}$$

Then W is the matrix of an equivalence relation. This means $u_{i,i} = 1$ and $u_{i,j} = u_{j,i}$ for all $i, j = 1, 2, \ldots, l$, and if D is a 2×2 submatrix with three nonzero entries then all four entries of D are nonzero. We will call matrices of this type equivalence matrices and W the equivalence matrix of w. There is a unique correspondence between words of length l and equivalence matrices of side l.

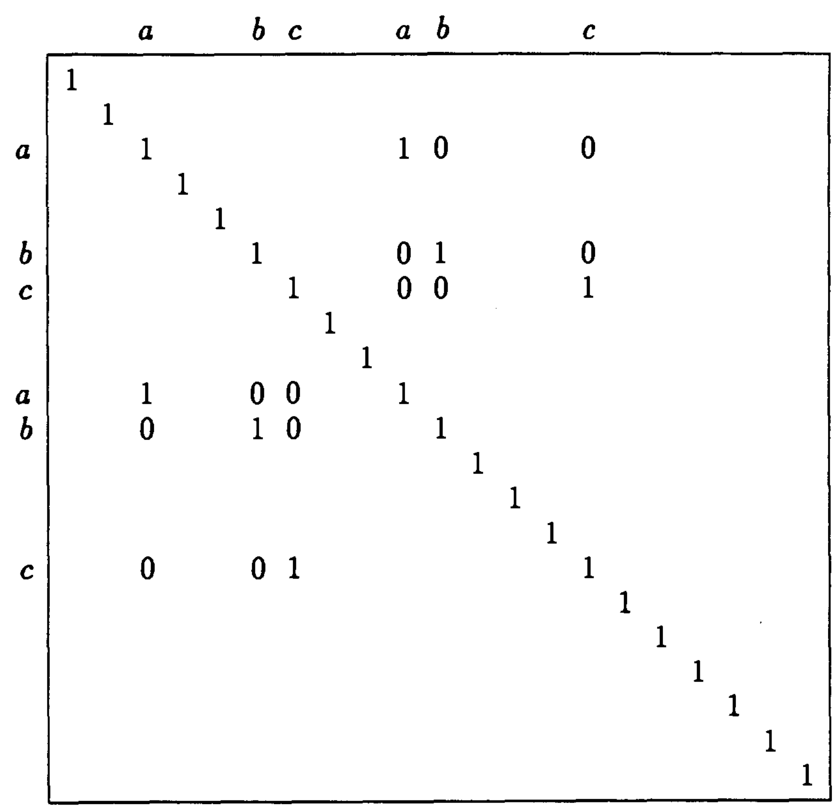
Definition 7 *A (0,1)-matrix is called k-sparse if every pair of two nonzero entries lying in the same row or column are separated by at least $k - 1$ zero entries.*

Observation 4 *A word w is k-regular if and only if its equivalence matrix is k-sparse.*

Proof: If w is not k-regular then there must exist two different positions i and j in w containing the same letter, such that $|j - i| < k$. Then $w_{i,j} = w_{i,i} = 1$ are two entries in the i^{th} row such that there are at most $|j - i| - 1 < k - 1$ cells between them, and there is no room for $k - 1$ zeros there. Now let W not be k-sparse. Since it is symmetric, we can without loss of generality assume that there exist two entries in the same row, say $w_{i,j}$ and $w_{i,m}$, that do not have $k - 1$ zeros between them. Take a pair with minimum $|j-m|$. There cannot be any nonzero entries between $w_{i,j}$ and $w_{i,m}$ because this would contradict minimality. Hence the number of cells between $w_{i,j}$ and $w_{i,m}$ is the same as the number of zeros, which is at most $k-2$. Then $|j-m| \leq k-1$. It follows from $w_{i,j}=1$ that $a_i = a_j$, and similarly $w_{i,m}=1$ implies $a_i = a_m$. Consequently $a_j = a_m$, $|j-m| < k$, and w is not k-regular.

We can associate the occurrence of forbidden subwords in w with the occurrence of forbidden submatrices in W.

Observation 5 *A 3-regular word w contains a subword isomorphic to $f = abcabc$ if and only if its equivalence matrix contains the identity matrix I_3 of side 3 on one side of the main diagonal.*



Remark 3 By saying "on one side of the main diagonal" we mean that for all $w_{i,j}$ lying on the identity submatrix the sign of $i-j$ is the same (i.e. either $i-j>0$ for all $w_{i,j}$, or $i-j<0$ for all $w_{i,j}$).

Proof: The situation is shown in the picture. Let $w = a_1a_2\dots a_l$, where $a_i \in A$. Let $a_ia_ja_ka_pa_qa_r \cong abcabc$, where $i<j<k<p<q<r$. Then in the equivalence matrix W we have

$$\begin{pmatrix} w_{i,p} & w_{i,q} & w_{i,r} \\ w_{j,p} & w_{j,q} & w_{j,r} \\ w_{k,p} & w_{k,q} & w_{k,r} \end{pmatrix} = I_3$$

and $k<p$, so all entries of I_3 lie in the upper right-hand corner of W. Conversely, let there be a submatrix I_3 on one side of the main diagonal of W. Owing to symmetry there is such a matrix in the upper right-hand corner of W. Denote its entries as in the identity matrix above. Then $k<p$, and the inequality $i<j<k<p<q<r$ holds. Consequently in the subword $a_ia_ja_ka_pa_qa_r$ of w we get $a_i=a_p$, $a_j=a_q$, $a_k=a_r$, and a_i, a_j, a_k are all different. Hence $a_ia_ja_ka_pa_qa_r \cong abcabc$, and we are done.

This means that the maximum length function $s(n)$ for the forbidden word $f = abcabc$ is the maximum side of a 3-sparse equivalence matrix not containing I_3 on either side of the main diagonal.

Füredi and Hajnal employ a slightly different approach in [8]. Instead of forbidden submatrices they use forbidden configurations.

Definition 8

- *A configuration $C=(c_{i,j})$ of the type $u \times v$ is a nonempty partial matrix with 1's and blanks for entries with no blank rows and columns.*

- *A (0,1)-matrix $M=(m_{i,j})$ contains the configuration C if there exists a submatrix $B=(b_{i,j})$ of M of the type $u \times v$ such that $c_{i,j}=1$ implies $b_{i,j}=1$ for all i and j.*

Definition 9

- *For a (0,1)-matrix M we denote by $|M|$ the number of nonzero entries in M.*
- *In the case of matrices, $s(n)$ denotes the maximum $|M|$ of an $n \times n$ matrix M avoiding a given forbidden configuration F. We retain the name maximum length function for $s(n)$.*

Theorem 12 *The maximum length function $s(n)$ is $O(n\alpha(n))$ for the forbidden configuration*

$$F = \begin{pmatrix} 1 & & 1 & \\ & 1 & & 1 \end{pmatrix}.$$

Proof: Let $M = (m_{i,j})$ be an $n \times n$ (0,1)-matrix avoiding F such that $|M| = s(n)$. We replace the first and the last nonzero entry in each row of M by zero and keep only the columns with at least two nonzero entries to get a matrix M'. We removed at most $2n$ ones in the first step and at most n ones in the second step, hence $|M'| \geq |M| - 3n$. We pick a number j from $\{1, 2, \ldots, n\}$ and consider the j^{th} column of M. Let $i_1 < i_2 < \ldots < i_l$ be the subscripts of nonzero entries in the j^{th} column. Formally $m_{i_1,j} = m_{i_2,j} = \ldots = m_{i_l,j} = 1$ and $m_{i,j} = 0$ for any other value of i. Define a word w_j on n letters as

$$w_j = i_1 i_2 \ldots i_l .$$

Obviously $l \geq 2$ because we left out the columns with less than two nonzero entries. Repeating this for each $j = 1, 2, \ldots, n$ we get words $w_1, w_2, \ldots w_n$ on the letter set $\{1, 2, \ldots, n\}$. (Here we assume that the integers we have used are all in A. We can easily avoid formal inconsistency by renaming them conveniently.) Finally we put

$$w' = w_1 w_2 \ldots w_n$$

and applly $A(2)$ to w' to obtain a word w. In other words we remove each immediate repetition from w' leaving a single occurrence of the letter instead. We claim that w avoids $ababa$. Assume the contrary. Then there exists a subword $g = xyxy$ in w such that $x < y$. We must have g in w' as well. The second and the third letters of g lie in two different factors w_j of w' because each of these factors is increasing. Now there are four possibilities for the occurrences of x and y in w', each corresponding to a configuration that must have occured in M' to yield it.

- The first two letters of g come from the same factor; the last two letters do not. The configuration is
$$\begin{pmatrix} 1 & 1 & \\ 1 & & 1 \end{pmatrix}.$$
- The last two letters of g come from the same factor; the first two letters do not. The configuration is
$$\begin{pmatrix} 1 & & 1 \\ & 1 & 1 \end{pmatrix}.$$
- Both the first and the last two letters of g come from the same factor. The configuration is
$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$
- The letters of g come from four different factors. The configuration is
$$\begin{pmatrix} 1 & & 1 & \\ & 1 & & 1 \end{pmatrix}.$$

In every case adding 1 at the beginning and the end of each row yields the forbidden configuration F, so F must have been contained in M. This is a contradiction. Denoting the maximum length function for *ababa* by $t(n)$ we get $|w| \leq t(n)$. The algorithm $A(2)$ left out at most one letter from each factor w_j because w_j is increasing, and $|w_j| \geq 2$ for all j. Hence $|w'| \leq t(n) + n$. This means $|M'| \leq t(n) + n$, and $s(n) = |M| \leq t(n) + 4n$. Since $t(n)$ is known to be $O(n\alpha(n))$, we get $s(n) = O(n\alpha(n))$ as well.

Füredi and Hajnal also show that the obtained upper bound is tight using the same double induction we used in Chapter 4 to prove superlinearity for *ababa*. They construct matrices of side n with $\Omega(n\alpha(n))$ ones for infinitely many values of n. This does not mean that $s(n)$ is $\Omega(n\alpha(n))$ in the sense of Definition 2. The correct interpretation of their result is that $s(n)$ is not $o(n\alpha(n))$. Their construction, referring to [3], can be found in[8].

Definition 10 *Let C be a configuration. The following two operations on C are called the elementary operations.*

- *deleting an entry*
- *attaching a new column or row to the boundary of C with exactly one entry* 1 *next to an existing one in C*

If there is a blank row or column in C after the deletion of an entry, we leave them out, too.

Definition 11 *We write $C \rightarrow D$ if D can be obtained from C by a finite sequence of elementary operations.*

Theorem 13 *Let C and D be configurations such that $C \rightarrow D$ in k steps, and $s(n)$ and $t(n)$ be the maximum length functions for C and D respectively. Then $t(n) \leq s(n) + kn$.*

Proof: It is enough to prove the statement for only one step. If the step is deleting an entry, the statement is trivial because a matrix avoiding D avoids C as well. If the step is adding a new column to the right, we consider an $n \times n$ (0,1)-matrix M with $|M| = t(n)$. We replace the last occurrence of 1 in each nonzero row of M by 0 to get a matrix M'. If M' contained C, then we could find C in corresponding cells of M and extend it to D by adding one of the nonzero entries that had been left out. Hence M' avoids C, and $|M'| \leq s(n)$. We left out at most n 1's from M to obtain M', so

$$t(n) = |M| \leq |M'| + n \leq s(n) + n\,.$$

The case of adding a new column to the left or adding a new row is virtually the same. Now, Theorem 13 follows easily by induction.

Observation 6 *If $F \neq (1)$ then the maximum length function $s(n)$ for F satisfies $s(n) \geq n$ for all n.*

Proof: If F has at least two columns then it is always possible to construct an $n \times n$ matrix with n nonzero entries avoiding F. Just set all the entries in one column equal to 1 and all the other entries equal to 0. If F has at least two rows, the construction is symmetric. The only case we did not cover is $F = (1)$ when $s(n) = 0$ for all n.

Corollary 6 *Let $s(n)$ and $t(n)$ be the maximum length functions for two configurations C and D respectively, and $C \neq (1)$.*

- *If $C \to D$ then $t(n) = O(s(n))$.*
- *If D can be obtained from C by elementary operations of only the second type, then $s(n)$ and $t(n)$ are of the same order.*

Proof: If $C \to D$ then $t(n) \leq s(n) + kn$ by Theorem 13. It follows from $s(n) \geq n$ that $kn = O(s(n))$, $s(n) + kn = O(s(n))$, and finally $t(n) = O(s(n))$. If we only use elementary operations of the second type to derive D, then of course $D \neq (1)$, and $C \to D$. Hence trivially $t(n) = O(s(n))$. At the same time we can get back from D to C by deletions of the added 1's. Therefore $D \to C$, $s(n) = O(t(n))$, and $s(n)$ and $t(n)$ are of the same order.

Corollary 7 *The maximum length function $s(n)$ is $O(n\alpha(n))$ for the following configurations.*

$$\begin{pmatrix} & & 1 & \\ 1 & & & \\ & 1 & & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & & & \\ & & 1 & \\ & 1 & & 1 \end{pmatrix}$$

$$\begin{pmatrix} & & 1 & \\ 1 & & & \\ & & & 1 \\ & 1 & & \end{pmatrix} \qquad \begin{pmatrix} 1 & & & \\ & & 1 & \\ & 1 & & \\ & & & 1 \end{pmatrix}$$

Proof: Each of the above four matrices can be easily derived by elementary operations from the forbidden configuration F used in Theorem 12.

Chapter 7

Computer Searches

A computer search has been conducted to find the maximum length function for $f = abcabc$ for $n \leq 7$. The search also generated the first maximum sequence in the lexicographic order. The results follow.

n	$s(n)$	*sequence*
2	2	1 2
3	5	1 2 3 1 2
4	10	1 2 3 1 4 3 2 4 1 2
5	13	1 2 3 1 2 4 1 5 4 2 5 1 2
6	18	1 2 3 1 4 3 2 4 1 2 5 1 6 5 2 6 1 2
7	21	1 2 3 1 2 4 1 5 4 2 5 1 2 6 1 7 6 2 7 1 2

The pattern in the table suggests that the maximum sequence on $n + 2$ letters can be obtained from the sequence on n letters by juxtaposing a sequence of the type 61762712 where 7 and 6 stand for n and $n - 1$. If this is true then $s(n) = 4n - 6$ for n even and $s(n) = 4n - 7$ for n odd greater than 1. The pattern is convincing enough to enable us to formulate a conjecture.

Conjecture 1 *The maximum length function for $f = abcabc$ is $\Theta(n)$.*

What adds plausibility to this conjecture is that all forbidden words that are known to be $\Theta(n)$ show similar regularity while $f = ababa$, which is known to be superlinear, does not. For $f = abbaab$ we get

n	$s(n)$	*sequence*
2	7	1 2 1 2 1 2 1
3	13	1 2 1 2 1 2 1 3 1 3 1 3 1
4	19	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1

For $f = abcab$ we get

n	$s(n)$	*sequence*
2	2	1 2
3	4	1 2 3 1
4	6	1 2 3 4 2 1
5	8	1 2 3 4 5 3 2 1
6	10	1 2 3 4 5 6 4 3 2 1
7	12	1 2 3 4 5 6 7 5 4 3 2 1

However, the following table for $f = ababa$ lacks symmetry.

n	$s(n)$	*sequence*
2	4	1 2 1 2
3	8	1 2 1 3 1 3 2 3
4	12	1 2 1 3 1 3 2 4 2 4 3 4
5	17	1 2 1 3 1 4 1 4 3 2 5 2 5 3 5 4 5

For $f = abab$ we can even prove easily that the upper bound we conjecture is correct and tight. The table is

n	$s(n)$	*sequence*
2	3	1 2 1
3	5	1 2 1 3 1
4	7	1 2 1 3 1 4 1
5	9	1 2 1 3 1 4 1 5 1

Conjecture 2 *For $f = abab$ the maximum length function is $s(n) = 2n - 1$.*

Proof: Obviously it is always possible to construct a word avoiding f of length $2n-1$ extending the pattern. Now we must prove that $2n-1$ really is an upper bound. The case of $n = 1$ is trivial. Let $n > 1$ and w be a word on n letters avoiding f. Denote the first letter by a. If a never appears again then by leaving out its only occurrence from w we get a word w' on $n-1$ letters. By the induction hypothesis $|w'| \leq 2(n-1)-1$ and $|w| \leq 2n-1$. If a does appear again then w can be written as $auav$ where u and v are disjoint words. Let $\operatorname{rank} u = t$. Then $\operatorname{rank} av \leq n-t$, $|av| \leq 2(n-t)-1$, $|u| \leq 2t-1$ and $|w| = 1 + |u| + |av| \leq 2n-1$.

Bibliography

[1] H. Davenport and A. Schinzel, *A Combinatorial Problem Connected with Differential Equations,* Amer J. Math. 86 (1965), 684–694.

[2] H. Davenport, *A Combinatorial Problem Connected with Differential Equations, II,* Acta Arith. 17 (1971), 363–372

[3] S. Hart and M. Sharir, *Nonlinearity of Davenport-Schinzel Sequences and of Generalized Path Compression Schemes,* Combinatorica 6 (1986), 151–177.

[4] M. Sharir, *Almost Linear Upper Bounds on the Length of General Davenport-Schinzel Sequences,* Combinatorica 7 (1987), 131–143.

[5] A. Wiernik, *Planar Realizations of Nonlinear Davenport-Schinzel Sequences by Segments,* in Proceedings, 27th IEEE Found. of Comput. Sci., (1986), 97-106.

[6] E. Szemerédi, *On a Problem of Davenport and Schinzel,* Acta Arith. 25 (1974), 213–224.

[7] P. Komjáth, *A Simplified Construction of Nonlinear Davenport-Schinzel Sequences,* J. of Comb. Theory, Series A 49 (1988), 262–267.

[8] Z. Füredi and P. Hajnal, *Davenport-Schinzel Theory of Matrices,* to appear.

[9] R. Adamec, M. Klazar and P. Valtr, *Forbidden Words,* KAM Series No. 155-89 (Charles University, Prague, Czechoslovakia, 1989).