ON DISCERNING WORDS BY AUTOMATA

P. Goralčík and V. Koubek
Faculty of Mathematics and
Physics, Charles University
Sokolovská 83, Prague, ČSSR

Two distinct words $u,v \in A^*$ over a finite alphabet A are said to be r-discernible, if there is an r-state automaton $(Q,\S)$ accepting u, v by two distinct states, respectively. Put otherwise, if for an action $\S: Q \times A^* \longrightarrow Q: (q,w) \longmapsto q\S w$, satisfying $(q\S w)\S a = q\S(wa)$ for $w \in A^*$, $a \in A$, of the free monoid A^* on an r-set Q of states, we have $q_o\S u \neq q_o\S v$ for some $q_o \in Q$. If, moreover, u, v are not s-discernible for any $s < r$, we write $D(u,v) = r$.

Ch. Choffrut proposed to us to study the "worst-case discernibility" function $f(n)$, defined as the minimum such that $D(u,v) \leq f(n)$ for any two distinct words u,v of lengths $|u|, |v| \leq n$. His question was, how small $f(n)$ is, as compared to n, especially for big n. In this note we give the main lines of the proof that $\lim f(n)/n = 0$. The full proof is rather long and will be published elsewhere.

For $|A| = 1$ the result is rather trivial. Using the well-known results on prime numbers, one can prove

Lemma 1. For a suitable constant c and integer $\tilde{n}$, to every $n > \tilde{n}$ there exists a non-divisor m of n with $m < c.\log n$.

Any pair u, v of distinct words of length $\leq n$ can thus be discerned by a counter modulo r for some $r < c.\log n$.

For $|A| > 1$ we observe that the discernibility of words in A^* in fact does not depend on the size of A. In the sequel we fix our alphabet to be $A = \{0,1\}$.
The function $f(n)$ is obviously non-decreasing and unbounded. We now make and will keep throughout a basic assumption that $\limsup f(n)/n = d > 0$. In what follows we shall refute this assumption by showing that $\limsup f(n)/n < d$. To do this, we have to consider quite a number of cases, each representing a certain set C of pairs of distinct words. We say that a case C can be settled, or is amenable, if we are able to give an integer n_C and a real number d_C, $0 < d_C < d$, such that $D(u,v) \leq |u| d_C$ for all $(u,v) \in C$ with $|u| > n_C$. Clearly, if we

can settle each of an exhausting set of cases separately, then max n_C and max d_C, taken over all the cases, will settle the union of the cases.

To start with, we settle the case of the pairs (u,v) of words of different lengths by Lemma 1. More generally, define the logarithmic case as the set of all pairs of words (u,v) of length n, for which $D(u,v) \leq c.\log n$.

We now describe certain types of automata and certain proof techniques based on them.

A translator T of size $|T| = k$ is a $(k+2)$-state automaton with k non-terminal states $Q = \{q_0,\dots,q_{k-1}\}$, q_0 initial, and two terminal states t_0, t_1 such that $t_0\S0 = t_0\S1 = t_0$, $t_1\S0 = t_1\S1 = t_1$. For every $w \in A^*$, define recursively its translation $T(w)$ by T: Set $T(w) = e$ - the empty word, and start reading w by T started in q_0. Whenever T reaches t_0 or t_1 rewrite $T(w)$ to $T(w)0$ or $T(w)1$, respectively, restart T in q_0 and go on reading the rest of w. The T-translation $T(w)$ is obtained upon the reading of the whole w.

Lemma 2. If $T(u) \neq T(v)$ for some translator T and $u,v \in A^+$, then

$$D(u,v) \leq |T|.\ D(T(u),T(v))$$

Proof. If $(P,\S)$ is an automaton discerning $T(u)$, $T(v)$, then the automaton $(Q\times P,\S)$ defined by $(q,p)\S a = (q\S a,p)$ for $t_0 \neq q\S a \neq t_1$, $(q,p)\S a = (q_0,p\S b)$ for $q\S a = t_b$, $b \in A$, for $(q,p) \in Q\times P$ and $a \in A$, discernes u, v.

A translator T is good for $w \in A^*$ if $|T|.|T(w)| \leq |w|$; T is said to be ε-perfect for a positive real $\varepsilon < 1$ if there exists a word w' obtained from w by deleting at least $\varepsilon|w|$ letters, with T good for w' and $T(w') = T(w)$. We have then $|T|.|T(w)| \leq |w|(1-\varepsilon)$. T is good (ε-perfect) for a pair (u,v) of distinct words, if $T(u) \neq T(v)$ and T is good (ε-perfect) for both u and v.

A counter K of size $|K| = k$ is a $(k+1)$-state automaton with k non-terminal states $Q = \{q_0,\dots,q_{k-1}\}$, q_0 initial, and a single terminal state t such that $t\S0 = t\S1 = t$. Any automaton M can be attached to K, so as to form their concatenation $K \circ M$. To this end, undefine the action of K on its terminal state t and identify t with the initial state of M. An s-th concatenation power, or s-power, of K is defined as $K \circ \dots \circ K$, s times.

A counter K is said to be small for $w \in A^*$ if $|K| \leq |w|d/2$. If an

s-power of K is small for w of length n, i.e. $s|K| \leq nd/2$, then we also say that the number of disjoint occurrences in w of words accepted by K is small. A counter K is small for a pair of words (u,v) if it is small for both u and v. For $w \in A^*$, let $K(w)$ denote the biggest suffix of w, $w = w'K(w)$, such that w' is accepted by K.

Lemma 3. Let C be a case. If for every $(u,v) \in C$, $|u| = |v| = n$, there is a small counter K such that $(K(u),K(v))$ is in the logarithmic case, then C is amenable.

Proof. $K \circ M$ with M minimal discerning $K(u)$, $K(v)$ makes $D(u,v) \leq$ $\leq nd/2 + c.\log n \leq 3nd/4$ for big enough n.

Lemma 4. Let C be a case, B an amenable case, I an integer. If for every $(u,v) \in C$, $|u| = |v| = n$, there exists a good translator T of size $|T| \leq I$ with $(T(u),T(v)) \in B$, then C is amenable.

Proof. If $|T(u)| \neq |T(v)|$ then $D(T(u),T(v)) \leq c.\log n$, hence by Lemma 2, $D(u,v) \leq |T|.c.\log n \leq Ic.\log n$, hence $D(u,v) \leq nd_B$ for big enough n. If $|T(u)| = |T(v)| \leq nd_B/|T|$ then $D(T(u),T(v)) \leq nd_B/|T|$, hence by Lemma 2, $D(u,v) \leq nd_B$. If $|T(u)| = |T(v)| > nd_B/|T|$ then $|T(u)| > nd_B/I$, hence $|T(u)| > n_B$ for big enough n, thus $D(T(u),T(v)) \leq$ $\leq |T(u)|.d_B$, and by Lemma 2, $D(u,v) \leq |T|.|T(u)|.d_B \leq nd_B$.

Lemma 5. Let C be a case, let $\varepsilon > 0$. If for every $(u,v) \in C$, $|u| =$ $= |v| = n$, there is an ε-perfect translator T of size $|T| = r =$ $= r(u,v) < n(1-\varepsilon)/I(\eta)$, for some $\eta < \varepsilon d/(1-\varepsilon)$ and an integer $I(\eta)$ such that $f(n) < n(d+\eta)$ for every $n > I(\eta)$ (existing by the basic assumption on $f(n)$), then C is amenable.

Proof. For big enough n, $|T(u)| < n(1-\varepsilon)/r > I(\eta)$, hence we have $D(T(u),T(v)) \leq n(1-\varepsilon)(d+\eta)/r$, therefore $D(u,v) \leq n(1-\varepsilon)(d+\eta) < nd_C$ for $(1-\varepsilon)(d+\eta) < d_C < d$.

Let $\tilde{A}$ denote the monoid of canonical forms for the monoid generated by $A = \{0,1\}$ subject to the defining relation $11 = 1$, $\tilde{A} = A^* - A^*11A^*$. Let $P(w)$ and $S(w)$ denote the set of all prefixes (left factors) and the set of all suffixes (right factors) of $w \in A^*$, respectively.

For every $w \in \tilde{A}$, define a counter $K_w = (P(w),\S)$, with e initial, w terminal, and $q\S a = \max(P(w) \cap S(qa))$ for q non-terminal, $a \in A$, and define a translator $T_w = (P(w0) \cup P(w1),\S)$, with e initial, $w0$, $w1$ terminal, and $q\S a = \max((P(w0) \cup P(w1)) \cap S(qa))$ for q non-terminal.

A counter $K_w^q(x)$, for $q,x \in P(w)-\{w\}$, $x \neq e$, is obtained from K_w as follows: undefine $\S$ in w, identify w with q, add a terminal

state $xa \in \tilde{A}$ for $a \in A$ such that $xa \notin P(w)$, and redefine $x\S a$ to be xa. A translator $T_w^q(x,y)$ is obtained from $K_w^q(x)$, for $y \in P(w)-\{e,x,w\}$, by adding a second terminal state yb, for $b \in A$ such that $yb \notin P(w)$, and redefining $y\S b$ to be yb.

Our Basic Case will henceforth consist of the pairs (u,v) of distinct words of equal length n, which can be written, for some m, as

$$u = 0^{i_0} 1^{j_1} 0^{i_1} 1^{j_2} \ldots 1^{j_m} 0^{i_m} \ldots$$

$$v = 0^{k_0} 1^{l_1} 0^{k_1} 1^{l_2} \ldots 1^{l_m} 0^{k_m} \ldots,$$

where $i_0 = k_0$, $j_1 = l_1$, $\ldots$, $j_m = l_m$, but $i_m < k_m$.

This assumption means that the first difference occurs in the exponents describing the blocks of zeros in u and v. The case when such a difference first occurs with the blocks of ones is perfectly symmetric and thus equivalent to this one.

Classification of the Basic Case:

Each case imposes specific conditions on the typical pair above. Let J be an integer such that $f(n) < 5nd/4$ for all $n > J$. Denote $B = 3/4J$.

Case 1. $12/d \leq i_m < Bn$

Let s denote the number of the blocks of zeros in u. If $s \leq nd/3$, then $\max(P(u) \cap P(v))$ contains a small number of occurrences of $w = 01$, hence for a small power K of K_w we have $(K(u),K(v))$ in the logarithmic case and Lemma 3 applies. If $s > nd/3$, then the average size of the zero blocks in u is smaller than n/s, thus than $3/d$. If now there is p blocks of size i_m in u, then there must be a certain number q of blocks of size less than $3/d$ (less than the average, in fact) since i_m of Case 1 is much over the average size. To get a lower bound for q, assume that $q = s-p$ is the number of blocks of size 1. Then $(q + 12p/d)/(q+p) < (q + p(i_m+1))/(q+p) < 3/d$, whence we get that $q > 3p$, $4q/3 > q+p = s > nd/3$, thus $q > nd/4$. For $w = 0^{i_m}$, T_w skips over blocks of zeros shorter than i_m, hence is $(d/4)$-perfect, and, in view of the choice of the upper bound on i_m in this case, Lemma 5 applies.

Case 2. $i_m > Bn$

We show how to settle this case with the aid of the following lemma, which we state without proof.

Lemma 6. Let q be an integer, $q > 1$. Then for every $n > (10q)^4$ and for every pair of integers $b,c \in [3n/q, n-1]$ there exists an integer $r \leq n/q$ such that $b \neq c$ both modulo r and modulo $r+1$ and the four

remainders of the division of b and c by r and $r+1$ are pairwise different and bigger than $(r+1)/2$.

Choosing $q = q(n) = (3/B).\log n$, we have i_m , $k_m \in [3n/q$, $n-1]$, thus Lemma 6 yields an integer $r \leq n/q$ such that

$i_m \equiv x \pmod r$, $k_m \equiv y \pmod r$, $(r+1)/2 < x,y < r$,

$i_m \equiv x' \pmod{r+1}$, $k_m \equiv y' \pmod{r+1}$, $(r+1)/2 < x',y' < r+1$,

for some pairwise different integers x , y , x', y'. Let $T = T^e_r(x,y)$, $T' = T^e_{r+1}(x',y')$ (here we identify integers z with 0^z). Both T and T' translate (u,v) into a pair of different words. If, say, T translates less zero blocks of u than T', then we can assign, in one-one fashion, to each block accepted by T a block accepted by T', thus skipped by T . To get one letter of $T(u)$, T has to read at least $r+1$ letters of u , thus T is good for u . By our choice of q , we have $i_m - 2r \geq Bn - 2n/q(n) = Bn(1-2/3\log n)$, hence for big enough n we have $i_m - 2r \geq Bn/2$. Take k such that $Bn/2 \leq kr < Bn/2 + r$ and delete, from both u and v , kr zeros from their m-th block of zeros, thus obtaining u' and v', respectively. Since $i_m - kr > r$, we have $T(u') = T(u)$, thus T is $(B/2)$-perfect with $r < Bn/3\log n$. Apply Lemma 5.

Case 3. $i_m < 12/d$, $i_m \neq k_m \pmod{i_m+1}$

Either a small counter of occurrences of 0^{i_j} with $i_j \equiv k_m \pmod{i_m+1}$ reduces (u,v) to the logarithmic case, or $T^e_r(x,y)$, with x and y being the remainders of the division of i_m and k_m by $r = i_m+1$, respectively, is perfect for (u,v) .

Case 4. $i_m < 12/d$, $i_{m-1} \not\equiv i_m \equiv k_m \pmod{i_m+1}$

Either the number of $i_j \equiv i_{m-1} \pmod{i_m+1}$ with $j<m$ is small or the translator from Case 3 is perfect.

Case 5. $i_m < 12/d$, $i_{m-1} \equiv i_m \equiv k_m \pmod{i_m+1}$, $i_{m-1} > 2i_m+1$

Either a small power of T_{2r} for $r = i_m+1$ reduces (u,v) to the logarithmic case, or T_w for $w = 0^{r-1}\,1\,0^{r-1}$ is perfect for (u,v).

Case 6. $i_m < 12/d$, $i_{m-1} = i_m \equiv k_m \pmod{i_m+1}$

Classification of Case 6:

Case 61. $i_j = i_{m-1}$ for all $j < m-1$

K_r for $r = i_m+1$ reduces (u,v) to the logarithmic case.

For all the remaining subcases of Case 6 let $p = \max\{j\ ;\ j<m-1\ ,\ i_j \neq$

$\neq i_{m-1}\}$.

Case 62. $m-1-p \leq 13/d$, $i_p \neq i_m+2$

Let $w = 01(0^{i_m}1)^{m-1-p}0^{i_m}$, $x = 10^{i_p-1}w$. Then either a small power of K_x reduces (u,v) to the logarithmic case, or T_w is perfect.

Case 63. $m-1-p \leq 13/d$, $i_p = i_m+2$, $i_p \not\equiv i_m \pmod{i_m+1}$

Either a small number of occurrences of $10^{i_p}1$ in $\max(P(u)\cap P(v))$, or T_r for $r = i_m$ is perfect.

Case 64. $m-1-p \leq 13/d$, $i_p = 3$, $i_m = 1$

Let $w = 1(01)^{m-1-p}0$, $x = 1000w$, and proceed as in Case 6,2.

Case 65. $m-1-p > 13/d$

For $r = i_m$, $(T_r(u),T_r(v))$ belongs to Case 2, since the T_r-translations differ first by a block of ones of length at least $13/d$. Apply Lemma 4.

Case 7. $i_m < 12/d$, $i_m \equiv k_m \pmod{i_m+1}$, $i_{m-1} = 2i_m+1$

Classification of Case 7:

Case 71. $i_j = i_{m-1}$ for all $j<m-1$

Same as Case 61. Define further p as in Case 6.

Case 72. $p < m-2$

For $r = i_m$, the T_r-translations of u and v differ first by blocks of ones, the shorter one has length 1 and is preceded by a block of length one, hence $(T_r(u),T_r(v))$ comes under one of the Cases 3 to 6.

Case 73. $p = m-2$, $i_p \not\equiv i_m \pmod{i_m+1}$

Either the number of $j<m$ with $i_j \equiv i_p \pmod{i_m+1}$ is small, or T_r for $r = i_m$ is perfect.

Case 74. $p = m-2$, $i_p \equiv i_m \pmod{i_m+1}$, $i_p \neq i_m$

Same as Case 72.

Case 75. $p = m-2$, $i_p = i_m$

Classification of Case 75.

Case 751. $i_j = i_m$ for all $j<m-2$

The second occurrence of $2i_m+1$ comes latter in u than in v , hence (u,v) comes under the logarithmic case.

For the remaining cases define $q = \max\{j\ ;\ j \leq m-2\ ,\ i_j \neq i_m\}$.

Case 752. $q \leq m-5$

For $r = i_m$, $(T_r(u), T_r(v))$ differ first by a block of ones, the shorter one has length 1 and is preceded by a block of length at least 4, hence the case comes under Cases 3 to 6.

Case 753. $q = m-4$ or $m-3$, $i_q \not\equiv i_m \pmod{i_m+1}$

Either the number of $j < m$ with $i_j \equiv i_q \pmod{i_m+1}$ or T_{i_m} is perfect.

Case 754. $q = m-3$, $i_q \equiv i_m \pmod{i_m+1}$

For $r = i_m$, $(T_r(u), T_r(v))$ comes under Case 4.

Case 755. $q = m-4$, $i_q \equiv i_m \pmod{i_m+1}$

For $w = 0^{2i_m+1}10^{i_m}$, either the number of occurrences of $w10^{i_m}1$ in $\max(P(u) \cap P(v))$ is small or T_w is perfect.

The Basic Case is settled, thus we have proved

Theorem. Any pair of distinct words of length at most n is $o(n)$- discernible.