MATHEMATICAL PROBLEMS IN COMPUTATION THEORY
BANACH CENTER PUBLICATIONS, VOLUME 21
PWN – POLISH SCIENTIFIC PUBLISHERS
WARSAW 1988

SQUARE-FREE AND OVERLAP-FREE WORDS

A. J. KFOURY

Department of Computer Science, Boston University, U.S.A.

Words not containing repeated subwords (twice or more times) have been studied in formal language theory. They also have various applications in mathematical games. Our interest in them is the result of recent work in logics of programs.

In [3] and [8], infinite words not containing overlaps are used to settle various open problems concerning the "unwind property" and first-order logics of programs. A close examination of [6] and [7] will show that their proof technique also depends on the existence of infinite words not containing repeated subwords.

Among other results in this report, we show that there are uncountably many doubly-infinite words over a binary alphabet which are overlap-free (Theorem 2.6), from which it easily follows that there are uncountably many doubly-infinite words over a ternary alphabet which are square-free (Corollary 2.7). We also show that there is a very simple $\mathcal{O}(n \log n)$ algorithm to test whether a word of length n over a binary alphabet is overlap-free (Theorem 3.10), from which we derive a (low) polynomial-in-n bound on the number of overlap-free words of length n over a binary alphabet (Theorem 3.11).

Several of the results presented below are already known, or closely related to results that are already known. Crochemore has developed a $\mathcal{O}(n)$ algorithm to test whether a word of length n is square-free [2]; this algorithm can be adapted to test whether a word of length n is overlap-free. However, our asymptotically slower $\mathcal{O}(n \log n)$ algorithm has a simple description (see subsection 3.9 below); and, moreover, it allows us to determine a polynomial bound on the number of overlap-free words over a binary alphabet which is tighter than the previously known bound (ours is $\mathcal{O}(n^e)$ for some $e \leqslant 2.8$ whereas Restivo's and Salemi's is $\mathcal{O}(n^{\log 15})$, mentioned in [1]). It is also known that there are uncountably many (singly) infinite overlap-free words over a binary alphabet, and uncountably many (singly) infinite square-free words over a ternary alphabet (see Problems 2.2.3, 2.3.6, and 2.3.7, in

[4]): both of these results are immediate consequences of our Theorem 2.6 and Corollary 2.7.

Acknowledgements. The questions examined in this report were raised in a NSF research proposal, jointly written by J. Tiuryn, P. Urzyczyn, and myself. Jean Berstel guided me through the recent literature on square-free and overlap-free words. Tom Orowan was responsible for putting this report in its final typewritten form.

1. Notation and preliminary results

We generally follow the notation and terminology of Chapter 1 of [5]. In addition, we reserve early Greek letters (α, β and γ) for specific words we shall define in the course of our presentation, and late Greek letters (π, ϱ, σ and τ) for variables ranging over the set of possible words. Late Roman letters (x, y and z) will stand for symbols from finite alphabets.

ω is the order type of the natural numbers, and $\omega^* + \omega$ that of the integers. $\boldsymbol{Z}$ denotes the set of integers. An ω-word (or a right-infinite word) over a finite alphabet Σ is a member of Σ^ω; an ω^*-word (or a left-infinite word) over Σ is a member of Σ^{ω^*}.

A word σ contains a *square* if it contains a finite subword of the form $\tau\tau$, where τ is non-empty. A word σ contains an *overlap* if it contains a finite subword of the form $\varrho\tau\varrho'$ such that $\varrho\tau = \tau\varrho'$, where ϱ, τ, and ϱ' are non-empty. Square-freeness implies overlap-freeness, but not the other way around. No word of length $\geqslant 4$ over a binary alphabet can be square-free; on the other hand, there are overlap-free words of unbounded length over a binary alphabet.

If σ is a word over a binary alphabet, $\bar{\sigma}$ denotes the complement of σ, i.e., the word obtained from σ by replacing every 0 by 1 and every 1 by 0.

1.1. DEFINITION. We define an infinite sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$, of words inductively:

$$\alpha_0 = 0 \quad \text{and} \quad \bar{\alpha}_0 = 1, \text{ and for all } n \geqslant 0, \quad \alpha_{n+1} = \alpha_n \bar{\alpha}_n \quad \text{and} \quad \bar{\alpha}_{n+1} = \bar{\alpha}_n \alpha_n,$$

α_n is a prefix of α_{n+1} for every n. The limit of the sequence $(\alpha_n | n \geqslant 0)$ is therefore well-defined: we denote the resulting ω-word by α_ω.

1.2. LEMMA. *α_ω is overlap-free (and, therefore, so are α_n and $\alpha_n\alpha_n$ for every $n \geqslant 0$).*

Proof. One proof is given in Chapter 1 of [5] ∎

The following is a useful characterization of overlap-freeness we shall use repeatedly in the sequel.

1.3. LEMMA. *A word σ is overlap-free $\Leftrightarrow$ σ does not contain a finite subword of the form $x\tau x\tau x \Leftrightarrow \sigma$ does not contain a finite subword of the form $\tau\tau x$ (or $x\tau\tau$) where τ is non-empty and x is the first (or last) symbol of τ.*

Proof. Straightforward. ∎

2. Infinite overlap-free words

We start with a lemma we shall use again in Section 3.

2.1. LEMMA. *Let $\sigma = x_1 x_2 \dots x_n \in \{0, 1\}^*$ be an overlap-free word of length $n \geqslant 5$.*

(1) *If $x_1 x_2 x_3 x_4 \in \{0110, 1001, 1010, 0101\}$ of if $x_1 x_2 x_3 x_4 x_5 \in \{00100, 11011\}$, then for all even j, $4 \leqslant j < n$, $x_j = \bar{x}_{j-1}$.*

(2) *If $x_1 x_2 x_3 x_4 \in \{0011, 1100, 0100, 1011\}$ or if $x_1 x_2 x_3 x_4 x_5 \in \{00101, 11010\}$, then for all odd j, $3 \leqslant j < n$, $x_j = \bar{x}_{j-1}$.*

It is easy to check that all prefixes $x_1 x_2 x_3 x_4$ not mentioned in (1) *and* (2) *contain an overlap.*

Proof. (1) By induction on even $j \geqslant 4$. The result is clear for $j = 4$. Let then k be an even integer, $4 < k < n$, and assume the result is already proved for every even j, $4 \leqslant j < k$. With no loss of generality, suppose $x_{k-3} x_{k-2} = 01$. This forces $x_{k-1} x_k \neq 11$, otherwise $x_{k-2} x_{k-1} x_k$ would be an overlap. We next show that $x_{k-1} x_k \neq 00$.

By the induction hypothesis, $x_{k-5} x_{k-4} = 01$ or 10. (In case $k = 6$, $x_{k-5} x_{k-4}$ may also be 11, but then in this case $x_{k-1} x_k = 10$.) If $x_{k-5} x_{k-4} = 01$ and $x_{k-1} x_k = 00$, then $x_{k-5} \dots x_{k-1}$ would be an overlap. If $x_{k-5} x_{k-4} = 10$ and $x_{k-1} x_k = 00$ (the latter condition forcing $x_{k+1} = 1$), then $x_{k-5} \dots x_{k+1}$ would be an overlap.

(2) By induction on odd $j \geqslant 3$. The proof is similar to (1) above, and we omit it. ∎

The following result is a useful characterization of doubly infinite words that are overlap-free.

2.2. LEMMA. *$\sigma \in \{0, 1\}^{\omega^* + \omega}$ is overlap-free $\Leftrightarrow$ for every $n \geqslant 0$, $\sigma \in \{\alpha_n, \bar{\alpha}_n\}^{\omega^* + \omega}$.*

Proof. ($\Rightarrow$) This follows from Lemma 2.1. Consider first the case $n = 1$. It is easy to see that, among the 12 patterns of length 4 or 5 mentioned in Lemma 2.1, neither 00100 nor 11011 can appear as a subword of a double-infinite overlap-free word. The occurrence in σ of any of the remaining 10 patterns implies that the infinite suffix (respectively, the infinite prefix) starting with the first or second symbol of this pattern can be viewed as a word in $\{01, 10\}^\omega$ (respectively, $\{01, 10\}^{\omega^*}$). Hence, $\sigma \in \{01, 10\}^{\omega^* + \omega} = \{\alpha_1, \bar{\alpha}_1\}^{\omega^* + \omega}$.

As a doubly-infinite word over the alphabet $\{\alpha_1, \bar{\alpha}_1\}$, σ must be overlap-free, otherwise it would not be so, as a word over $\{0, 1\}$. By the reasoning used above, we conclude again that $\sigma \in \{\alpha_1 \bar{\alpha}_1, \bar{\alpha}_1 \alpha_1\}^{\omega^* + \omega} = \{\alpha_2, \bar{\alpha}_2\}^{\omega^* + \omega}$. Proceeding inductively, we obtain the desired result for every $n \geqslant 0$.

($\Leftarrow$) Suppose σ is not overlap-free and that for every $n \geqslant 0$, $\sigma \in \{\alpha_n, \bar{\alpha}_n\}^{\omega^* + \omega}$, and we shall get a contradiction. If σ contains an overlap τ, then τ must appear as a subword in one of the following: α_k, $\bar{\alpha}_k$, $\alpha_k \alpha_k$, or $\bar{\alpha}_k \bar{\alpha}_k$, for some sufficiently large k. But α_k, $\bar{\alpha}_k$, $\alpha_k \alpha_k$ and $\bar{\alpha}_k \bar{\alpha}_k$ are all overlap-free. ∎

2.3. Corollary. *There are exactly* 4 *doubly-infinite overlap-free words over* $\{0, 1\}$ *which contain* α_ω *or* $\bar{\alpha}_\omega$ *as an infinite suffix. These are*:

$$\beta_1 = \alpha_\omega^R \alpha_\omega, \quad \beta_2 = \alpha_\omega^R \bar{\alpha}_\omega, \quad \beta_3 = \bar{\alpha}_\omega^R \alpha_\omega, \quad \beta_4 = \bar{\alpha}_\omega^R \bar{\alpha}_\omega,$$

where α_ω^R *denotes the reverse of* α_ω *(i.e.,* α_ω *written backwards).*

Proof. This easily follows from the left-to-right implication in 2.2, and is omitted. ∎

For the analysis to follow we need infinite binary sequences, which we call *profiles*. We thus call any member of $\{+, -\}^\omega$ a *profile*. (We use $\{+, -\}$ instead of $\{0, 1\}$ for notational convenience.)

Suppose $\sigma \in \{0, 1\}^{\omega^* + \omega}$. We think of σ as a doubly-infinite sequence of symbols indexed with the integers, $\sigma = (x_i | \ i \in \mathbf{Z})$. If σ is overlap-free, then for every $n \geqslant 0$, σ may be viewed as a word over $\{\alpha_n, \bar{\alpha}_n\}$, by the preceding lemma. The profile of σ at i, denoted PROFILE (σ, i) will give the relative position of symbol x_i in α_n or $\bar{\alpha}_n$, for every $n \geqslant 1$, when σ is written as a sequence in $\{\alpha_n, \bar{\alpha}_n\}^{\omega^* + \omega}$.

The following is a precise definition of PROFILE$(\sigma, i) = y_1 y_2 y_3 \cdots$ For every $n \geqslant 1$, write σ as a sequence in $\{\alpha_n, \bar{\alpha}_n\}^{\omega^* + \omega}$. If x_i occurs in the first half of α_n or $\bar{\alpha}_n$, then $y_n = +$; if x_i occurs in the second half of α_n or $\bar{\alpha}_n$, then $y_n = -$.

PROFILE(σ, i) is uniquely defined because when σ is written as a sequence σ' over $\{\alpha_n, \bar{\alpha}_n\}$, $n \geqslant 1$, there is a unique way of doing it. Indeed, using the fact that σ' is overlap-free, it is easy to see that no occurrence of α_n or $\bar{\alpha}_n$ can be of the form $\pi \varrho \pi'$, where $\varrho \in \{\alpha_{n-1}, \bar{\alpha}_{n-1}\}$ while π and π' are both non-empty.

Suppose, for example, we consider x_i in σ and the few symbols preceding and following x_i say:

$$\begin{array}{cccccccccccccc} \cdots & x_{i-2} & x_{i-1} & x_i & x_{i+1} & x_{i+2} & x_{i+3} & x_{i+4} & x_{i+5} & x_{i+6} & x_{i+7} & x_{i+8} & x_{i+9} \\ \cdots & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\begin{array}{ccccc} x_{i+10} & x_{i+11} & x_{i+12} & x_{i+13} & \cdots \\ 1 & 0 & 0 & 1 & \cdots \end{array}$$

then PROFILE$(\sigma, i) = + - + + \ldots$, because $x_i x_{i+1} = \alpha_1$, $x_{i-2} x_{i-1} x_i x_{i+1} = \bar{\alpha}_2$, $x_{i-2} \ldots x_{i+5} = \alpha_3$, $x_{i-2} \ldots x_{i+13} = \alpha_4$, etc.

A profile p in $\{+, -\}^\omega$ *stabilizes to* $+$ (respectively, $-$) if all but finitely many of its symbols are $+$ (respectively, $-$).

2.4. LEMMA. *Let $p \in \{+, -\}^\omega$ be an arbitrary profile.*

(1) *If p stabilizes to $+$ (respectively, $-$), there is a unique non-negative (respectively, negative) integer k such that $p =$ PROFILE (β_1, k), where the symbols of the suffix α_ω of β_1 are indexed with the non-negative integers and those of the prefix α_ω^R with the negative integers.*

(2) *If p does not stabilize to either $+$ or $-$, there is a doubly-infinite overlap-free word $\sigma = (x_i | \ i \in \mathbf{Z})$ with $x_i \in \{0, 1\}$ and an integer k such that $p =$ PROFILE(σ, k). σ is unique up to the isomorphism $\sigma \mapsto \bar{\sigma}$ (which replaces 0 by 1 and 1 by 0).*

Proof. We only prove part (2), the proof for (1) is similar. In (2) we let $k = 0$, and we construct the desired $\sigma = (x_i | \ i \in \mathbf{Z})$ in stages. The segment of σ constructed at *Stage* $n \geqslant 0$ will have length 2^n. *Stage* 0: Let $x_0 = 0$. *Stage* $n+1$: Suppose the word constructed at *Stage* n is $x_i x_{i+1} \ldots x_j$, where $i \leqslant 0 \leqslant j$ and $1 + j - i = 2^n$. If the $(n+1)$st symbol of p is $+$, we construct

$$x_i \ldots x_j x_{j+1} \ldots x_{j+1+j-i},$$

where $x_{j+1} \ldots x_{j+1+j+i} = \bar{x}_i \ldots \bar{x}_j$. If the $(n+1)$st symbol of p is $-$, we construct

$$x_{i-1-j+i} \ldots x_{i-1} x_i \ldots x_j,$$

where $x_{i-1-j+i} \ldots x_{i-1} = \bar{x}_i \ldots \bar{x}_j$. It is straightforward to check that the resulting σ is unique up to the renaming of the alphabet symbols 0 and 1. ∎

Let $p, q \in \{+, -\}^\omega$. p and q are *equivalent*, in symbols $p \approx q$, if as functions from the natural numbers to $\{+, -\}$ they agree almost everywhere. In particular, all profiles that stabilize to $+$ are equivalent to each other, and likewise for all profiles that stabilize to $-$.

2.5. LEMMA. *Let $p, q \in \{+, -\}^\omega$ two profiles which do not stabilize to either $+$ or $-$. Then $p \approx q \Leftrightarrow$ there is a doubly-infinite overlap-free word $\sigma = (x_i | \ i \in \mathbf{Z})$, $x_i \in \{0, 1\}$, and integers j and k such that $p =$ PROFILE(σ, j) and $q =$ PROFILE(σ, k).*

Proof. ($\Rightarrow$) The proof is similar to that of Lemma 2.4. Increasingly larger segments of the desired σ are constructed in stages. If p and q disagree only up to their nth symbol (inclusive), then in *Stage* 0, *Stage* 1, ..., and *Stage* $n-1$, two separate segments of σ are constructed. At *Stage* n, the two segments each of length 2^{n-1} are merged into one of length 2^n. After *Stage* n, only one segment is constructed at every stage. The details are omitted.

($\Leftarrow$) If $p =$ PROFILE(σ, j) and $q =$ PROFILE(σ, k) do not stabilize to

either + or −, then for a sufficiently large n, x_j and x_k will appear in the same α_n or $\bar{\alpha}_n$ (when σ is written as sequence over $\{\alpha_n, \bar{\alpha}_n\}$). ∎

The right-to-left implication in 2.5 maps every indexed doubly-infinite overlap-free word $\sigma = (x_i | \ i \in \mathbf{Z})$ to an equivalence class of $\approx$. Suppose we have another indexing of the same word, i.e., $\sigma' = (y_i | \ i \in \mathbf{Z})$ and there is a fixed j such that $y_i = x_{i+j}$ for all i. It is easily seen that σ and σ' are mapped to the same equivalence class of $\approx$. In the next proof we shall therefore ignore the indexing of doubly-infinite overlap-free words.

2.6. THEOREM. *There are uncountably many overlap-free words in* $\{0, 1\}^{\omega^* + \omega}$.

Proof. Consider the set $A \subset \{+, -\}^{\omega}$ of all profiles which do not stabilize to either + or −. We denote by $A/\approx$ the set A modulo the equivalence relation $\approx$. Let B be the set of all doubly-infinite overlap-free words over $\{0, 1\}$ excluding the words β_1, β_2, β_3, and β_4 of Corollary 2.3; and define the equivalence relation $\sim$ on B by: $\sigma \sim \tau$ iff $\tau = \bar{\sigma}$. Each equivalence class in $B/\sim$ contains exactly two words. By Lemma 2.4 and 2.5, there is a bijection between $A/\approx$ and $B/\sim$. It is easy to verify that $A/\approx$ is uncountable, which in turn implies that $B/\sim$ is uncountable. ∎

2.7. COROLLARY. *There are uncountably many square-free words in* $\{0, 1, 2\}^{\omega^* + \omega}$.

Proof. As in the proof of 2.6, let B be the set of all doubly-infinite overlap-free words over $\{0, 1\}$ excluding the words β_1, β_2, β_3 and β_4. Let C be the set off all doubly-infinite square-free words over $\{0, 1, 2\}$. We define an injection f from B to C, from which the desired conclusion will follow.

Suppose $\sigma \in B$. Note that σ does not contain subwords of the form 000 nor of the form 111. $f(\sigma)$ is obtained by changing every subword 00 to 02, and every subword 11 to 12. It is easy to see that f is a one-one map. That $f(\sigma) \in C$ follows from: $f(\sigma)$ not square-free $\Rightarrow$ σ not overlap-free, which is easily established by considering the three possible cases of a square $\tau\tau$ in $f(\sigma)$, namely, τ may start with a 0, or a 1, or a 2. Details omitted. ∎

From the previous results it immediately follows that there are uncountably many overlap-free words in $\{0, 1\}^{\omega}$, and uncountably many square-free words in $\{0, 1, 2\}^{\omega}$.

3. Finite overlap-free words

The following is a restatement of Lemma 2.1, more convenient for the analysis of this section.

3.1. LEMMA. *If* $\sigma \in \{0, 1\}^*$ *is overlap-free of length* ≥ 7, *there is a unique way of decomposing* σ *into three parts:* $\sigma = \pi\varrho\pi'$ *where* $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$ *and* $\varrho \in \{01, 10\}^+$. ∎

The π and π' in 3.1 are determined according to the 12 patterns mentioned in Lemma 2.1. Clearly, once π is determined, so is π'. If σ has as a prefix one of the first four patterns in 2.1.1:

$$0110,\ 1001,\ 1010,\ 0101,$$

then $\pi = \lambda$; if σ has a prefix one of the remaining two patterns in 2.1.1:

$$00100,\ 11011,$$

then $\pi = 00$ or 11; and if σ has as a prefix one of the six patterns in 2.1.2:

$$0011,\ 1100,\ 0100,\ 1011,\ 00101,\ 11010,$$

then $\pi = 0$ or 1.

For the case of overlap-free words σ of length $\leqslant 6$ not in $\{0, 1, 00, 11\}$, a decomposition of σ in the form prescribed by Lemma 3.1 is always possible, although it may not be unique (e.g., let $\sigma = 001011$ which admits two such decompositions, according to whether $\pi = 0$ or $\pi = 00$).

A unique decomposition in the form prescribed by Lemma 3.1 is sometimes possible for words σ that are not overlap-free (for example, let $\sigma = 001100110$), and this is the reason why algorithm $\mathscr{A}$ below may terminate successfully even when the input word σ contains an overlap.

3.2. LEMMA. *Let* $\varrho \in \{01, 10\}^+$ *and* ϱ' *be obtained from* ϱ *by mapping consecutive occurrences of* 01 *and* 10 *into* 0 *and* 1, *respectively. Then* ϱ *is overlap-free* $\Leftrightarrow$ ϱ' *is overlap-free.*

Proof. Straightforward. ∎

3.3. ALGORITHM $\mathscr{A}$. We would like to develop an algorithm to test whether an arbitrary $\sigma \in \{0, 1\}^+$ is overlap-free, based on the following strategy. At the first iteration we set $\sigma_1 = \sigma$. At the nth iteration, $n \geqslant 1$, we carry out the following steps:

1. If $\sigma_n \in \{0, 1, 00, 11\}$, terminate successfully.

2. Decompose σ_n as $\pi_n \varrho_n \pi_n'$, with $\pi_n, \pi_n' \in \{\lambda, 0, 1, 00, 11\}$ and $\varrho_n \in \{01, 10\}^+$. If this is not possible, terminate unsuccessfully. If σ_n has more than one such decomposition, take π_n as short as possible (e.g., if $\sigma_n = 001011$, take $\pi_n = 0$ rather than $\pi_n = 00$).

3. Define σ_{n+1} from ϱ_n by mapping consecutive occurrences of 01 and 10 into 0 and 1, respectively, and go to the $(n+1)$st iteration. ∎

By Lemma 3.2, if the initial σ is overlap-free, $\mathscr{A}$ must terminate successfully. However, $\mathscr{A}$ may also terminate successfully even if the initial σ is not overlap-free. Thus if $\sigma = 001100110$ (not overlap-free), $\mathscr{A}$ terminates successfully at the third iteration:

1st iteration:	$\sigma_1 = 001100110$
	$\pi_1 = 0$
	$\varrho_1 = 01100110$
	$\pi'_1 = \lambda$
2nd iteration:	$\sigma_2 = 0101$
	$\pi_2 = \lambda$
	$\varrho_2 = 0101$
	$\pi'_2 = \lambda$
3rd iteration:	$\sigma_3 = 00$

At every iteration of $\mathscr{A}$ in the above example, there is a unique decomposition of σ_i in the form prescribed by Lemma 3.1, even though the initial σ is not overlap-free.

Our next task is to derive an algorithm $\mathscr{B}$ from $\mathscr{A}$ which will terminate successfully just in case the initial σ is overlap-free.

3.4. LEMMA. *Let* $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$, $\varrho \in \{01, 10\}^+$ *and* $|\varrho| > 4$. *Then* $\pi\varrho\pi'$ *is overlap-free* $\Leftrightarrow$ *both* $\pi\varrho$ *and* $\varrho\pi'$ *are overlap-free.*

Proof. The left-to-right implication is immediate. For the converse, assume both $\pi\varrho$ and $\varrho\pi'$ are overlap-free but that $\pi\varrho\pi'$ is not, and we shall get a contradiction. Under this assumption, $|\pi| \neq 0$ and $|\pi'| \neq 0$. Because ϱ is of even length, not both $|\pi| = 1$ and $|\pi'| = 1$, otherwise $\pi\varrho\pi'$ would not be of the form $\tau\tau x$ where x is the leftmost symbol of τ. With no loss of generality, assume $|\pi| = 2$. This implies that $\pi = xx$ and $\pi' = x$ (or xx), where $x \in \{0, 1\}$, with the shortest overlap in $\pi\varrho\pi'$ being $xx\varrho x$ (or $x\varrho xx$). But if $\varrho \in \{01, 10\}^+$, it is now easily checked that both $\pi\varrho$ and $\varrho\pi'$ contain an overlap, contradicting the initial assumption. ∎

In the preceding lemma we cannot ignore the condition $|\varrho| > 4$ in the hypothesis. For example, if $\pi = 00$, $\varrho = 1001$, and $\pi' = 00$, both $\pi\varrho$ and $\varrho\pi'$ are overlap-free but $\pi\varrho\pi'$ is not.

3.5. LEMMA. *Let* $x, y \in \{0, 1\}$ *and* $\varrho \in \{01, 10\}^+$. *Then*:

(1) $x\varrho$ *overlap-free* $\Leftrightarrow \bar{x}x\varrho$ *overlap-free,*

(2) ϱy *overlap-free* $\Leftrightarrow \varrho y\bar{y}$ *overlap-free.*

Proof. The right-to-left implications are trivially true. We prove the left-to-right implication in (1) only; the same proof applies to (2). Assume then that $x\varrho$ is overlap-free, $\bar{x}x\varrho$ is not, and we get a contradiction. Under this assumption, the shortest overlap in $\bar{x}x\varrho$ contains the leftmost $\bar{x}$; i.e., $\bar{x}x\varrho$ contains a prefix $\tau\tau\bar{x}$ where $\tau \neq \lambda$. Given that $\varrho \in \{01, 10\}^+$, it is easily seen that τ cannot be of odd length. And if τ is of even length, $x\varrho$ has already a prefix $\tau'\tau'x$, with $|\tau'| = |\tau|$, which contradicts the assumption that $x\varrho$ is overlap-free. ∎

3.6. LEMMA. *Let* $x, y \in \{0, 1\}$ *and* $\varrho \in \{01, 10\}^+$. *Then*:

(1) $xx\varrho$ *overlap-free* $\Rightarrow \bar{x}x\varrho$ *overlap-free,*

(2) *ϱyy overlap-free $\Rightarrow \varrho y\bar{y}$ overlap-free,*

(3) *$\bar{x}x\varrho$ overlap-free $\Rightarrow$ if $xx\varrho$ has an overlap then xxx or $xx\bar{x}xx\bar{x}x$ is a prefix of $xx\varrho$,*

(4) *$\varrho y\bar{y}$ overlap-free $\Rightarrow$ if ϱyy has an overlap then yyy or $y\bar{y}yy\bar{y}yy$ is a suffix of ϱyy.*

Proof. The proofs for (1) and (2) are similar to those of the left-to-right implications in the preceding lemma. We prove (3) only; the proof for (4) is similar.

For (3), assume $\bar{x}x\varrho$ is overlap-free and $xx\varrho$ is not. Hence $xx\varrho$ contains a prefix $\tau\tau x$, with $\tau \neq \lambda$. Given that $\varrho \in \{01, 10\}^+$ and that τ is x or starts with xx, the length of τ must be odd. Given that $x\varrho$ is overlap-free, this forces the prefix $\tau\tau x$ to be xxx or $xx\bar{x}xx\bar{x}x$. ∎

We define a partial function φ on $\{0, 1\}^+$. φ is defined for all words of the form $\pi\varrho\pi'$ where $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$ and $\varrho \in \{01, 10\}^+$ by:

$$\varphi(\pi\varrho\pi') = \begin{cases} \varrho, & \text{if } \pi = \pi' = \lambda; \\ \bar{x}x\varrho, & \text{if } \pi = x \text{ or } xx, \text{ with } x \in \{0, 1\}, \text{ and } \pi' = \hat{\lambda}; \\ \varrho y\bar{y}, & \text{if } \pi' = y \text{ or } yy, \text{ with } y \in \{0, 1\}, \text{ and } \pi = \lambda; \\ \bar{x}x\varrho y\bar{y}, & \text{if } \pi = x \text{ or } xx, \pi' = y \text{ or } yy, \text{ with } x, y \in \{0, 1\}. \end{cases}$$

3.7. LEMMA. *Let $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$, and $\varrho \in \{01, 10\}^+$. Hypothesis:*

(1) If ($\pi = x$ or xx) and ($\pi' = y$ or yy), then $x\varrho y \neq \tau\tau$ for all $\tau \in \{0, 1\}^+$;

(2) If $\pi = xx$ then neither xxx nor $xx\bar{x}xx\bar{x}x$ is a prefix of $\pi\varrho$;

(3) If $\pi' = yy$ then neither yyy nor $y\bar{y}yy\bar{y}yy$ is a suffix of $\varrho\pi'$.

Conclusion: $\pi\varrho\pi'$ overlap-free $\Leftrightarrow \varphi(\pi\varrho\pi')$ overlap-free.

Proof. We consider the case when $|\varrho| > 4$, so that Lemma 3.4 can be used. For the case when $|\varrho| \leqslant 4$, the lemma is established exhaustively.

By Lemmas 3.4, 3.5, and 3.6, it is easy to see that all we need to prove is the implication: $\bar{x}x\varrho$ and $\varrho y\bar{y}$ overlap-free $\Rightarrow \bar{x}x\varrho y\bar{y}$ overlap-free. Assume that $\bar{x}x\varrho$ and $\varrho y\bar{y}$ are overlap-free, but that $\bar{x}x\varrho y\bar{y}$ is not. A shortest overlap (an expression of the form $\tau\tau z$ with z the first symbol of τ) in $\bar{x}x\varrho y\bar{y}$ is therefore $x\varrho y\bar{y}$, or $\bar{x}x\varrho y$, or $\bar{x}x\varrho y\bar{y}$. Because $\bar{x}x\varrho y\bar{y}$ is of even length, it cannot be a shortest overlap. Hence, $x\varrho y\bar{y}$ or $\bar{x}x\varrho y$ is a shortest overlap. But in both cases, this contradicts the fact that $x\varrho y \neq \tau\tau$ for all τ. ∎

We cannot omit condition (1) in the hypothesis of the preceding lemma. For example, let $\pi = 0$, $\pi' = 1$, and $\varrho = 011001$, so that $\pi\varrho\pi' = \tau\tau$ with $\tau = 0011$. Here $\pi\varrho\pi'$ is overlap-free, but $\varphi(\pi\varrho\pi')$ is not.

We define another partial function ψ on $\{0, 1\}^+$. ψ is defined for all words of the form $\pi\varrho\pi'$ where $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$ and $\varrho \in \{01, 10\}^+$, by:

$$\psi(\pi\varrho\pi') = \begin{cases} \varrho, & \text{if } \pi = \pi' = \lambda; \\ \bar{x}x\varrho, & \text{if } \pi = x \text{ or } xx, \text{ with } x \in \{0, 1\}, \text{ and } \pi' = \lambda; \\ \varrho y\bar{y}, & \text{if } \pi = \lambda \text{ or } x \text{ or } xx, \pi' = y \text{ or } yy, \text{ with } x, y \in \{0, 1\}. \end{cases}$$

3.8. LEMMA. *Let* $\pi, \pi' \in \{\lambda, 0, 1, 00, 11\}$, *and* $\varrho \in \{01, 10\}^+$. *Hypothesis*:

(1) *If* ($\pi = x$ *or* xx) *and* ($\pi' = y$ *or* yy), *then* $x\varrho y = \tau\tau$ *for some* $\tau \in \{0, 1\}^+$;

(2) *If* $\pi = xx$ *then neither* xxx *nor* $xx\bar{x}xx\bar{x}x$ *is a prefix of* $\pi\varrho$;

(3) *If* $\pi' = yy$ *then neither* yyy *nor* $y\bar{y}yy\bar{y}yy$ *is a suffix of* $\varrho\pi'$.

Conclusion: $\pi\varrho\pi'$ *overlap-free* $\Leftrightarrow \psi(\pi\varrho\pi')$ *overlap-free.*

Proof. We consider the case when $|\varrho| > 4$, so that Lemma 3.4 can be used. For the case $|\varrho| \leqslant 4$, the lemma is established exhaustively. By Lemmas 3.4, 3.5, and 3.6, it is easy to see that we only need to prove that: $x\varrho y$ contains an overlap $\Rightarrow$ both $x\varrho$ and ϱy contain an overlap.

Assume $x\varrho y$ contains an overlap, but not ϱy, and we get a contradiction. (The proof that $x\varrho$ contains an overlap is similar.) By hypothesis, $x\varrho y = \tau\tau$ so that $x\varrho y = x\tau' y x\tau' y$ for some $\tau' \in \{0, 1\}^+$. Because ϱy is overlap-free, $x\varrho y$ has a prefix $x\tau''\tau''$ for some $\tau'' \in \{0, 1\}^+$ which ends with symbol x. Hence $\tau''\tau''$ is a non-empty prefix of $\tau' y x\tau' y = \varrho y$. If $\tau''\tau''$ is also a prefix of $\tau' y$ then ϱy contains the sub-expression $x\tau''\tau''$ which is an overlap, contradicting the initial assumption. If $\tau' yx$ is a prefix of $\tau''\tau''$, itself a prefix of $\tau' yx\tau' y$, it is not difficult to see that $\tau' xy\tau' y$ contains an overlap – another contradiction. ∎

Observe that $\varphi(\pi\varrho\pi')$ and $\psi(\pi\varrho\pi')$ are words in $\{01, 10\}^+$, and $\varphi(\pi\varrho\pi') = \psi(\pi\varrho\pi')$ whenever $\pi = \lambda$ or $\pi' = \lambda$.

3.9. ALGORITHM $\mathscr{B}$. The input is an arbitrary $\sigma \in \{0, 1\}^+$. At the first iteration we set $\sigma_1 = \sigma$. At the nth iteration, $n \geqslant 1$, we carry out the following steps:

1. If $\sigma_n \in \{0, 1, 00, 11\}$, terminate successfully.

2. Decompose σ_n as $\pi_n \varrho_n \pi'_n$, with $\pi_n, \pi'_n \in \{\lambda, 0, 1, 00, 11\}$ and $\varrho_n \in \{01, 10\}^+$. If this is not possible, terminate unsuccessfully. If σ_n has more than one such decomposition, take π_n as short as possible.

3. If $\pi_n = xx$, with $x \in \{0, 1\}$, and xxx or $xx\bar{x}xx\bar{x}x$ is a prefix of $\pi_n\varrho_n$, terminate unsuccessfully.

4. If $\pi'_n = yy$, with $y \in \{0, 1\}$, and yyy or $y\bar{y}yy\bar{y}yy$ is a suffix of $\varrho_n\pi'_n$, terminate unsuccessfully.

5. If ($\pi_n = x$ or xx) and ($\pi'_n = y$ or yy) and ($x\varrho_n y = \tau\tau$ for some $\tau \in \{0,1\}^*$) then go to 6 else go to 7.

6. Define σ_{n+1} from $\psi(\pi_n\varrho_n\pi'_n)$ by mapping consecutive occurrences of 01 and 10 into 0 and 1, respectively, and go to the $(n+1)$st iteration.

7. Define σ_{n+1} from $\varphi(\pi_n\varrho_n\pi'_n)$ by mapping consecutive occurrences of 01 and 10 into 0 and 1, respectively, and go to the $(n+1)$st iteration. ∎

3.10. THEOREM. *Algorithm* $\mathscr{B}$ *terminates successfully on input* $\sigma \in \{0, 1\}^+ \Leftrightarrow \sigma$ *is overlap-free. If* $|\sigma| = n$, *there* $\mathscr{B}$ *executes at most* $\mathcal{O}(n \log n)$ *steps.*

Proof. The correctness of algorithm $\mathscr{B}$ follows from Lemmas 3.7 and 3.8.

Its time complexity is obtained by observing that $\mathscr{B}$ executes at most $\mathcal{O}(\log n)$ iterations, and at every iteration $\mathscr{B}$ needs to scan a word of length $\leqslant n$ at most three times. ∎

3.11. THEOREM. *There are at most* $\mathcal{O}(n^e)$ *overlap-free words of length* n, *where* $e \leqslant 2.8$.

Proof (outlined). The analysis is simpler if we modify Algorithm $\mathscr{B}$. Let $U \subset \{0, 1\}^{\leqslant 4}$ be the set of overlap-free words of length $\leqslant 4$. We replace Step 1 in $\mathscr{B}$ by the following:

1. If $\sigma_n \in U$, terminate successfully; and if $\sigma_n \in \{0, 1\}^{\leqslant 4} - U$, terminate unsuccessfully. We call $\mathscr{C}$ the algorithm obtained from $\mathscr{B}$ after this modification.

The bound mentioned in the statement of the theorem is a bound on the number of words of length n on which $\mathscr{C}$ terminates successfully. If $\mathscr{C}$ terminates successfully on $\sigma \in \{0, 1\}^+$ and $|\sigma| = n \geqslant 1$, then $\mathscr{C}$ executes k iterations for some $k \leqslant \lceil \log(n-2) \rceil$. Indeed, it is not difficult to check that $|\sigma_i| \leqslant (n+2^i-2)/2^{i-1}$ for $i \geqslant 1$, so that the largest possible value of i such that $|\sigma_i| \leqslant 4$ is $\lceil \log(n-2) \rceil$.

Assume $\mathscr{C}$ executes k iterations and terminates successfully. Hence, the tests of Steps 2, 3, and 4, are always false in the course of this execution. The only test which may switch from false to true, or vice-versa, is that of Step 5. (The test of Step 1 is false throughout except in the kth iteration.)

Case 1. The test of Step 5 remains false throughout the execution of $\mathscr{C}$ (the first $k-1$ iterations of $\mathscr{C}$). In this case, the input σ is fully determined by the values of $\pi_1, \pi'_1, \pi_2, \pi'_2, \ldots, \pi_{k-1}, \pi'_{k-1}$, and σ_k; i.e., for a given value of $\sigma_k \in U$ by running the algorithm in "reverse" we uniquely reconstruct the input σ, if we also know the values of $\pi_1, \pi'_1, \ldots, \pi_{k-1}, \pi'_{k-1}$. Although each π_i (resp. π'_i) may assume one of 5 values, $1 \leqslant i < k$, only three cases may arise in relation to π_i (resp. π'_i) for a given value of σ_{i+1}. Indeed, if the leftmost (resp. rightmost) symbol of σ_{i+1} is $x \in \{0, 1\}$, such an x being mapped from $x\bar{x}$ at the end of the ith iteration, then:

either (A) $\pi_i = \lambda$ (resp., $\pi'_i = \lambda$),

or (B) $\pi_i = \bar{x}$ (resp. $\pi'_i = x$),

or (C) $\pi_i = \bar{x}\bar{x}$ (resp. $\pi'_i = xx$).

Further, if $\pi_i = \bar{x}\bar{x}$ (case C) then $\pi_{i+1} = x$ or xx (case B or C); i.e., an instance of C at the left end (resp., right end) at the ith iteration cannot be followed by an instance of A at the left end (resp., right end) at the $(i+1)$st iteration. Further, if $\pi_i = \bar{x}\bar{x}$ (case C) and $\pi_{i+1} = x$ (case B), it is not difficult to see that $\pi_{i+2} = \lambda$ (case A) and $\pi_{i+3} = \lambda$ or $\bar{x}$ (case A or B). Hence, if $F(k-1)$ is the number of sequences in $\{A, B, C\}^{k-1}$ not containing any occurence of the following patterns $\{CA, CBB, CBC, CBAC\}$, then for each value of $\sigma_k \in U$ the sequences $\pi_1 \pi_2 \ldots \pi_{k-1}$ and $\pi'_1 \pi'_2 \ldots \pi'_{k-1}$ may each assume no

more than $F(k-1)$ values. There are therefore at most $16 \cdot F(k-1) \cdot F(k-1)$ input words for which $\mathscr{C}$ will terminate successfully after k iterations, 16 being the number of overlap-free words of length 3 or 4 – under the assumption that the test of Step 5 remains false throughout the execution of $\mathscr{C}$.

Case 2. The test of Step 5 is true in the ith iteration, for some $i < k-1$. In this case it is not difficult to show that in the jth iteration, for $j = i+1, \ldots, k-1$: if the test of Step 5 is false then $\pi_j = \pi'_j = \lambda$; if the test of Step 5 is true then $\pi_j = x$ and $\pi'_j = \bar{x}$ for some $x \in \{0, 1\}$ (details omitted). Hence, compared to Case 1, the possible values for the sequences $\pi_1 \pi_2 \ldots \pi_{k-1}$ and $\pi'_1 \pi'_2 \ldots \pi'_{k-1}$ are further reduced, and cannot therefore exceed the bound $16 \cdot F(k-1) \cdot F(k-1)$ mentioned above.

From the preceding analysis, the number of input words for which $\mathscr{C}$ terminates successfully in k iterations cannot exceed $(k-1) \cdot 16 \cdot F(k-1) \cdot F(k-1)$. Case 2 covers $k-2$ subcases, one for every value of $i \in \{1, \ldots, k-2\}$.

It remains to evaluate the function $F(v)$, for $v \geqslant 1$. It is not difficult to prove that this function satisifies the recurrence relation: $F(v+2) = 3F(v+1) - F(v)$, for $v \geqslant 1$, with initial conditions $F(1) = 3$ and $F(2) = 8$. Using standard techniques, the solution of this recurrence relation is:

$$F(v) = (1/10)\left[(5+3\sqrt{5})\left(\frac{3+\sqrt{5}}{2}\right)^v + (5-3\sqrt{5})\left(\frac{3-\sqrt{5}}{2}\right)^v\right].$$

The second term contributes very little to the rate of growth of $F(v)$. In fact, for all $v \geqslant 1$:

$$F(v) \leqslant 1.172(2.618)^v.$$

Hence, the desired bound $16 \cdot (k-1) \cdot F(k-1) \cdot F(k-1)$ does not exceed:

$$16 \cdot (\log n) \cdot (1.172)^2 \cdot (2.618)^{2\log n} \approx 22 \cdot (\log n) \cdot n^{2.78},$$

which is $\mathcal{O}(n^e)$ for any $e > 2.78$. ∎

A more careful analysis should make the exponent e even smaller in the preceding theorem. We conjecture that $e \leqslant 2$. Putting Theorems 2.6 and 3.11 together, we deduce that: In the infinite complete binary tree (represented by $\{0, 1\}^\omega$) there are uncountably many overlap-free ω-paths, which are also sparse (in the sense that, out of the 2^n nodes at level n, only $\mathcal{O}(n^e)$ may occur along these ω-paths).

References

[1] J. Berstel, *Some recent results on squarefree words*, Proc. of Symposium on Theoretical Aspects of Computer Science, 11–13 April 1984, Paris.

[2] M. Crochemore, *Linear searching for a square in a word*, Bulletin of EATCS, No. 24, October 1984, 66–72.

[3] A. J. Kfoury, *Definability by deterministic and non-deterministic programs* (with applications to first-order dynamic logic), Information and Control (to appear).
[4] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, 1983.
[5] A. Salomaa, *Jewels of Formal Language Theory*, Computer Science Press, 1981.
[6] Stolboushkin, *Deterministic context-free dynamic logic is strictly weaker than context-free dynamic logic*, Information and Control 59, 1–3 (1983), 94–107.
[7] –, Taitslin, *Deterministic dynamic logic is strictly weaker than dynamic logic*, Information and Control 57, 1 (1983), 48–55.
[8] P. Urzyczyn, *Non-trivial definability by flowchart programs*, Information and Control 58, 1–3 (1983), 59–87.

Presented to the semester
Mathematical Problems in Computation Theory
September 16–December 14, 1985