The alternation number and a dot hierarchy of regular sets

STEPHEN L. BLOOM

In this note we introduce a property of languages we call the alternation number. This property is used to deduce several facts about regular languages in particular. One of these facts is related to what might be called the "*generalized dot height*" of a regular language. The complexity of regular expressions is usually determined by their "*star height*" [E], although other complexity measures have also been considered [EK]. In all of the papers we know of, a nontrivial argument is needed to establish that the complexity of '*regular expressions*' must grow in order to name all regular sets. (Our definition of '*regular expression*' allows an atomic name for each finite set, see below.) In the present note, we give a simple proof that the "*dot height*" (or depth of concatenation signs) of a regular expression also must grow. (The "*dot-depth*" considered in [BK] is unrelated to our dot height.) The dot height is related to the alternation number. We will show that if the alternation number of a regular language L is n, then any regular expression which denotes L contains (roughly) at least $\log n$ dots.

Define a function f from the set of nonnegative integers $\mathbf{N}$ to collections of regular subsets of Σ^* (for some fixed alphabet Σ) as follows:

$$f(0) = \text{all finite languages};$$

$$f(n+1) = f(n) \cup \{L\colon\ L = L_1 + L_2, L = L_1 \cdot L_2, \quad L = L_1^*, L_i \in f(n)\}.$$

Thus a language is regular iff it belongs to $f(n)$ for some n. We will show first that for each n, there is a language in $f(n+1) - f(n)$. The truth of this fact follows easily from the well-known star-height hierarchy theorem (see [DS] for one proof). The argument given here shows that the hieararchy of regular languages depends also on the operation of concatenation. As a corollary we will obtain the dot height hierarchy. Lastly, we mention an automaton characterization of the alternation number.

First we assume that Σ has at least two letters, say a and b. A language L *admits alternations of size* n if: for each $k > 0$ (or, equivalently, for infinitely many $k > 0$) there is a word w in L of the form

$$w(k) = u_0 x_0 \dots x_0 u_1 x_1 \dots x_1 u_2 \dots u_n x_n \dots x_n u_{n+1}$$

where

1. $u_0, u_1, \ldots, u_{n+1}$ are arbitrary words in Σ^*;
2. x_i are letters in Σ, and for each $i<n$, x_i and x_{i+1} are distinct;
3. there are k consecutive occurences of the letters x_i, $i=0, 1, \ldots, n$. (We will say that a word of the form $w(k)$ has "n alternations of length at least k".) We say a language admits alternations of size 0 if it admits no alternations. For example, finite languages admit no alternations. $\{a\}^*$, $\{b\}^*$ admits alternations of size 1, as does $L\cdot\{a\}^*\cdot L'\cdot\{b\}^*\cdot L''$, for any nonempty languages L, L' and L''. Note that if L admits alternations of length $n+1$, then L also admits alternations of length n.

Definition. The alternation number of L, $a(L)$, is n if L admits alternations of size n but not of size $n+1$. Let $a(L)=\infty$ if for each n, L admits alternations of size n.

Proposition 1. Assume $a(L)=x$, and $a(K)=y$, where $x, y\in \mathrm{N}\cup\{\infty\}$. Then

$$a(L+K)=\max(x,y);\quad x+y\leqq a(L\cdot K);\quad \text{if}\quad a(L^*)>0, \tag{1.1}$$

$$\text{then}\quad a(L^*)=\infty.$$

$$a(L\cdot K)\leqq x+y+1; \tag{1.2}$$

(Of course, $n<\infty$ and $n+\infty=\infty$, for all $n\in\mathrm{N}$.)

Proof. We prove only the last part of 1.1. If $a(L^*)>0$, for each k, there is a word w in L^* which has 1 alternation of length at least k. But then ww has at least 3 alternations of length at least k, and www has at least 5 alternations of length at least k, etc. Thus $a(L^*)=\infty$.

The proof of 1.2 is longer. Assume that $L\cdot K$ admits alternations of size s. We will show $s\leqq a(K)+a(L)+1$. For each $k>0$ there is a word in $L\cdot K$ of the form

$$w(k)=u_0x_0\ldots x_0u_1x_1\ldots x_1u_2\ldots u_sx_s\ldots x_su_{s+1}$$

as described above. We may factor each word $w(k)$ as $l(k)\cdot v(k)$, with $l(k)\in L$ and $v(k)\in K$. Suppose that $i(k)$ is the greatest integer i, $-1\leqq i\leqq s$, such that

$$u_0x_0^k\ldots u_ix_i^k$$

is an initial segment of $l(k)$ ($i(k)=-1$ if there is no such initial segment). Thus at least one of the integers between -1 and s is the value of $i(k)$ for infinitely many k; let n be the maximum of these integers, so that for infinitely many values of k, $i(k)=n$. (If $n=-1$ or $n=s$, then we may easily show that $s-1\leqq a(K)$ and $s\leqq a(L)$, respectively, so that from now on, we assume $0\leqq n<s$.)

For infinitely many values of k (say $k\in I$) we may write

$$l(k)=u_0x_0^k\ldots u_nx_n^k\,l'(k),$$

$$v(k)=l''(k)u_{n+2}x_{n+2}^k\ldots x_s^ku_{s+1},$$

where $l'(k)l''(k)=u_{n+1}x_{n+1}^k$.

Case 1. For infinitely many values of k, say k in I', $l'(k)$ is an initial segment of u_{n+1}. Then, for k in I', we may write

$$v(k) = v'(k)\, x_{n+1}^{k} u_{n+2} x_{n+2}^{k} \dots x_{s}^{k} u_{s+1},$$

so that $s-(n+1) \leqq a(K)$; also, $n \leqq a(L)$, so that $s-i \leqq a(K)+a(L)$, as claimed.

Case 2. Otherwise. Then, for infinitely many k, u_{n+1} is an initial segment of $l'(k)$. Hence, for these values of k there is a number $h(k)$ for which

$$l(k) = u_0 x_0^{k} \dots x_n^{k} u_{n+1} x_{n+1}^{k-h(k)};$$

$$v(k) = x_{n+1}^{h(k)} u_{n+2} \dots u_s x_s^{k} u_{s+1}.$$

We now have two further subcases.

Case 2a. The numbers $h(k)$ are unbounded. Then, we may write

$$v(k) = x_{n+1}^{h(k)} u'_{n+2} x_{n+2}^{h(k)} \dots u'_s x_s^{h(k)} u'_{s+1},$$

for infinitely many k, which shows that $s-(n+1) \leqq a(K)$; clearly, $n \leqq a(L)$, so that again, $s-1 \leqq a(L)+a(K)$.

Case 2b. Otherwise. In this case, the numbers $k-h(k)$ are unbounded, so that for infinitely many k,

$$l(k) = u'_0 x_0^{k-h(k)} u'_1 \dots x_n^{k-h(k)} u'_{n+1} x_{n+1}^{k-h(k)}.$$

Since the numbers $k-h(k)$ are unbounded, $n+1 \leqq a(L)$; clearly, $s-(n+2) \leqq a(K)$, so that $s-1 \leqq a(L)+a(K)$, completing the proof.

Lemma 2. For each $n \in \mathrm{N}$ define

$$g(n) = \max \{a(L)\colon\ L \in f(n) \quad \text{and} \quad a(L) < \infty\}.$$

Then $g(0)=g(1)=0$; $g(2)=1$ and for $n>0$, $g(n)=2^{(n-1)}-1$.

Proof. All languages in $f(0)$ are finite, so that $g(0)=0$; the sum and product of finite languages are finite, and $a(L^*)$ is either 0 or ∞, so that $g(1)=0$ also. Clearly $g(2)$ is at least 1 and by Proposition 1, $g(2)$ is at most 1. Assume $g(n)=2^{n-1}-1$. If $L \in f(n+1)$ and $a(L)<\infty$, then using the proposition, the largest $a(L)$ can be is $2g(n)+1=2[2^{n-1}-1]+1=2^n-1$. But it is easy to see that $g(n+1)$ is not less than $2g(n)+1$ also, completing the induction.

Theorem 3. For each positive $n \in \mathrm{N}$, there is a language in $f(n)-f(n-1)$.

Proof. Let L be a language in $f(n)$ with $a(L)=g(n)$. Then, if $n>1$, L is not in $f(n-1)$, since $g(n-1)<g(n)$. The statement is trivial for $n=1$.

What about the case that Σ is a singleton, say $\{a\}$, so that Σ^* may be identified with N? In this case, if L is an ifinite regular set, there is a finite set F and a fixed integer n and numbers $k1, \dots, kt$ such that

$$L = F \cup \{a^{k1}\} \cdot \{a^n\}^* \cup \{a^{k2}\} \cdot \{a^n\}^* \cup \dots \cup \{a^{kt}\} \cdot \{a^n\}^* = F \cup \{a^{k1}, \dots, a^{kt}\} \cdot \{a^n\}^*.$$

Hence all regular subsets of **N** are in $f(3)$.

In order to avoid the trivial cases, we assume that the regular expressions (over Σ) are built from the atomic letters (i.e. a symbol for each *finite subset of* Σ^*) and the function symbols $+$, $\cdot$, and $*$ in the usual way. (Thus, a finite set of words may be denoted by a regular expression with none of the function signs $+$, $\cdot$, $*$.) Let $|\alpha|$ be the language denoted by the regular expression α. If α is a regular expression, let the '*dot height of* α', $\mathrm{dh}(\alpha)$, be 0, when α is an atomic letter; $\mathrm{dh}(\alpha+\beta)=$ $=\max(\mathrm{dh}(\alpha), \mathrm{dh}(\beta))$; $\mathrm{dh}(\alpha^*)=\mathrm{dh}(\alpha)$, and lastly,

$$\mathrm{dh}(\alpha\cdot\beta) = 1+\max(\mathrm{dh}(\alpha), \mathrm{dh}(\beta)).$$

Let $R(n)$ denote the family of regular expressions α with $\mathrm{dh}(\alpha)<n$.

Proposition 4. Suppose that $n>0$, that L is a language denoted by α, $\alpha\in R(n)$ and that $a(L)<\infty$. Then

$$a(L) \leqq g(n).$$

Proof. By induction on n. If $n=1$ and L is denoted by a regular expression α having no dot symbols, then either α is an atomic symbol or has the form

$$\beta+\sigma, \quad \text{or} \quad \beta^*$$

for some other regular expressions β, σ with dot height 0. By induction on the structure of α, one sees that either $a(L)=\infty$ or $a(L)=0=g(1)$.

Now assume that the proposition holds for n and that L is a language denoted by a regular expression in $R(n+1)-R(n)$ and $a(L)<\infty$. If α is of the form $\beta+\sigma$ or β^*, it is easily seen by induction on the structure of α that $a(L)\leqq g(n+1)$. If α is of the form $\beta\cdot\sigma$ then $\mathrm{dh}(\beta), \mathrm{dh}(\sigma)<n$. Thus, by the induction hypothesis, $a(|\beta|)$ and $a(|\sigma|)$ are at most $g(n)$, and by proposition 1, $a(L)$ is at most $2g(n)+1=$ $=g(n+1)$, completing the proof.

Corollary 5. (The '*dot height*' hierarchy). For each $n>0$, there is an infinite regular language L not denoted by a regular expression in $R(n)$.

Proof. Any regular language L with $g(n)<a(L)<\infty$ will do, by Proposition 2.

The alternation number of a regular language may be described by certain properties of a finite automaton which accepts it. Let $\mathrm{M}=(Q, i, F)$ be a finite Σ-automaton (with state set Q, initial state i and final states F; we denote the action of a word u in Σ^* on the state q by $q\cdot u$). A state q in Q is "*accessible*" if $q=i\cdot u$, for some word u in Σ^*. If x is a letter in Σ, we call a state q *x-stable* if $q\cdot u=q$, where u is some positive power of x (i.e. $u=x$ or xx or xxx, etc.); q is stable if q is x-stable for some letter x. The "*behavior of q*", $|q|$, is the set $\{u\in\Sigma^*\colon q\cdot u\in F\}$.

We now define by induction the notion of an "n-state", for $0\leqq n$.

Definition. a) The state q will be called "*a 0-state via the letter x*" if

1. $|q|$ is nonempty;
2. q is x-stable.

b) q is an "*$n+1$ state via x*" if

1. q is x-stable;
2. there is some word v such that $q\cdot v$ is an n-state via y, for some letter $y\neq x$;

A state is an "n-state" if it is an n-state via x, for some letter x.

The easy proof of the next fact is omitted.

Lemma 6. Let $\mathbf{M}=(Q, i, F)$ be an automaton which accepts the language L. Then, for $1 \leqq n$, L admits alternations of size n iff there is some accessible n-state in Q.

Corollary 7. Let $\mathbf{M}=(Q, i, F)$ be an automaton which accepts the language L. Then, for $n>0$, $a(L)=n$ iff Q contains an accessible n-state but no accessible k-state with $k>n$. If Q has m states, then $a(L)=\infty$ iff there is an accessible m-state in Q.

Proof. We need only prove that if the cardinality of Q is m, and Q has an accessible m-state, say q_0, then $a(L)=\infty$. But, there is a sequence of words $u_0, u_1, \ldots, u_m$ such that if $q_{i+1}=q_i \cdot u_i$, for $i=0, 1, \ldots, m$, then q_i is an $m-i$ state via x_i, with $x_i \neq x_{i+1}$; since the states q_i cannot all be distinct, let s and t, $t>0$, be least such that $q_s=q_{s+t}$. It is easy to see that q_s is also an $n-s+kt$ state, for all $k>0$; hence $a(L)=\infty$.

Corollary 8. There is an algorithm to determine, given a regular language L, what the alternation number of L is.

Proof. Suppose one is given an accessible finite automaton with n states which accepts L. First one finds all the 0-states, by considering only paths of length $\leqq n$, then 1-states, etc. until one knows all the n-states. Then one applies the previous Corollary.

Questions: Is there an algorithm to determine, given a regular language L, the least n such that $L \in f(n)$? Is there an algorithm to determine the dot height of a regular language?

Acknowledgements

It is a pleasure to thank Douglas Bauer and Klaus Sutner for several helpful conversations. The author especially thanks Zoltan Esik for carefully reading the manuscript, correcting several errors and suggesting several improvements.

DEPARTMENT OF COMPUTER SCIENCE
STEVENS INSTITUTE OF TECHNOLOGY
HOBOKEN, NJ 07030
U.S.A.

References

[BK] J. A. Brzozowski and R. Knast, "The dot-depth hierarchy of star-free languages is infinite", J. Comp. and System Sci., 16 (1978) 37—55.

[DS] F. Dejean and M. P. Schutzenberger, "On a question of Eggan", Information and Control vol. 9 (1966) 23—25.

(E) L. C. Eggan, "Transition graphs and the star-height of regular events", Michigan Math. J. 10 (1963) 385—397.

[EZ] A. Ehrenfeucht and P. Zeiger, "Complexity measures for regular expressions", J. Comp. System Sci., 12 (1976) 134—146.

(Received 25 May, 1985)