Maria José Serna

Asymptotical behaviour of some non-uniform measures

<http://www.numdam.org/item?id=ITA_1989__23_3_281_0>

ASYMPTOTICAL BEHAVIOUR OF SOME NON-UNIFORM MEASURES

by Maria José SERNA ([1])

Communicated by J. DIAZ

Abstract. – *We show the existence of the Shannon effect for some well-known non-uniform measures such as regular complexity, initial index and context-free cost. We also obtain the asymptotical behaviour of the hardest finite languages for these measures.*

Résumé. – *On présente l'existence de l'effet Shannon pour certaines mesures non uniformes comme l'index initial et le coût grammatical.*

0. INTRODUCTION

An important characteristic in the study of non-uniform mesures is the existence of the Shannon effect; namely that "almost all problems of size n lying in a given class have almost identical complexity, which is asymptotically equal to the complexity of the hardest problem of size n" (the notion of the Shannon effect was introduced in [Lu, 70]). There has been a large number of papers dealing with the Shannon effect for various complexity measures in various models of computation [We, 87].

Recall that in 1958 O. B. Lupanov [Lu, 58] proved that every boolean function f of n variables has a boolean cost, denoted by $c(f)$, upper bounded above by

$$\frac{2^n}{n} + o\left(\frac{2^n}{n}\right)$$

Previously, C. E. Shannon and J. Riordan [Ri, Sha, 42], shown in 1942 that the cardinality of the set

$$\left\{ f \mid c(f) > \frac{2^n}{n} - o\left(\frac{2^n}{n}\right) \right\}$$

has an order of magnitude of 2^{2^n}. This result implies that the bound of Lupanov's theorem is tigth.

We conclude that the asymptotical behaviour of circuit size is $2^n/n$.

Theorems of this kind have been proved for other non-uniform measures. For instance, in the case of the boolean formula size, it has been proved that; $2^n/\log n$ holds asymptotically [Lu, 62]; G. Goodrich, R. Ladner and M. Fischer [Go, La, Fi, 77] proved the Shannon effect for some classes of straight line programs computing finite languages over an alphabet Σ of s symbols, the cost of a program being defined as the number of instructions.

M. R. Kramer and J. Van Leeuwen [Kr, Le, 84] proved similar results for the area measure of *VLSI*. More exactly they gave an upper bound of $O(2^n)$ and a lower bound of $\Omega(2^n)$ to the cost of a finite language.

Let us consider another approach to the description of finite languages. Given an alphabet Σ, let us consider the language $L \subset \Sigma^n$ in which all the words have length n. We can describe L as a binary string of length s^n. This string codifies the characteristic function of L. Thus, the complexity of L can be evaluated through the complexity of its associated string.

The first work in this direction was done by G. J. Chaitin [Ch, 66]. He considered Turning machines as the computational model. Given a Turing machine its cost is defined as the number of states. Chaitin proved that the cost is asymptotically $s^n/(k-1) n \log s$, where k is the number of tape symbols.

Similar work has been done by J. Berstel and S. Brelk [Be, Br, 87]. They used word chains as the computational model, the complexity being defined as the length of the chain. They proved that the cost of the hardest language oscillates between $s^n/n \log s$ and $s^n/n \log_3 (s)$.

A lot of research in this field has taken place in Russia, *see* for example [Ne, 65], [Ug, 76].

In this paper, we study non-uniform measures which have the Shannon effect. Theorems about upper and lower bounds are proved for *regular cost* [Er, Ze, 74], *initial index* [Ba, Di, Ga, 85] and *context-free cost* [Bu, Cu, Ma, Wo, 81].

Let Σ be an alphabet. For a string $w \in \Sigma^*$, let $|w|$ denote its length. Let Σ^n denote the set of all words of length n, and let the term "finite language" denote a subset of Σ^n for some $n > 0$.

Throughout this paper, N denotes the set of non-negative integers.

DEFINITION 0.1: The notion "almost all languages $L \subset \Sigma^n$ have property P" stands for the following asertion:

$$\text{Card}\{L \subset \Sigma^n \,|\, L \text{ has } P\}/\text{Card}(\Sigma^n) \to 1 \quad \text{as } n \to \infty.$$

DEFINITION 0.2: Given a complexity measure α, we denote by $S_\alpha(n)$ the complexity of the hardest problem in Σ^n, namely $\max\{\alpha(L) \,|\, L \subset \Sigma^n\}$.

DEFINITION 0.3: We say that the Shannon effect holds for a complexity measure α if $\alpha(L) \geqq S_\alpha(n) - o(S_\alpha(n))$ for almost all $L \in \Sigma^n$.

This definition is closed to the adopted in [We, 87]

1. REGULAR COMPLEXITY

We consider first a non-uniform measure obtained from regular expressions over the basis $\{+, .\}$. This is a particular case of various types of measures based on regular expressions, which were introduced by Ehrenfeucht and Zeiger [Er, Ze, 74].

Let E be a regular expression; let $\|E\|$ denote the number of operators " $+ .$ " in E. Given a language L we define the *regular cost* of L by:

$$\text{reg}(L) = \min\{\|E\| \,|\, E \text{ is a regular expression such that } L(E) = L\}$$

In order to give the upper bound, we first examine some results for languages with a bounded number of words.

LEMMA 1.1: *Let* k, $q \in N$ *and* $A \subset \Sigma^k$ *with* $\text{Card}(A) \leqq q$ *then,* $\text{reg}(A) \leqq k . q - 1$.

If we limit the number of words of the language we obtain

LEMMA 1.2: *Let* $n, k \in N$ *with* $0 \leqq k \leqq n$, $L \subset \Sigma^n$ *and* $B \subset \Sigma^k$ *with* $\text{Card}(B) \leqq q$, *then*

$$\text{reg}(L \cap (B . \Sigma^{n-k})) \leqq 2^q (kq + 1) + s^{n-k} (n - k).$$

Proof: Let $\Sigma^{n-k} = \{x_1, \ldots, x_t\}$, $t = s^{n-k}$, L can be written as

$$L = R_1 \cdot x_1 + \ldots + R_t \cdot x_t$$

with

$$R_i = \{w \in \Sigma^k \,|\, wx_i \in L\} \qquad \text{for} \quad i = 1 \ldots t$$

then $L \cap (B \cdot \Sigma^{n-k}) = (R_1 \cap B) \cdot x_1 + \ldots + (R_t \cap B) \cdot x_t$.

Note that as $\text{Card}(B) \leq q$, the number of subsets of B is less than or equal to 2^q. By grouping the terms with $R_i \cap B = R_j \cap B$ we have:

$$L \cap (B \cdot \Sigma^{n-k}) = B_1 \cdot A_1 + \ldots + B_p \cdot A_p \qquad \text{with} \quad p \leq 2^q.$$

As $B_i \subset B$, then $\text{Card}(B_i) \leq q$ and by lemma 1.1 we get

$$\text{reg}(B_i) \leq k \cdot q - 1 \qquad \text{for} \quad i = 1 \ldots p$$

where $A_1, \ldots, A_p$ is a partition of Σ^{n-k}; then

$$\Sigma \, \text{reg}(A_i) \leq (n-k) s^{n-k} - 1.$$

Therefore

$$\text{reg}(L \cap (B \cdot \Sigma^{n-k})) \leq \Sigma \, \text{reg}(B_i) + \Sigma \, \text{reg}(A_i) + 2^q + 2^q - 1$$

$$\leq 2^q (kq - 1) + (n-k) s^{n-k} - 1 + 2^q + 2^q - 1. \quad \blacksquare$$

Using the last lemma, we can establish the promised upper bound.

THEOREM 1.1: *Let Σ with $\text{Card}(\Sigma) = s$. For any $\varepsilon > 0$, for all sufficiently large n, and for any $L \subset \Sigma^n$ we have:*

$$\text{reg}(L) \leq (1 + \varepsilon) \frac{s^n}{\log s}.$$

Proof: Let k, $0 \leq k \leq n$, and p, $0 \leq p \leq s^k$, be parameter to be specified later. Also let $\Sigma^{n-k} = \{x_1, \ldots, x_t\}$ $t = s^{n-k}$.

Consider a partition of Σ^k into pairwise disjoint sets $P_1, \ldots, P_r$, $r = [s^k/p]$, each of cardinality less than or equal to p.

Then we have $\Sigma^n = \sum_{i=1 \ldots r} P_i \cdot \Sigma^{n-k}$ so $L = \sum_{i=1 \ldots r} L \cap (P_i \cdot \Sigma^{n-k})$

therefore:

$$\text{reg}(L) \leq r \cdot \max_{1 \leq i \leq r} \{\text{reg}(L \cap (P_i \cdot \Sigma^{n-k}))\} + r - 1.$$

From lemma 1.2, by letting $q = p$ we have:

$$\operatorname{reg}(L \cap (P_i \Sigma^{n-k})) \leq 2^p (kp+1) + s^{n-k} (n-k))$$

therefore

$$\operatorname{reg}(L) \leq \frac{s^n}{p}(n-k) + \frac{s^k}{p}(2^p(kp+1)+1).$$

By letting $k = \log_s n$ and $p = -2 \log n + n \log s$ we have that for all sufficiently large n:

$$\operatorname{reg}(L) \leq (1+\varepsilon)\frac{s^n}{\log s}. \quad \blacksquare$$

In the next lemma, we compute an upper bound for the number of regular expressions with bounded cost. This result is needed to prove the desired lower bound.

LEMMA 1.3: *For some constant c we have:*

Card $\{E \mid E$ *is a regular expression with* $\|E\| \leq R\} \leq (cs)^{R+1}$.

Proof: First we compute an upper bound for Card $\{E \mid \|E\| = r\}$.

A regular expression of cost r can be represented as a binary tree with r nodes and $r+1$ leaves. We can describe this tree by a string of length $2r+1$. For these $2r+1$ positions we have:

$$\text{positions to operators} +, \quad \binom{2r+1}{r}$$

$$\text{two operators in } r \text{ positions, } 2^r$$

$$s \text{ symbols in } r+1 \text{ positions, } s^{r+1}$$

Therefore,

$$\text{Card}\{E \mid \|E\| = r\} \leq 2^r \binom{2r+1}{r} s^{r+1} \leq 2s(8s)^r$$

and the sum over all possible values of r give us:

$$\text{Card}\{E \mid \|E\| \leq R\} \leq \Sigma \, 2s(8s)^r \leq (cs)^{R+1}$$

where c is a positive constant. $\quad \blacksquare$

We can get an expression for the lower bound:

THEOREM 1.2: *Let* Σ *with* Card $(\Sigma) = s$. *For any* $\varepsilon > 0$ *and for almost all* $L \subset \Sigma^n$ *we have* $\mathrm{reg}(L) > (1 - \varepsilon) s^n / \log s$.

Proof: We prove that the fraction of subsets L such that

$$\mathrm{reg}(L) \leqq (1 - \varepsilon) \frac{s^n}{\log s}$$

approaches 0 as $n \to \infty$.

From lemma 1.3, we have an upper bound for regular expression with cost R or less. We search for an R such that it is solution of the equation:

$$(2 s (8 s)^{R+1}) / 2^{s^n} \leqq 2^{-\delta s^n} \quad \text{for some} \quad \delta > 0$$

we obtain that

$$R = (1 - \varepsilon) \frac{s^n}{\log s}. \quad \blacksquare$$

And from theorem 1.1 and theorem 1.2 we have:

THEOREM 1.3: *Let* Σ *with* Card $(\Sigma) = s$. *The Shannon effect holds for the measure* reg, *and* $S_{\mathrm{reg}}(n)$ *is asymptotically equal to* $s^n / \log s$.

Theorem 1.2 has pleasant consequences. Recall the connection between the size of the regular expressions and the area of *VLSI* given in [F1, U11, 82]. Let $A(n)$ be the area alotted to a circuit for a regular expression of length n;

THEOREM 1.4: *There exist positive constants* d, e *and* f *such that for every regular expression of length* $n \geqq 2$, *we get* $A(n) \leqq dn - e \sqrt{n} - f$.

By considering theorem 1.2 and theorem 1.4 we obtain as corollary a theorem given in [Kr, Le, 84].

THEOREM 1.5: *Every boolean function of* n *variables can be computed by a VLSI circuit with area* $O(2^n)$.

We can extend the regular complexity to regular expressions over other bases, as in [Sto, 79]. For example to the bases: $\{ +, ., \cap \}$ or $\{ +, ., - \}$, where $-$ denotes complementation relative to Σ^n. The results obtained for both are the same; namely that every regular expression can be computed with $s^n / \log s$ connectives.

This means that bounds for extended regular expressions are in some sense independent of the basis.

Remark: Intuitively regular expressions seem to be close to boolean formula in the sense that in both measures the associated graph is a tree with operations at the nodes. It is somewhat surprising to obtain different bounds and prove that the regular cost is asymptotically more expensive than the formula size.

2. RATIONAL AND CONTEXT-FREE COMPLEXITIES

The second non-uniform measure which we consider is the initial index of finite automata. This measure was introduced in [Ga, 83].

Let A be a finite automaton. By $\|A\|$ we mean the number of transitions of A. Given a language L we define the *initial index* of L as:

$$a(L) = \min \{ \|A\| \mid A \text{ is a non-deterministic automaton with } L(A) = L \}.$$

To prove the upper bound, we first construct an automaton which has one accepting state for each word in Σ^k. This construction can be made recursively from corresponding automaton for Σ^{k-1}. Then we have:

LEMMA 2.1: *There exist an NFA of complexity s^{k+1} which for each word of Σ^k has a definite accepting state.*

Consider now a language with a limited number of words, then in the following lemma, we derive a cost's upper bound to the automaton which recognizes any subset of this language.

LEMMA 2.2: *Let $A \subset \Sigma^k$ with $\text{Card}(A) = q$, there exist an NFA of complexity less than or equal to $s^{k+1} + q 2^{q-1}$ wich for each subset of A has a definite accepting state.*

Proof: We consider the automaton of the figure.

This automaton is obtained from the automaton constructed in the previous lemma. We add one state for each A_j, subset of A, and λ-transitions to connect each word of A_j with $A_{j'}$.

Then the cost of computing all subsets of A is:

$$
\begin{array}{l}
1 \text{ for subsets of cardinality } 1 \\
2 \text{ for subsets of cardinality } 2 \\
. \text{ for subsets of cardinality } . \\
. \text{ for subsets of cardinality } . \\
. \text{ for subsets of cardinality } . \\
q \text{ for subsets of cardinality } q
\end{array}
$$
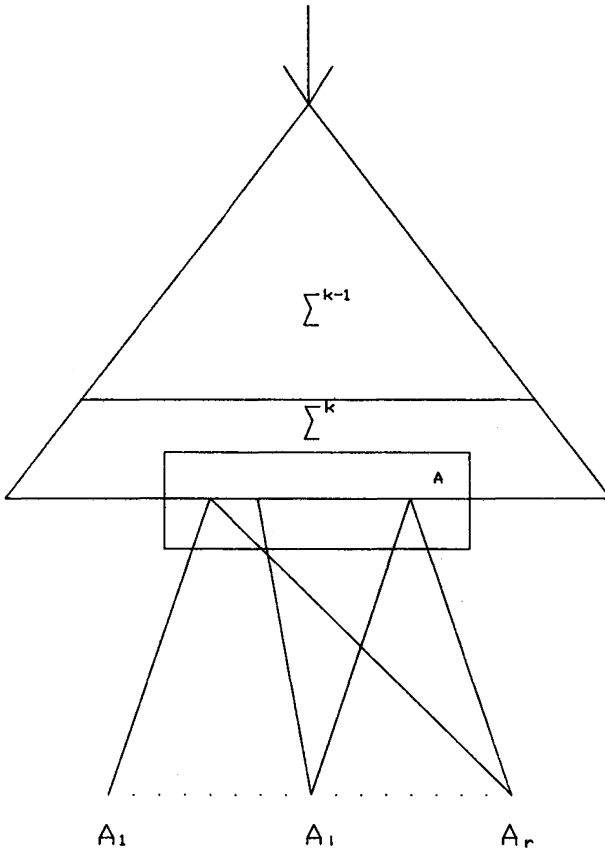
and then less than or equal to:

$$q+2\binom{q}{2}+\ldots+q\binom{q}{q}=q+q\binom{q-1}{1}+\ldots+q\binom{q-1}{q-1}=q\,2^{q-1}.\quad\blacksquare$$

We shall use the above lemma to obtain the upper bound for the initial index;

THEOREM 2.1: *Let Σ with Card $(\Sigma)=s$. For any $\varepsilon>0$, for all sufficiently large n, and for any $L\subset\Sigma^n$ we have:*

$$a(L)\leqq(1+\varepsilon)\frac{s^n}{n\log s}.$$

Proof: Let k, $0 \leq k \leq n$ and p, $0 \leq p \leq s^k$, be parameters to be specified later. Let $\Sigma^{n-k} = \{x_1, \ldots, x_t\}$, $t = s^{n-k}$. Then L can be written as:

$$L = R_1 x_1 + \ldots + R_t x_t$$

with

$$R_i = \{w \in \Sigma^k \mid wx_i \in L\} \quad \text{for} \quad i = 1, \ldots, t.$$

Consider a partition $P_1, \ldots, P_r$, of Σ^k, with $r = [s^k/p]$, and such that the cardinality of every P_i is at most p.

We have $R_i = R_i \cap P_1 + \ldots + R_i \cap P_r$ and we can construct the automaton of the figure.

By techniques used in lemma 2.2, we can construct an automaton which has an accepting state for each subset of P_1. We do this construction for each P_i. Notice that as the underlying automaton for Σ^k is not replicated, the total cost is bounded by $r \cdot p \cdot 2^{p-1} + s^{k+1}$.

In the next step we construct the sets R_i from de sets $R_i \cap P_j$ constructed precedently. This is done by adding λ-transitions. Then the number of λ-transitions increase in at most $r \cdot s^{n-k}$.

Finally we need to connect the automaton constructed in the previous step with the automaton for Σ^{n-k}, described in lemma 2.1. Then we have:

$$a(L) \leq s^{k+1} + (p+1) 2^p \frac{s^k}{p} + \frac{s^n}{p} + (s+1) s^{n-k}.$$

Letting $k = [2 \log_s n]$ and $p = [(\log s)(n - 4 \log_s n)]$ we get $a(L) \leq (1 + \varepsilon) s^n/n \log s$ for all n sufficiently large. ∎

To prove the lower bound, we must first compute an upper bound for the number of automata with bounded cost.

LEMMA 2.3: *There exists a positive constant c such that*

$$\text{Card}\{A \mid A \text{ is an NFA with } \|A\| \leq R\} \leq (cR)^R.$$

Proof: First we compute the number of automata with q states and $\|A\| = r$. We can describe an automaton of this type as a string of length $2r$, which is formed by r symbols of Σ and r states. Then we have that the total number of such automatas is less than or equal to
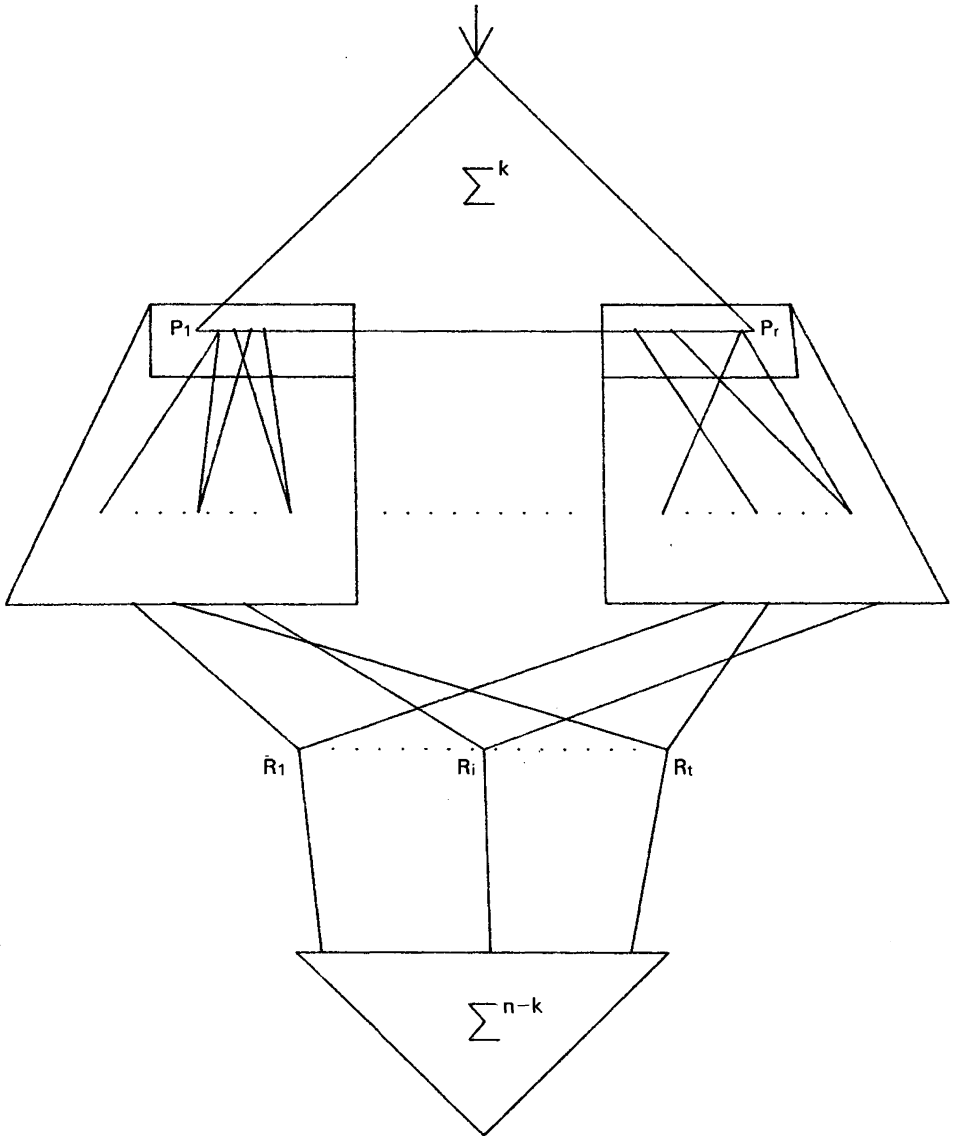
$$\binom{2r}{q} q^r s^r,$$

So the total number of automatas with cost r, is less than or equal to:

$$\sum \binom{2r}{q} q^r s^r \leqq s^r r^r \sum \binom{2r}{q} \leqq s^r r^r 2^{2r} \leqq (4sr)^r.$$

It follows that $(cR)^R$ is an upper bound to the number of automata, with $\|A\| \leq R$. ∎

Using the same techniques developed in [Go, La, Fi, 77], and applying the previous lemma, we prove the following result which give us the lower bound to the initial index.

THEOREM 2.2: *Let Σ with* Card $(\Sigma) = s$. *For any $\varepsilon > 0$ $a(L) > (1 - \varepsilon) s^n / n \log s$ for almost all $L \subset \Sigma^n$.*

From this last theorem it follows in a straightforward manner that:

THEOREM 2.3: *Let Σ with* Card $(\Sigma) = s$. *The Shannon effect holds for the measure a and $S_a(n)$ is asymptotically equal to $s^n / n \log s$.*

Let us turn into non-uniform measure based on context-free grammars.

Let G be a context-free grammar, and let $\|G\|$ denote the sum of the lengths of all the right terms in the rules of G.

Given a language L we define the *context-free cost* of L as:

$$\mathrm{cf}(L) = \min \{ \|G\| \,|\, G \text{ is a context-free grammar such that } L(G) = L \}.$$

This measure was introduced in [Bu, Cu, Ma, Wo, 81].

Note from the definition it follows that for all languages $L \subset \Sigma^n$ we have $a(L) \leq 2 \,\mathrm{cf}(L)$. An easy way to derive an upper bound for context-free cost is to construct the grammar associated to the automaton given in theorem 2.1. Then we have:

THEOREM 2.4: *Let Σ with* Card $(\Sigma) = s$. *For any $\varepsilon > 0$, for all sufficiently large n, and for any $L \subset \Sigma^n$ we have:*

$$\mathrm{cf}(L) \leq (1 + \varepsilon) \frac{s^n}{n \log s}.$$

The next lemma, give us an upper bound to the number of grammars with bounded cost.

LEMMA 2.4: *There exists a positive constant c such that* Card$(\{ G \,|\, G$ *is a context-free grammar with* $\|G\| \leq R \}) \leq (cR)^R$.

Proof: First we compute the number of grammars, with m productions and q variables, which have size r. We describe these grammars by the concatenation of all rigth rules. The total number of grammars is at most:

$$\binom{r}{m} \binom{m}{q} (s + q)^r.$$

Therefore an upper bound to the number of grammars with cost r and m productions, is

$$\Sigma \binom{r}{m} \binom{m}{q} (s+q)^r \leq \binom{r}{m} (s+m)^{r2^m}$$

and an upper bound to the number of grammars with $\|G\| = r$, is

$$\Sigma \binom{r}{m} (s+m)^{r2^m} \leq (3(s+r))^r.$$

Therefore $(cR)^R$ is an upper bound to the number of grammars with $\|G\| \leq R$. ∎

By the same considerations made in theorem 2.2, we can state the following results:

THEOREM 2.5: *Let Σ with* Card $(\Sigma) = s$. *For any $\varepsilon > 0$ and for almost all $L \subset \Sigma^n$ we have:* cf $(L) > (1-\varepsilon) s^n/n \log s$.

THEOREM 2.6: *Let Σ with* Card $(\Sigma) = s$. *The Shannon effect holds for the measure* cf *and $S_{cf}(n)$ is asymptotically equal to $s^n/n \log s$.*

The relationship between initial index, context-free cost and circuit size has been studied in [Ba, Di, Ga, 85]. In that paper it is proved that initial index is less powerful than context-free cost, and that context-free cost is less powerful than circuit size. Less powerful means that exist languages with succint descriptions in one measure that blow up in the other. It is interesting to note that regarding their Shannon's functions, all these measures are in some sense equivalent.

The existence of languages with cost near to optimal is established in the following publications; for regular expressions in [Eh, Ze, 74]; for initial index in [Ga, 83] and for context-free cost in [Go, La, Fi, 77].

REFERENCES

[Ba, Di, Ga, 85] J. L. BALCAZAR, J. DIAZ and J. GABARRO, *Uniform Characterizations of Non-Uniform Measures*, Inf. and Cont., Vol. 67, 1985, pp. 53-69.

[Be, Br] J. BERSTEL and S. BRLEK, *On the Length of Word Chains*, I.P.L., Vol. 26, 1987, pp. 23-28.

[Bu, Cu, Ma, Wo, 81] W. BUCHER, K. CULIK, H. MAURER and D. WOTSCHKE, *Concise Description of Finite Languages*, Theor. Comp. Sci., Vol. 14, 1981, pp. 227-246.

[Ch, 66] G. J. CHAITIN, *On the Length of Programs for Compute Finite Binary Sequences*, Jour. A.C.M., Vol. 13, No. 4, 1966, pp. 547-569.

[Eh, Ze, 74] A. EHRENFEUCHT and P. ZEIGER, *Complexity Measures for Regular Expressions*, J. Comp. and Sys. Sci., Vol. 12, No. 2, 1976, pp. 134-146.

[Fl, Ull, 82] R. FLOYD and J. ULLMAN, *The Compilation of Regular Expressions into Integrated Circuits*, Jour. A.C.M., Vol. 29, No. 2, 1982, pp. 603-622.

[Ga, 83] J. GABARRO, *Initial Index: a New Complexity Function for Languages*, I.C.A.L.P. 83 L.N.C.S. 154, Springer-Verlag, 1983, pp. 226-236.

[Go, La, Fi, 77] G. GOODRICH, R. LADNER and M. FISHER, *Straight Line Programs to Compute Finite Languages*, Conf. on Theor. Comp. Sci., Waterloo, 1977.

[Kr, Le, 84] M. R. KRAMER and J. VAN LEEUWEN, *The VLSI Complexity of Boolean Functions*, L.N.C.S. 171, Springer-Verlag, 1984, pp. 397-407.

[Lu, 58] O. B. LUPANOV, *A Method of Circuit Synthesis*, Izv. V.U.Z. Radiofiz, Vol. 1, 1958, pp. 120-140.

[Lu, 62] O. B. LUPANOV, *Complexity of Formula Realization of Functions of Logical Algebra*, Prob. Kibern., Vol. 3, 1962, pp. 782-811.

[Lu, 70] O. B. LUPANOV, *On Circuits of Functional Elements with Delay*, Problems of Cybernetics [in Russian], Vol. 23, 1970, pp. 43-81.

[Ne, 65] E. I. NECHIPORUK, *On the Design of Logical Networks in Incomplete and Degenerate Bases*, Problems of Cybernetics [in Russian], Vol. 14, 1965, pp. 111-160.

[Ri, Sha, 42] J. RIORDAN and C. E. SHANNON, *The Number of Two-Terminals Series Parallel Networks*, J. on Math. Phys., Vol. 21, 1942, pp. 83-93.

[Sto, 79] L. J. STOCKMEYER, *Classifying the computational Complexity of Problems*, M.I.T. Research Report R.C. 7606, 1979.

[Ug, 76] A. B. UGOLNIKOV, *On the Realization of Monotone Functions by Circuits of Functional Elements*, Problems of Cybernetics [in Russian], Vol. 31, 1976, pp. 167-185.

[We, 87] I. WEGENER, *The Complexity of Boolean Functions*, Wiley-Teubner, 1987.