A LOWER BOUND ON THE SIZE OF ϵ-FREE NFA CORRESPONDING TO A REGULAR EXPRESSION

Yuri Lifshits
St. Petersburg State University

October 25, 2006

Abstract

Hromkovič et al. showed how to transform a regular expression of size n into an ϵ-free nondeterministic finite automaton (which defines the same language as the expression) with $O(n)$ states and $O(n\log^2(n))$ transitions. They also established a lower bound $\Omega(n\log(n))$ on the number of transitions. We improve the lower bound to $\Omega(\frac{n\log^2 n}{\log\log n})$.

Keywords. Formal languages, regular expressions, combinatorial problem, epsilon-free nondeterministic automata.

AMS Subject Classification. 68Q45, 68Q25, 68W01, 68W10.

1 Introduction

There are several equivalent definitions of regular languages, out of which we will consider two classical ones. One way is to use regular expressions, the second one is to use finite automata. It is well known that for each automaton there is a regular expression which defines the same language as recognized by the given automaton. The converse assertion is true, too. We are interested in the problem: how small can be an automaton corresponding to a regular expression of size n. The size of automaton is defined as the number of transitions, while the size of a regular expression is the number of symbols occurring in it. Computing such automata is important, for example, for intersection or membership tests. In this paper we consider nondeterministic finite automata without ϵ-transitions (ϵ-free NFA for short).

A well-known method for constructing ϵ-free NFA from regular expressions is based on position automata (Glushkov automata). This classical construction yields NFA of quadratic size (see [1]). A substantial improvement on this construction was achieved in [3] where a nondeterministic version of the position automata construction was shown to yield ϵ-free

NFA with $O(n\log^2(n))$ transitions. This is optimal up to a $\log n$ factor, as shown also in [3] by proving the lower bound:

$$\Omega(n\log n). \tag{1}$$

In this paper we improve the lower bound for the conversion of a regular expression of size n into an ϵ-free NFA to $\Omega(n\log^2 n/\log\log n)$. To obtain the new lower bound we use the regular expression which has been introduced in [3] and then used by Hagenah and Muscholl in [2], namely,

$$E_n \;=\; (a_1+\epsilon)(a_2+\epsilon)\;\ldots\;(a_n+\epsilon).$$

In [2] the authors introduced a special kind of graphs which we call *universal* graphs (the definition is given in Section 2). Hagenah and Muscholl proved the following result ([2, Lemma 6.1]):

The number of transitions in ϵ-free NFA which recognize the regular language defined by E_n is greater than or equal to the number of edges in some n-universal graph.

The paper is organized as follows. In Section 2 we give the definition of universal graphs (UG for short) and a graphical representation for them. In Section 3 we convert UG into a special *arrow structure* and prove some properties of such structures. The main result (a lower bound on the size of UG) is stated in Section 4 together with some necessary constructions. The proof is completed in Section 5.

2 Universal graphs

For a positive integer n let Γ_n denote the class of directed graphs with $n+1$ vertices having the following properties:

1. the vertices are *numbered* by $0,1,\ldots,n$;
2. the graph edges are *marked* by numbers $1,2,\ldots,n$; let $\mu(e)$ denote the mark of the edge e;
3. for every edge e the number of its start vertex $\alpha(e)$, the mark $\mu(e)$ and the number of its end vertex $\omega(e)$ satisfy the inequalities

$$\alpha(e) < \mu(e) \le \omega(e). \tag{2}$$

A sequence of edges

$$e_1,\ldots,e_k \tag{3}$$

in a graph G from Γ_n is called an *oriented route* if

$$\omega(e_1)=\alpha(e_2),\ldots,\omega(e_{k-1})=\alpha(e_k). \tag{4}$$

It follows from (2) and (4) that for every oriented route (3) the inequalities

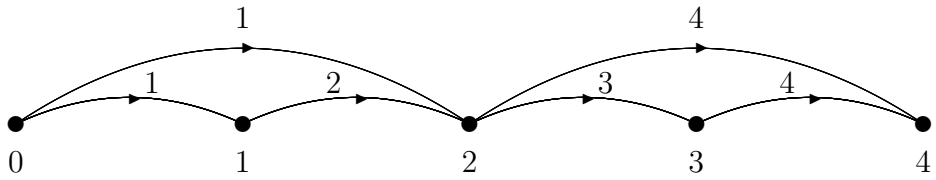
$$1 \le m_1 < \ldots < m_k \le n \tag{5}$$

hold where

$$m_1 = \mu(e_1), \ldots, m_k = \mu(e_k). \tag{6}$$

A graph G from Γ_n is called *n-universal*, if for arbitrary numbers $m_1, \ldots, m_k$ which satisfy (5) there exists an oriented route (3) satisfying (6). A graph G is called *universal* if there exists some n such that G is n-universal.

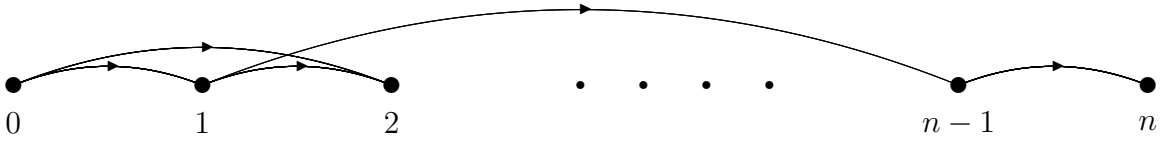
An example of an n-universal graph is given by the complete oriented graph K_{n+1} in which for every edge e we set $\mu(e) = \omega(e)$. This graph has $\frac{n(n+1)}{2}$ edges. A different example of 4-universal graph is presented on the picture below.



Let $\lambda(n)$ denote the minimal number of edges in a n-universal graph. First we will prove that $\lambda(n) = \Omega(n \log(n))$, which was the basis for the known bound (1). Then we improve it.

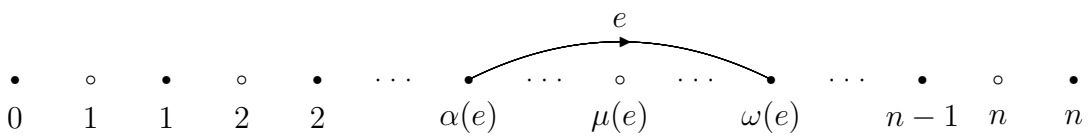
We first introduce a graphical representation of universal graphs and some new definitions.

Let G be a universal graph. Let us represent the vertices of the graph by points situated on a horizontal line in the natural order. Now we draw all edges of G. By assumption all of them go from left to right. We represent edges as oriented arcs that link corresponding pairs of points on our line. We draw all edges *above* the line.



Let us introduce *antivertices* which will be numbered by $1, 2, \ldots, n$. The antivertex k will be placed on the line between the $(k-1)$-th and the k-th vertex.

Proposition 1 *In our graphical representation for every edge e the antivertex $\mu(e)$ is situated under the arc representing the edge e.*



Proof: This statement follows directly from inequalities (2) and from the definition of numbering of antivertices. □

We define now the notion of *arrow structure* (AS). An AS is an oriented graph with $n+1$ vertices and n antivertices situated on a line and numbered in the same way as in the graphical representation of the UG. We also assume that every edge starts from an antivertex and ends in a vertex. In the sequel, the word "edge" is related only to UG while for the edges of AS we use the word "arrow".

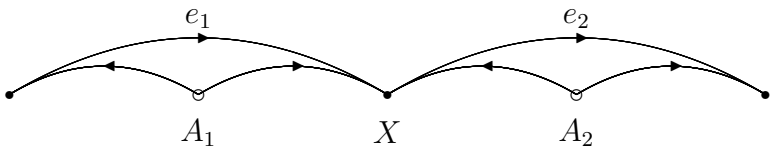
We call an arrow structure *perfect* (PAS) if it has the following property: for every pair of antivertices A and B there exists a vertex X situated between them such that there are arrows from A to X and from B to X.

3 Arrow structure associated with an universal graph

We associate an arrow structure with every universal graph. To this end, we replace every edge e from the UG by two arrows starting from the antivertex $\mu(e)$ and ending in the start vertex $\alpha(e)$ and in the end vertex $\omega(e)$ of the edge, respectively. If some identical arrows appear during the construction, we consider them as a single one, i.e. we merge them.

Proposition 2 *The arrow structure generated from an UG in this way is perfect.*

Proof: Consider two arbitrary antivertices A_1 and A_2 (w.l.o.g. $A_1 < A_2$). Apply the basic property of UG to these numbers. We get two edges e_1 and e_2 marked by A_1 and A_2 such that the end of the first edge coincides with the start vertex of the second one. Let X denote this start-end vertex. Then, by the construction of the AS, there are arrows from the antivertices A_1 and A_2 to the vertex X:



□

Theorem 1 *In every PAS with n vertices there exist $\Omega(n \log n)$ arrows.*

Proof: Let $f(n)$ denote the minimal number of arrows in a PAS with n vertices. Let us prove the following inequality:

$$f(2n+2) \geq 2f(n) + n. \tag{7}$$

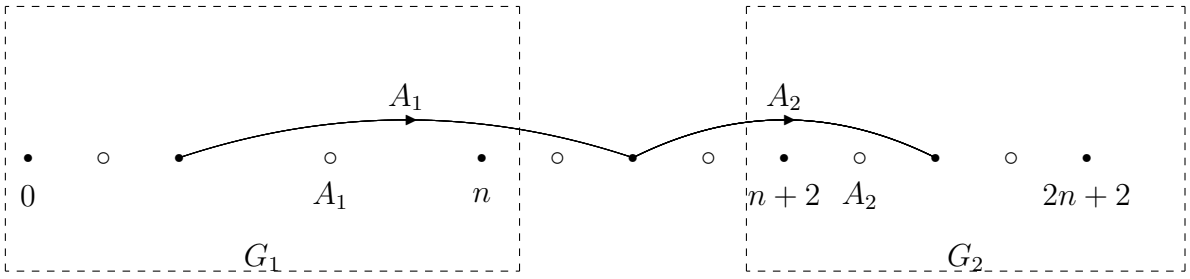
Consider an arbitrary PAS F with $2n+2$ vertices. Notice that the AS (we call it F_1) which consists of the $(n+1)$ first vertices and of the n first antivertices (i.e. vertices with numbers

$0, \ldots, n$ and antivertices with numbers $1, \ldots, n$) is perfect in its own. Indeed, it is sufficient to apply the basic property of the initial PAS F to every pair of antivertices of F_1 and we obtain the basic property of PAS for F_1. Similarly, the AS F_2 which consists of the last $(n+1)$ vertices and the last n antivertices (i.e. vertices with numbers $n+2, \ldots, 2n+2$ and the antivertices with numbers $n+3, \ldots, 2n+2$) is perfect. We call *internal* every arrow which is entirely situated inside one of arrow structures F_1 or F_2. We call *external* every arrow which is not an internal one. By the definition of the function f, there are at least $f(n)$ arrows in each of the structures F_1 and F_2.

Moreover, one of the following alternatives must hold:

- either for every antivertex in F_1 there exists an external arrow starting from it,
- or for every antivertex in F_2 there exists an external arrow starting from it.

To see this, suppose the contrary, i.e. that there is some antivertex A_1 in F_1, from which no external arrows start and some antivertex A_2 in F_2 with the same property. In this case the basic condition of PAS would not hold for A_1 and A_2.



Therefore, either for every antivertex in F_1 there is an external arrow starting from it, or for every antivertex in F_2 there is an external arrow starting from it. In any case, F contains at least $2f(n)$ internal arrows and at least n external ones (since both of F_1 and F_2 contain n antivertices). The inequality (7) is thus proved.

We omit the standard inference of Proposition 3 from the inequality (7). □

Corollary The minimal number of edges in UG is $\Omega(n \log n)$, since we constructed two arrows from every edge and, eventually, merged some of them. Therefore, the number of edges in UG is at least as large as the half of number of arrows in the associated PAS.

Remark: The lower bound from Theorem 1 is optimal up to a numerical factor, namely, it can be proved that $f(n) = O(n \log n)$.

4 Main result and auxiliary constructions

We aim to prove the following inequality which is the main result of the paper:

Theorem 2

$$\lambda(N) \geq \frac{c\ N \ln^2 N}{\ln \ln N} \tag{8}$$

with the constant $c = 0.014$.

It is sufficient to consider only the case $N > 50$, since for $N \leq 50$ we have

$$\lambda(N) \geq N \geq \frac{0.08\ N \ln^2 N}{\ln \ln N}\ .$$

4.1 Some notation and constructions

We use the principle *reductio ad absurdum* and consider an UG G with the minimal number of vertices such that its number of edges does not satisfy inequality (8). By the above remark $N > 50$. Let G have $2n + 3$ vertices (the case of even number of vertices can be considered similarly). Let us construct again the perfect arrow structure F from our universal graph and select the AS F_1 and F_2 in it by proceeding as in the proof of Theorem 1. Consider also the graph G_1 which consists of the first $(n+1)$ vertices of graph G and of the edges between them, and the graph G_2 which consists of the last $(n+1)$ vertices. There is one more vertex number $n + 1$ between these graphs.

Proposition 3 *The graphs* G_1 *and* G_2 *are universal.*

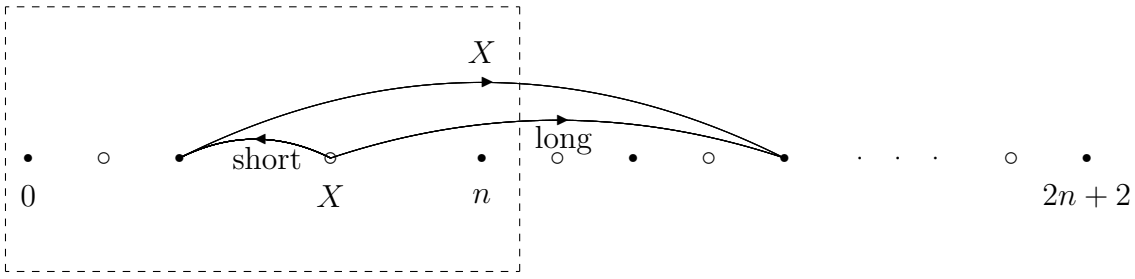
Proof: We check the condition of universality for G_1 (one can consider G_2 similarly[1]). Consider a sequence $1 \leq m_1 < \cdots < m_k \leq n$. We append to this sequence an element $m_{k+1} = n+1$ and apply to the extended sequence the condition of universality of G. Indeed, the first k edges of the obtained route provide the desired route for G_1. In fact, the edge with mark m_k ends up in the vertex from which the edge with mark $n + 1$ starts, i.e. in the vertex with number not larger than n. Therefore, these first k edges are situated entirely in G_1. □

We call *internal* the edges of graph G that belong either to G_1 or to G_2. All other edges are called *external*. Internal and external arrows are defined as in the previous section.

We aim now to show that the number of external edges in G is at least $\frac{c_1 N \ln N}{\ln \ln N}$ where $N + 1 = 2n + 3$ is the number of vertices in G with $c_1 = 0.025$.

Definition: Consider an external edge starting from a vertex in G_1 and marked by $X \in [1..n]$. While replacing it with two arrows, we add one left arrow which is internal for F_1 and call it *short*, and one external arrow which we call *long*.

[1]Strictly speaking, G_2 becomes universal only after the subtraction of $n+2$ from all numbers of its vertices and from the marks of its edges.



Short and long arrows for F_2 are defined similarly.

Proposition 4 *If one deletes all left arrows except the short ones from F_1 and all right arrows except the short ones from F_2, then either the resulting AS F_1' will be still perfect or F_2' will be perfect.*

Proof: We show the claim by contradiction. Assume that the condition of perfectness is not verified for F_1' for some antivertices with numbers A and B, and for some antivertices C and D in F_2' (without loss of generality $A < B < C < D$). Consider the route which consists of the four edges with marks A, B, C, and D. Such a route exists, since G is universal. But then either the edge with mark B is not internal for G_1, or the edge with mark C is not internal for G_2. Again without loss of generality we assume that the first of the two statements holds. Then the arrows constructed from the edges with marks A and B provide the condition of perfectness for AS F_1 for the antivertices A and B because the left arrow from B is short, since the edge with mark B is external for G_1. We arrive at a contradiction with the definition of the antivertices A and B. □

In the sequel, we may and will assume that it is the AS F_1' that is perfect.

A short digression. We have already proved that the number of arrows in PAS is $\Omega(n \ln n)$. If the similar bound would be true for the number of arrows going only in one direction, we could apply it to F_1' and would obtain the lower bound $\Omega(n \log n)$ on the number of short arrows, hence for the number of external edges in G, since each short arrow corresponds to its own external edge. Therefore, we would obtain the equality $\lambda(2n+2) = 2\lambda(n) + \Omega(n \log n)$, which immediately yields $\lambda(n) = \Omega(n \ln^2 n)$. Unfortunately, the desired bound on the number of arrows going in one direction in PAS is not true.

4.2 Constructing the arrow structure H

We construct a new arrow structure H. We start from the PAS F_1', as defined in Proposition 4. Let us define an integer j by the formula

$$j = \left\lceil \frac{16cN \ln^2 N}{n \ln \ln N} \right\rceil .$$

Since the total number of edges in the UG G is by assumption less than $\frac{cN\ln^2 N}{\ln\ln N}$, the number of antivertices of degree higher than $j/2$ in F_1' does not exceed

$$\frac{2cN\ln^2 N}{(j/2)\ln\ln N} \leq \frac{n}{4}\ .$$

We delete from F_1' all antivertices of degree higher than $j/2$. We also delete all arrows starting from deleted antivertices. After that, we merge those vertices between which no antivertices remain, and those arrows which duplicate each other.

The resulting AS (let us denote it H) will contain at least $3n/4$ antivertices. Indeed, there were n antivertices in F_1', while not more than $n/4$ of them were deleted. Notice also that the degrees of all antivertices in H do not exceed $j/2$, since the degrees of antivertices do not increase during the construction.

Proposition 5 *The arrow structure H is perfect.*

Proof: Consider two arbitrary antivertices in H. Both were antivertices in F_1', and there was a vertex X between them such that there were arrows from those antivertices to X. What could break down? The antivertices remained on their places, yet the vertex X could merge with some other(s). But it means that we should consider as a new X the vertex obtained by merging from the former X. Therefore, the perfectness condition of AS does not suffer during this construction process. □

4.3 A lower bound on the number of left arrows in H

Proposition 6 *If $n > 24$, then the number of left arrows in H is at least $\frac{c_1 N(\ln N)}{\ln\ln N}$, where $c_1 = 0.025$.*

Proof:

Let m denote the number of antivertices in H. Let $k = \lfloor \log_{j+1} m \rfloor$. Then there exist integers q and r such that

$$m = q(j+1)^k + r, \quad 1 \leq q \leq j, \quad r < (j+1)^k.$$

Forget about the last r antivertices. We split the remaining antivertices of H in q groups of "zero level", with $(j+1)^k$ antivertices in each group. In each of these groups we perform the following splitting: for every $1 \leq s \leq k$ we split the antivertices of the zero level group in $(j+1)^s$ groups of "s-th level" with $(j+1)^{k-s}$ antivertices in each group. We add to each group for each antivertex its left neighboring vertex. We say that a left arrow is of *s-th order* if its start vertex and the end vertex are situated in the same group of the $(s-1)$-th level but in different groups of the s-th level.

Let us consider a group of level s and forget for a while about the other groups. Consider the leftmost antivertex A in it. There are at most $j/2$ right arrows starting from A (since by the construction of H the degree of each antivertex does not exceed $j/2$). We call a group

of level $s+1$ a *good* one, if it is not the leftmost one (i.e. it does not contain the antivertex A) and if there is no arrow going from A to vertices of this group. Obviously, among the $j+1$ groups of the level $s+1$ we have at least $j/2$ good ones. By applying the definition of PAS (recall that H is PAS) for all pairs of antivertices $\langle A, B\rangle$ (where B is an antivertex of a good group), we obtain that for every antivertex of each good group of level $s+1$ there is at least one left arrow going from this antivertex to some vertex of another group. It follows that in each group of order s there are at least $(j/2)(j+1)^{k-s-1}$ arrows of order $s+1$.

Thus there are q groups of level 0, $q(j+1)$ groups of level 1 and, in general, $q(j+1)^s$ groups of level s. By summing up the obtained estimates over all groups and over arrows of all orders, and using the inequality $m \geq 3n/4$, we obtain that the number of left arrows in H is at least

$$q((j/2)(j+1)^{k-1}) + (q(j+1))((j/2)(j+1)^{k-2}) + \cdots + (q(j+1)^{k-1})(j/2) =$$

$$kq(j+1)^{k-1}(j/2) = \frac{kj}{2(j+1)}\, q(j+1)^k \geq$$

$$\frac{kj}{2(j+1)}\, \frac{m}{2} \geq \frac{kj}{2(j+1)}\, \frac{3n}{8} = \frac{3kjn}{16(j+1)}\ .$$

Plugging in here the estimate

$$k = \left\lfloor \frac{\ln(m)}{\ln(j+1)} \right\rfloor \geq \left\lfloor \frac{\ln(3n/4)}{\ln(j+1)} \right\rfloor ,$$

we find that the number of left arrows in H is at least

$$Z(n) = \frac{3\left\lfloor \frac{\ln(3n/4)}{\ln(j+1)} \right\rfloor jn}{16(j+1)}\ .$$

It is not too difficult to show that for all $n \geq 24$ it holds

$$\frac{Z(n)\ln\ln N}{N\ln N} \geq 0.025\ .$$

□

5 Proof of the main theorem

Since both G_1 and G_2 are universal graphs, the number of internal edges in G is at least $\frac{2cn\ln^2 n}{\ln\ln n}$. The number of external edges in G is greater than or equal to the number of short arrows in F. Hence, it is greater than or equal to the number of short arrows in F_1. The latter value is, in its turn, equal to the number of left arrows in F_1' (by the rule of its construction)

which is greater than or equal to the number of left arrows in H. By Proposition 6, this is at least $\frac{c_1 N(\ln N)}{\ln\ln N}$. Therefore we get for the number of edges in G the lower bound

$$\frac{2c\,n\ln^2 n}{\ln\ln n} + \frac{c_1 N\ln N}{\ln\ln N} \geq \frac{2c\,n\ln^2 n + c_1 N\ln N}{\ln\ln N}.$$

On the other hand, by assumption, the number of edges in G does not exceed $\frac{c\,N\ln^2 N}{\ln\ln N}$. Hence,

$$2c\,n\ln^2 n + c_1 N\ln N \leq c\,N\ln^2 N.$$

Using the definition $N = 2(n+1)$ and letting $c_2 = \frac{c_1}{c} \approx 1.78$, we rewrite the obtained inequality in the form

$$\frac{n}{n+1}\ \ln^2 n + c_2\ln N \leq \ln^2 N.$$

Notice that

$$\begin{aligned}\ln^2 N - c_2\ln N &\leq \ln^2(n+1) + (2\ln 2 - c_2)\ln(n+1)\\ &\leq (\ln n + \frac{1}{n})^2 + (2\ln 2 - c_2)\ln(n+1).\end{aligned}$$

By joining these two inequalities, we obtain

$$(c_2 - 2\ln 2)\ln(n+1) \leq \frac{\ln^2 n}{n+1} + \frac{2\ln n}{n} + \frac{1}{n^2}\ .$$

However, for $n \geq 24$ the opposite inequality holds. Therefore, the graph G cannot serve as counterexample to (8). □

Acknowledgements

I thank Yuri Matiyasevich who proposed this problem to me.

References

[1] V.M. Glushkov, The abstract theory of automata, *Russian Math. Surveys*, **16** (1961), 1–53, translation from *Usp. Mat. Nauk* **16**, (1961) No.5(101), 3–62; Correction **17**, (1962) No.2(104), 270.

[2] Ch. Hagenah and A. Muscholl, Computing ϵ-free NFA from regular expressions in $O(nlog^2(n))$ time, *R.A.I.R.O. Theoretical Informatics and Applications*, **34** (2000), 257–277.

[3] J. Hromkovič, S. Seibert, and Th. Wilke, Translating regular expressions into small ϵ-free nondetermenistic finite automata, in *Proc. of the 14th Annual Symposium on Theoretical aspects of Computer Science (STACS'97, Lübeck, Germany)*, edited by R.Reischuk et al. Lecture Notes in Comput. Sci., **1200**, Springer (1997) 55–56.

Yuri Lifshits
Department of Algebra
Faculty of Mathematics and Mechanics
St-Petersburg State University
198504, Stary Peterhof
Bibliotechnaya pl., 2
Russia
lifts@mail.rcom.ru