Cartesian and statistical approches of the satisfiability problem

Israël-César Lerman

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

1992
25ème
anniversaire

N° 1685

Programme 3
*Intelligence artificielle, Systèmes cognitifs et
Interaction homme-machine*

CARTESIAN AND STATISTICAL APPROACHES OF THE SATISFIABILITY PROBLEM

Israël-César LERMAN

Mai 1992

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTEMES ALEATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CEDEX FRANCE
Tél. : 99 84 71 00 - Télex : UNIRISA 950 473 F

CARTESIAN AND STATISTICAL APPROACHES OF THE SATISFIABILITY PROBLEM

I.C. LERMAN
IRISA
Campus Universitaire de Beaulieu
35042 RENNES CEDEX FRANCE

Publication Interne n° 643 - Mars 1992, 58 pages, Programme 3

SUMMARY

An english new version of [Lerman 91_b] is proposed here. The present text is more accurate and more complete by including new points. On the other hand, its conclusion is more decisive.

This work is devoted to a new and systematic treatment of the different questions arised in the original study [Simon & Dubois 1989]. We consider both formal and statistical aspects with a set theoretic and geometrical representation. On the other hand we adopt the combinatoric and statistical point of view which is usual in our approach of data classification. Our synthetic formulation makes clearer and improves the algorithms or calculations, previously considered (cf. above reference). On the other hand and mainly, we propose new algorithms and new calculations which enable to "see" the limit of the complexity reduction that we can expect. For this respect, we essentially distinguish the problem of the evaluation of the number of solutions and that one of the recognition of the satisfiability instances. Two cases are studied concerning real observed and random systems of clauses. In our formalization we represent a clause by a logical and geometrical object that we call a "pinpoint cylinder". On the other hand, the "inclusion and exclusion" formula plays an important rôle in our evaluations. The new algorithms that we present take into account the marginal statistical distributions of the variables. On the other hand, we introduce in these algorithms parallel procedures and -in a relevant way- our approach in data classification. The latter can play an important role in complexity reduction of a SAT problem. In each of both aspects : evaluation of the number of solutions and recognition of the satisfiability, significant results are obtained in the context of the generation of a random system of clauses. The randommess is according to a model that we usually consider in our approach of data classification, for measuring associations (between "pinpoint cylinders", here) ; and that we call, "hypothesis of no relation".

Key word : NP complete ; Logical exclusivity and statistical independence between clauses ; Statistical estimation ; Computing complexity ; Clustering.

APPROCHES CARTESIENNE ET STATISTIQUE DU PROBLEME DE LA SATISFIABILITE

RESUME

Ce rapport reprend dans une version anglaise un précédent rapport [Lerman 91_b]. Les résultats que nous présentons ici sont toutefois plus précis et plus complets, par l'apport de points nouveaux. D'autre part, la conclusion a un caractère plus décisif.

Rappelons que dans ce travail nous reconsidérons de façon systématique l'étude originale de J.C. Simon et O. Dubois [Simon & Dubois 1989] du problème SAT ; et ce, aussi bien dans ses aspects formels que statistiques. Nous apportons un traitement de la question à partir d'une représentation d'un type ensembliste et géométrique, en adoptant une approche combinatoire et statistique qui nous est usuelle en classification automatique. Ainsi, à partir d'une vision synthétique, nous reformulons, avec des apports nouveaux, ou bien une clarification du sens des résultats, les algorithmes ou calculs déjà exprimés dans la référence ci-dessus. Mais aussi et surtout, nous proposons de nouvelles algorithmiques et nous effectuons des calculs originaux qui permettent de "voir" les limites de la réduction de la complexité qu'on peut espérer. A cette fin, qu'il s'agisse d'un système réel observé, ou bien résultant d'une génération aléatoire, de clauses, nous distinguons de façon essentielle, le problème de l'évaluation du nombre de solutions, de celui de la reconnaissance de la satisfiabilité. Pour pouvoir entreprendre notre approche, nous sommes conduits à représenter une clause par un "cylindre logique ponctuel" et à faire jouer un rôle déterminant à la formule "d'inclusion et d'exclusion". Les nouveaux algorithmes que nous présentons tiennent compte des caractéristiques statistiques marginales de la distribution des variables. Nous y introduisons le parallélisme et, de façon pertinente, notre approche de la classification automatique qui peut jouer un rôle important dans la réduction de la complexité d'un problème SAT. Sur chacun des deux aspects : évaluation et reconnaissance, des résultats significatifs sont obtenus, dans le cadre d'un système aléatoire de clauses, conformément à un modèle que nous avons coutume de considérer dans notre approche pour l'évaluation des liens (ici entre "cylindres ponctuels"), en classification, sous l'appellation "hypothèse d'absence de liaison".

Mots-clés. NP-complet ; Exclusivité logique et indépendante statistique entre clauses ; Estimation statistique ; Complexité calcul ; Classification automatique.

I. INTRODUCTION

This work is devoted to a new and systematic treatment of the different questions arised in the original paper of J.C. Simon and O. Dubois [Simon & Dubois 1989] on the SAT problem ([Cook 1971,1983], [Garey & Johnson 1979]). Our synthetic formulation makes clearer and improves the algorithms or calculations previously considered (cf. above reference). On the other hand and mainly, we propose new algorithms and new calculations which enable to "see" the limit of the complexity reduction that we can expect.

Two general aspects are developed in our study ; the former, which concerns the case of a real observed system of clauses, is purely formal and algorithmic. The latter, which concerns a random system of clauses (introduction of a random model to generate a sequence of clauses), is of combinatorial nature [see section III for the first aspect and section IV for the second aspect].

Whatever is the case considered, faced with satisfiability instances, we very clearly distinguish the two following NP problems:

(i) exact, approximate or estimated evaluation of the number of solutions ;
(ii) recognition of the satisfiability.

Obviously, an answer of non statistical nature to the first problem (i) can provide an algorithm to the second one (ii). But, it may exist a resolution algorithm for the second problem (ii) which cannot be relevant for the first one.

Relative to the generation of a sequence of random clauses, the previous fundamuntal conceptual distinction between (i) and (ii), makes clear the difference between the experimental verification and the theoretical expected value of the number of clauses, from which the system becomes contradictory [see section 3.3. of Simon & Dubois 1989]. In fact -as it is expressed by the authors, the latter theoretical value is incorrect ; but, the evaluation of the average number of solutions is correct. In this framework, we obtain significant results by considering different forms of the random model of clause generation. As a matter of fact, we have introduced the same type of model in our approach of data classification [Lerman 1981, 1991a], and that, in order to significantly evaluate the associations between qualitative attributes.

For the both aspects : formal (see section III) and statistical (see section IV), we will consider (i) first and (ii) afterwards. On the other hand, we adopt in our combinatorial approach a set theoretic and geometrical representation. In fact, we will work at the level of the logical cube $\{0,1\}^n$, where n is the number of boolean variables. In these conditions, we associate to a given r-clause C, comprising r instanciated variables (i.e. where binary values 1 or 0 are assigned to exactly r among the n variables), its anti-clause that we denote by C and that we represent in $\{0,1\}^n$ by what we call a "pinpoint cylinder of order r" (see section II below). The proposed evaluations and algorithms can take into account the geometrical structure of the latter mathematical object. In this context, the "inclusion and exclusion formula" will play an important role.

The random model of clause generaion mentionned above, corresponds to a hypothesis of "no relation" or "independence" in the probabilistic sense. This model will intervene with very different points of view at two levels. First, in the context of a real observed system of clauses, where we show the pertinent role of our classification method, to notably reduce the computational complexity, for determining -exactly or approximatively- the number of solutions. In this method, the association coefficient between the structures to be compared ("pinpoint cylindres", here) is of probabilistic nature. It is established with respect to a probabilistic hypothesis of independence. In fact, the latter provides a probability scale to significantly measure the associations or similarities.

Secondly (see section IV), the random model will intervene to generate a system of independent -in probabilistic sense- clauses. Relative to the latter random system of clauses, the purpose is then to study the problems (i) and (ii) mentionned above.

In this latter study the notion of independence in probability between pinpoints cylinders -respectively associated to random clauses- will play a very fundamental role. In these conditions, we must emphasize that the notion considered in [Simon & Dubois 1989] of "independent" clauses, corresponds exactly to that one of disjunction or exclusion -in a set theoretic meaning- between the pinpoints cylinders, respectively associated (see section II). Then, we will reserve the term "independence" exclusively to its probabilistic sense.

Let us now make clear the subjects treated in the following paragraphs. In section II, we will precise the basic notions and our mathematical representation of the two fundamental problems set [see (i) and (ii) above]. These two problems are analyzed in section III, from formal and algorithmic computational points of view, in a real observed case. But, the random case is consistently studied in section IV. The last section V is devoted to a conclusion where we will try to give -on the basis of this work and our experiment in combinatorial data classification -an evaluation of the reduction possibility of the computational complexity to treat a given NP problem.

Let us indicate that a new version of [Lerman 91_b] is proposed here. The present text is more accurate and more complete by including new points. On the other hand, its conclusion is more decisive.

II. THE BASIC NOTIONS ; SETTING UP THE TWO PROBLEMS

II.1. Pinpoint cylinder associated to a clause

Let $X=\{x_1,x_2,...,x_n\}$ be a set of n booelan variables and let $\tilde{X}=\{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n\}$ be the set of the complemented variables ($\tilde{x}_i=1-x_i$, $1\leq i\leq n$). Let us recall that a clause is a disjunction of variables belonging to $X \cup \tilde{X}$; where, for each i, $1\leq i\leq n$, we have exactly one of the three following exclusive cases :

(1) only x_i appears ;
(2) only $\tilde{x}_i$ appears ;
(3) neither x_i nor $\tilde{x}_i$ are present.

Thus, a formula representing a clause, involves at most, n literals, respectively separated by the disjunction sign v. For example, by supposing n greater than 4, a clause C^3 of order 3, can be :

$$C^3 = x_1 \vee \tilde{x}_3 \vee x_4$$

More generally, we denote a clause C^r of order r, under the following form :

$$C^r = y_{i1} \vee y_{i2} \vee ... \vee y_{ir} , \qquad (1)$$

where $r\leq n$, where -without loss of generality- $1\leq i_1\leq i_2\leq ...\leq i_r\leq n$ and where $y_{ip} = x_{ip}$ or (exclusively) $y_i = \tilde{x}_i$, $1\leq p\leq r$.

We associate to C^r, its opposite, the anti-clause $\tilde{C}^r$:

$$\tilde{C}^r=\tilde{y}_{i1}\wedge\tilde{y}_{i2}\wedge\cdots\wedge\tilde{y}_{ir} \qquad (2)$$

where $\wedge$ denotes the conjunction and where $\tilde{y}_{ip} = \tilde{x}_{ip}$ (resp.x_{ip}) if $y_{ip} = x_{ip}$ (resp. $\tilde{x}_{ip}$), $1\leq p\leq r$.

Then, we identify $\tilde{C}^r$ with the set of the points of the logical cube $\{0,1\}^n$, which satisfy the formula defined by (2). We will denote by $E(\tilde{C}^r)$ -or more simply E in case of non ambiguity- this subset of $\{0,1\}^n$. E defines a cylinder in $\{0,1\}^n$, of which the basis is a single point in a given cartesian subspace. This is the reason why we call E, a "pinpoint cylinder of order r". More precisely, for $\tilde{C}^r$ [see (2) above], the cartesian subspace is generated by the components $i_1,i_2,...,i_r$; and the basis of $E(\tilde{C}^r)$ is the point $(\alpha_{i1} ,\alpha_{i2},...,\alpha_{ir}$), where $\alpha_{ip}=1$ (resp. 0) if $y_{ip} = \tilde{x}_{ip}$ (resp. x_{ip}), $1\leq p\leq r$. In other words, E is the set of points of $\{0,1\}^n$ for which, the sequence of the values of the components is $(\alpha_{i1},\alpha_{i2},...,\alpha_{ir})$. In these conditions, we will denote :

$$E\{(i_1,i_2,...,i_r),(\alpha_{i_1},\alpha_{i_2},...,\alpha_{i_r})\} \qquad (3)$$

In the previous exemple concerning the above clause C^3, we have

$$E=\{(1,3,4),(0,1,0)\} \qquad (4)$$

To be more explicit, if we suppose n=8, we may denote

$$E=\{(0,\varepsilon,1,0,\varepsilon,\varepsilon,\varepsilon,\varepsilon)\}, \qquad (5)$$

where ε is an undetermined boolean, belonging to {0,1}.

The volume (i.e. number of points) of a r-pinpoint cylinder [see (3) above] is equal to 2^{n-r} :

$$\mathrm{vol}[E(\tilde{C}^r]=2^{n-r} \qquad (6)$$

Thus, the volume of E -defined by (4) or (5) above- is equal to 2^5.

Let now E_1 and E_2 be two pinpoints cylinders having, repectively, the orders r_1 and r_2 :

$$E_1=\{(i_1,i_2,\ldots,i_{r1}),(\alpha_{i1},\alpha_{i2},\ldots,\alpha_{ir1})\} \quad (7)$$

$$E_2=\{(j_1,j_2,\ldots,j_{r2}),(\beta_{j1},\beta_{j2},\ldots,\beta_{jr2})\} \quad (8)$$

We have the following properties :

Properties :

(a) $E_{12}=E_1 \cap E_2$ is a pinpoint cylinder.

(b) $E_{12}=\phi$ and $\{h_1,h_2,\ldots,h_s\}=\{i_1,i_2,\ldots,i_{r_1}\} \cap \{j_1,j_2,\ldots,j_{r2}\}$

$$\Rightarrow \alpha_h = \beta_h \ , \ 1 \leq u \leq s \qquad (9)$$

and $E_{12}=\{(k_1,k_2,\ldots,k_{r1+r2-s}),(\gamma_{k1},\gamma_{k2},\ldots,\gamma_{r1+r2-s})\}$ (10)

where $(k_1,k_2,\ldots,k_{r1+r2-s})$ is the strictly increasing sequence of the subscripts of the union set

$$\{i_1,i_2,\ldots,i_{r1}\} \cup \{j_1,j_2,\ldots,j_{r2}\} \qquad (11)$$

and where, for a given k_v, we necessarily have one of the three exclusive cases :

(i) k_v is an h_u [cf. (9)] and then, $\gamma_{kv} = \alpha_{kv} = \beta_{kv}$;
(ii) k_v is an i_p but non a j_q [cf. (9)] and then, $\gamma_{kv} = \alpha_{kv}$;
(iii) k_v is a j_q but non an i_p [cf. (9)] and then, $\gamma_{kv} = \beta_{kv}$;
$1 \leq v \leq r_1+r_2-s$.

(c) $\mathrm{vol}(E_{12})=2^{n-r_1-r_2+s}$ (12)

These properties are easy to see. On the other hand, if $i_1<i_2<\ldots<i_{r1}$ and $j_1<j_2<\ldots<j_{r2}$, the maximum number of comparisons to sort (11), for establishing $(k_1,k_2;\ldots,k_{r1+r2-s})$ is max (r_1,r_2).

II.2. Logical "independence" between clauses

According to [Simon & Dubois 1989], two clauses C and C′ are "independent" iff no assignment of the n variables contradicts both clauses. Since, the contradiction of a given clause C is equivalent to the statisfiability to the anti-clause $\tilde{C}$, the previous notion of "independence" corresponds exactly to the disjunction in the set theoretic sense, between the two pinpoint s cylinders $E(\tilde{C})$ and $E(\tilde{C}')$:

logical "independence" between C and C′ <=> $E(\tilde{C}) \cap E(\tilde{C}')=\phi$(13)

The condition (13) is equivalent to the following :

$$(\exists i,\ 1\leqslant i\leqslant n);(E(\tilde{C})\subset\{i,1\})\&(E(\tilde{C}')\subset\{i,0\}) \text{ or } (E(\tilde{C})\subset\{i,0\})\&(E(\tilde{C}')\subset\{i,1\}). \qquad (14)$$

More precisely, the two pinpoints cylinders of order 1, $\{i,1\}$ and $\{i,0\}$ are two coordinate hypeplans, which partition the space $\{0,1\}^n$ into two complementary subspaces :

$$\{0,1\}^n=\{i,1\}+[i,0\}, \qquad (15)$$

where the sum is of set theoretic nature and also, corresponds to the direct sum between subspaces.

According to above, the logical "dependence" between two clauses C and C′ can be expressed by :

$$E(\tilde{C}) \cap E(\tilde{C}')\neq\phi \qquad (16)$$

A particular case of logical "dependence" between clauses corresponds to the implication

$$C\Rightarrow C' \qquad (17)$$

which expresses that every literal which is present in C, also appears in C′, under the same form (positive or negative). The relation (17) can be translated as follows :

$$E(\tilde{C})\supset E(\tilde{C}') \qquad (18)$$

In fact -providing $\{0,1\}^n$ by a probability measure P- the previous logical "independence" notion between two clauses C and C′, corresponds to a complete probabilistic dependence between the two pinpoints cylinders $E(\tilde{C})$ and $E(\tilde{C}')$ respectively associated. Effectively, in the latter case, we have for the conditional probabilities

$$P[E(\tilde{C})/E(\tilde{C}')]=P[E(\tilde{C}')/E(\tilde{C})]=0 \qquad (19)$$

Then -as mentionned in the introduction- henceforth, the term independence will be devoted to its probabilistic sense, whereas the previous logical sens of independence between two clauses will be expressed by the equivalent notion of exclusion or disjunction between the pinpoints cylinders associated [compare with (13) above].

II.3. The inclusion and exclusion formula

Let Ω be a finite set of elements and let $\{E_j/1\leqslant j\leqslant k\}$ a set of k non empty subsets of Ω. Let us recall that the inclusion and exclusion formula enables to determine the cardinal of the union

$$U\{E_j/1\leqslant j\leqslant k\}$$

of all the parts E_j of Ω, as a function of the intersections q by q, $1\leqslant q\leqslant k$, of the different sets E_j. More precisely, we have

$$\begin{aligned}
\mathrm{card}(\bigcup_{1\leqslant j\leqslant k} E_j) &= \sum_j \mathrm{card}(E_j) - \Sigma\left\{\mathrm{card}(E_{j_1} \cap E_{j_2}) / \{j_1, j_2\}\right\} \\
&\ldots + (-1)^{2p}\Sigma\left\{\mathrm{card}(E_{j_1} \cap \ldots \cap E_{j_{2p-1}}) / \{j_1, \ldots, j_{2p-1}\}\right\} \\
&+ (-1)^{2p+1}\Sigma\left\{\mathrm{card}(E_{j_1} \ldots \ldots \ldots E_{j_{2p}}) / \{j_1, \ldots, j_{2p}\}\right\} \\
&+ \ldots + (-1)^{k+1}\mathrm{card}(E_1 \cap E_2 \cap \ldots \cap E_k). \quad (20)
\end{aligned}$$

In this formula, an expression as $\{j_1,\ldots,j_q\}$ which indexes a given sum, denotes a generic element of the set $P_q(\{1,2,\ldots,k\})$; the latter being defined as the set of all subsets of $\{1,2,\ldots,k\}$, of which the cardinality is q, $1\leqslant q\leqslant k$. The cardinal of $P_q(\{1,2,\ldots,k\})$ is the binomial coefficient $\binom{k}{q}$, $1\leqslant q\leqslant k$.

Let us denote by S(q) the expansion of the second member of (20) up to the q^{th} term, which concerns the intersections q by q of the E_j, $1\leqslant q\leqslant k$. In the above formula (20) we have clearly expressed the two consecutive $(2p-1)^{th}$ and $2p^{th}$ terms. The former is positive and the latter, negative.

Let us now introduce the notion of elementary class of order h $(h\leqslant k)$ with respect to the set $\{E_j/1\leqslant j\leqslant k\}$ of parts of Ω. A such class is defined by the following expression :

$$\begin{aligned}
C(j_1,\ldots,j_h) &= (E_j \cap \ldots \cap E_{jh}) \cap (E_{jh+1} \cup \ldots \cup E_{jk})^c \\
&= (E_{j1} \cap \ldots \cap E_{jh}) \cap (E_{jh+1}{}^c \cap \ldots \cap E_{jk}{}^c), \quad (21)
\end{aligned}$$

where $\{j_1,j_2,\ldots,j_h\}$ is a subset of h elements of $\{1,2,\ldots,k\}$ and where $\{j_{h+1},j_{h+2},\ldots,j_k\}$ is the complementary subset in $\{1,2,\ldots,k\}$. Notice [cf. (20)} that we have denoted by X^c the complementary subset of X in $U\{E_j/1\leqslant j\leqslant k\}$.

Denoting by $\nu(q,h)$ the number of times where the cardinal of a given $C(j_1,\ldots,j_h)$ is counted in $S(q)$. We can prove that

$$\nu(2p-1,h)=1+\binom{h-1}{2p-1}$$
$$\nu(2p,h)=1-\binom{h-1}{2p} \qquad (22)$$

where the binomial coefficient $\binom{u}{v}$ is null if the integer u is strictly less than the integer v. We have

$$\nu(2p-1,h)-\nu(2p,h)=\binom{h}{2p} \qquad (23)$$

and we conclude

$$S(2p)\leqslant \mathrm{card}(\bigcup_{1\leqslant j\leqslant k} E_j)\leqslant S(2p-1) \qquad (24)$$

By (23) we see that, for fixed h, the difference between $\nu(2p-1,h)$ and $\nu(2p,h)$ is decreasing with respect to p, for p greater than h/4. Therefore, the higher is the value of p, the more accurate is the interval approximation given by (24).

Otherwise and in general, the higher is h, the more we can expect a low value of the cardinality of an elementary class of order h. This property -intuitively expressed- will be quantified in section IV where we will introduce an random model.

If the E_j, $1\leqslant j\leqslant k$, are pinpoints cylinders of the logical cube $\{0,1\}^n$ [see section II.1 above] ; then, the intersection

$$E_{j1} \bigcap E_{j2} \bigcap \ldots \bigcap E_{jq} \qquad (25)$$

of q among them, is also a pinpoint cylinder. Assuming each pinpoint cylinder associated to a clause, the latter pinpoint cylinder [cf. (25)] is empty iff there exists at least one variable presented under two opposite forms (x_i and $\tilde{x}_i$) in two distinct clauses. If not and if t is the total number of variables which appear in at least one clause, we have

$$\mathrm{card}(E_{j1} \bigcap E_{j2} \bigcap \ldots \bigcap E_{jq})=2^{n-t} \qquad (26)$$

We may notice that the last term of (20) above is necessarily equal to 0 or to $(-1)^{k+1}$. As a matter of fact, each of the n variables appears at least once [under its positive or (non exclusively) negative forms (x_i or $\tilde{x}_i$, $1\leqslant i\leqslant n$)] in the different clauses ; otherwise, the parameter n of the problem is greater than necessary.

II. 4. Setting up the two problems

Given a system of k clauses $\{C_j/1\leqslant j\leqslant k\}$, consider the logical following expression :

$$C_1\wedge C_2\wedge\cdots\wedge C_j\wedge\cdots\wedge C_k \qquad (27)$$

A <u>solution</u> is an assignation of the n boolean variables of $X=\{x_1,x_2,\ldots,x_i,\ldots,x_n\}$ (cf. section II.1. for which the expression (26) is true or <u>satisfiable.</u>

As we have already announced in the introduction (cf. section I), the two following problems will be studied, with our new formulation:

(i) number N of solutions ;
(ii) existence of a solution.

According to section II.1, we associate to each clause C_j, its anticlause $\widetilde{C}_j$, that we represent by the pinpoint cylinder $E_j=E(C_j)$ of the cube $\{0,1\}^n$, $1\leqslant j\leqslant k$. In these conditions, the two above problems (i) and (ii) become :

(i) evaluate the cardinality of the union $U\{E_j/1\leqslant j\leqslant k\}$;
(ii) does the union $U\{E_j/1\leqslant j\leqslant k\}$ cover the entire space $\{0,1\}^n$?

In fact, we have :

$$N=2^n-\mathrm{card}(\bigcup_{1\leqslant j\leqslant k} E_j) \;; \qquad (28)$$

then, the covering of the all space $\{0,1\}^n$ by $U\{E_j/1\leqslant j\leqslant k\}$ corresponds to the <u>non satisfiability</u> of the formula (26).

Thus, in the following, our expression will be stated only in terms of pinpoints cylinders of $\{0,1\}^n$, or -when some flexibility is needed, with respect to the geometrical structure of our objects- in terms of subsets of a finite set.

We have said above that each x_i of the n boolean variables, appears necessarily, at least once in the different clauses, under either its positive form x_i or negative form $\widetilde{x}_i$, $1\leqslant i\leqslant n$. Therefore, necessarily, one at least of the following two pinpoint cylinders of order 1, $\{i,1\}$ and $\{i,0\}$ [see (3)], is non empty ; that is to say, includes at least one pinpoint cylinder E_j, associated to a clause C_j, $1\leqslant j\leqslant k$.

On the other hand, we can notice that we can assume that there does not exist in the set $\{E_j/1\leqslant j\leqslant k\}$, two pinpoint cylinders, such that one of them is included in the other one.. Effectively, it is always possible to reduce the system $\{E_j/1\leqslant j\leqslant k\}$ to an equivalent one, by the following simplification algorithm :

1- Order the cylinders by decreasing volume :

$(E_{j1},E_{j2},...,E_{jk})$;

$vol(E_{j1})\geqslant vol(E_{j2})\geqslant...\geqslant vol(E_{jk})$.

2- Compare E_{j_1} with the ordered sequence of the others following cylinders. If $E_{jh}(h\geqslant 2)$ is included in E_{j1}, delete E_{jh}.

3- If E_{jl} is the first cylinder preserved of the sequence $(E_{j2},...,E_{jk}$), go back to 2, with E_{jl} instead of E_{j1}.

4- An so on, until end.

III. EVALUATION OF N AND RECOGNITION OF THE SATISFIABILITY IN A REAL OBSERVED CASE

III.1. Evaluation aspect

According to the above formula (27), we seek to determine

$$\tilde{N}=2^n-N=card(\bigcup_{1\leqslant j\leqslant k} E_j) \qquad (1)$$

III.1.1. Exact evaluation from a partitionning of $\cup\{E_j/1\leq j\leq k\}$

The construction of the partition will take into account the geometrical structure of a pinpoint cylinder ; and in fact, each element of the decomposition will be a pinpoint cylinder. The proposed algorithm is recursive and each of its steps concerns a couple (E_g,E_h) of non disjoint pinpoint cylinders, chosen in an optimal way, faced with the all pinpoint cylinders developed from $\{E_j/1\leqslant j\leqslant k\}$.

According to expression (3) of section II, E_g and E_h are denoted as follows :

$$E_g=\{i_1,i_2,...,i_p,i_{p+1},...,i_q),(\alpha_{i1},...,\alpha_{ip},\alpha_{ip+1..},...,\alpha_{iq})\}(2)$$

and

$$E_h=\{(i_1,i_2,...,i_p,j_{p+1},j_r),(\alpha_{i1},...\alpha_{ip},\beta_{j_{p+1}},...,\beta_{jr})\}, \quad (3)$$

where q is supposed less than r ; that is to say,

$$vol(E_g)\geqslant vol(E_h) \qquad (4)$$

In this algorithm, E_h is partitionned with respect to E_g, into a sequence of exactly q-p+1 pinpoint cylinders, which are mutually disjoint and -except the last- exclusive from E_g. As for the last $(q-p+1)^{th}$ element of the latter decomposition, it concerns a pinpoint cylinder contained in E_g.

Precisely, E_h is decomposed -with respect to E_g- in the following form:

$$E_h=E_h{}^1+\ldots+E_h{}^1+\ldots+E_h{}^{q-p}+F_h, \qquad (5)$$

where

$$E_h{}^{l} = E_h \cap \{(i_{p+1},\ldots,i_{p+l}),(\alpha_{ip+1},\ldots,\alpha_{ip+1-1},\tilde{\alpha}_{ip+1})\}$$

$(1\leqslant l\leqslant q-p)$ and

$$F_h=E_h\cap\{(i_{p+1},\ldots,i_q),(\alpha_{ip+1},\ldots,\alpha_{iq})\}. \qquad (6)$$

As mentionned just above, on the one hand, E_g and the different $E^l{}_h$, $1\leqslant l\leqslant q-p$, are mutually disjoint ; and on the other hand, F_h is included in E_g. Therefore,

$$\mathrm{card}(E_g \cup E_h)=2^{n-q}+ \sum_{1\leqslant l\leqslant q-p} 2^{n-r-1} \qquad (7)$$

We must notice that the number of new pinpoint cylinders introduced in (5) is

$$p-q=\mathrm{Log}_2[\mathrm{vol}(E_h)/\mathrm{vol}(E_g \cap E_h)] \qquad (8)$$

and corresponds exactly -when $\mathrm{vol}(E_g \cap E_h)=\phi$- to the number of instanciated variables, which intervene in E_g but not in E_h.

In these conditions, beforehand applying the s^{th} step of the algorithm, consider that we are faced with the following sequence of pinpoint cylinders

$$\{E_1,\ldots,E_j,\ldots,E_{k(s)}\} \qquad (9)$$

which are supposed ordered according to the decreasing value of the volume :

$$\mathrm{vol}(E_1)\geqslant\mathrm{vol}(E_2)\geqslant\ldots\geqslant\mathrm{vol}(E_{k(s)} \qquad (10)$$

Then, establish the table

$$\{p(g,h)/1\leqslant g<h\leqslant k(s)\} \qquad (11)$$

where

$$p(g,h)=\begin{cases} \mathrm{Log}_2\left[\frac{vol(E_h)}{vol(E_g \cap E_h)}\right] \text{ if } E_g \cap E_h \neq \phi \\ \infty \text{ if } E_g \cap E_h = \phi \end{cases} \qquad (12)$$

Hence, the s^{th} step of the reduction algorithm consists of :

(i) locating -by means of a sorting- a couple $(g_0,h_0), 1\leq g_0<h_0\leq k(s)$, for which $p(g,h)$ is minimal ;

(ii) partitionning E_{h0} with respect to E_g , as in (5) above and then, replacing E_{h_0} by

$$\{E^1{}_{h0} , E^2{}_{h0}, \ldots, E_{h0}{}^{p(g0,h0)}\} \; ; \qquad (13)$$

(iii) reactualization of the sequence (9) by introducing (13) instead of E_{h0} and by reordering its elements according to (10) where $k(s)$ has to be replaced by

$$k(s+1)=k(s)+p(g_{0,}h_0)-1 \; ; \qquad (14)$$

(iv) reactualization of (11), by comparing each element of (13) with each element of (9), except E_{h0}, wich is deleted and E_{g0} for which the $p(g_0,h_{0)}$ values are equal to infinity. $k(s)$ has to be replaced by $k(s+1)$(cf. (14)).

We have $k(1)=k$. On the other hand, the algorithm will necessarily converge to the state where all the entries of the table (11) are filled with the infinity value. The latter case indicates the completion of the process of decomposition which of course may reach an exponential complexity from computational aspect. But it is of importance to note that the exponential nature of the computational problem is relative to the number of clauses k, but not in relation with the number n of variables. Because, the expansion (13) obtained from E_h is linear with respect to the number n of variables.

The seed idea of the decomposition of a clause C with respect to a clause C', is clearly considered in [Simon & Dubois 1989]. In our treatment where a geometrical and synthetic view is given, we have to handle, in an optimal way and globally, a set of clauses. If, in the preceding decomposition, we obtain an exclusive system of K pinpoint cylinders, such as k_i of them have the order r_i, $1\leq i\leq p$ and

$$K=k_1+k_2+\ldots+k_p, \qquad (15)$$

then, we may write the following formula, considered in the above reference :

$$\mathrm{card}\Big(\bigcup_{1\leq j\leq k} E_j\Big)= \sum_{1\leq i\leq p} k_i 2^{n-r_i} \qquad (16)$$

III.1.2. Exact or approximate evaluations from the inclusion and exclusion formula.

We now return to the formula (20) of section II.3 and more specifically to the q^{th} term of the second number which may be written.

$$(-1)^{q+1}\,\Sigma\;\mathrm{card}(E_{j_1}\cap\;\ldots\;\cap E_{j_q}) \qquad (17)$$

where the notations have been already specified [see below formula (20) §II.3] and where $\{E_j/1\leq j\leq k\}$ are supposed to be pinpoint cylinders. Consider a given term of (17) :

$$\mathrm{card}(E_{j1}\cap\;\ldots\;\cap E_{jq}) \qquad (18)$$

and suppose that $r_1,r_2,\ldots$ and r_q are the respective orders of the pinpoint cylinders E_{j1}, $E_{j2},\ldots$ and E_{jq}. If we assume ordered the components of each pinpoint cylinder -as in (3) of section II.1, where $i_1<i_2<\ldots<i_r$- then, the maximum number of comparisons to establish the value of (18), is bounded by $r_1,r_2+\ldots+r_q$, which is linear with respect to n. The latter value is given by (26) of section II.3.

A sum as (17) above comprises $\binom{k}{q}$ terms and all the computational complexity in using the formula (20) (section II.3.), is provided by the necessity to examine each of them. But, some simplifications may arise in a recursive scheme if some of the elements of the sum (17) vanish.

To be more explicit, let us suppose that $E_{j1}\;..\;E_{j2}\;..\;...\;..E_{jq}=\phi$, for a particular subset $\{j_1,j_2,\ldots,j_q\}$ of q elements among $1,2,\ldots,k$. Then, we may eliminate from the $(q+u)^{th}$ term of the expression (20), the examination of $\binom{k-q}{u}$ elements, each latter being necessarily null, because, having the following form :

$$\mathrm{card}(E_{j1}\cap\;\ldots\;\cap E_{jq}\cap E_{jq+1}\cap\;\ldots\;\cap E_{jq+u}) \quad (19)$$

Thus, in all, we eliminate the consideration of

$$\binom{k-q}{1}+\binom{k-q}{2}+\;-\;-\;-\;+\binom{k-q}{k-q}=2^{k-q}-1 \qquad (20)$$

It is now interesting to study the elimination process in passing from the q^{th} term to the $(q+1)^{th}$ term in the expansion (20) (section II.3). For this purpose, consider the following set

$$V_q=\{\{j_1,\ldots,j_q\}/E_{j1}\cap\ldots\cap E_{jq}=\phi\}, \qquad (21)$$

which is a subset of what we have denoted by P_q ({1,2,...,k}) (see below expression (20) in section II.3).

A given element $\{j_1,j_2,...,j_q\}$ of V_q generates (k-q) elements of V_{q+1}. Each of the latter is obtained by inserting an integer j_{q+1} belonging to the complementary of $\{j_1,...,j_q\}$ with respect to $\{1,2,...,k\}$. Let us denote by

$$E(\{j_1,j_2,...,j_q\}) \qquad (22)$$

the subset of $P_{q+1}(\{1,2,...,k\})$ such that $\{j_1,...,j_q\}$ is included in each of its elements ; which represents a (q+1) subset of $\{1,2,...,k\}$. Clearly,

$$E(\{j_1,...,j_q\})=\{\{j_1,...,j_q,h\}/1\leq h\leq k,h \notin \{j_1,...,j_q\}\}.(23)$$

Then, we have the following properties :

(i) $\text{card}[E(\{j_1,...,j_q\})=k-q$; (24)

(ii) for $\{j_1,...,j_q\}\neq\{j'_1,...,j'_q\}$,

$\text{card}[E(\{j_1,...,j_q\})\cap E(\{j'_1,...,j'_q\})]=0(\text{resp}.1)$ iff

$\text{card}(\{j_1,...,j_q\}\cap\{j'_1,...,j'_q\})<q-1$ (resp.=q-1) (25)

In these conditions, consider the table indexed by the set of unordered object pairs of V_q, the entries of this table are filled with 0 or 1 according to the above condition (25). If we denote by t the total number of 1, then the number of elements which necessarily belong to V_{q+1}, as a consequence of the previous analysis of V_q, is equal to

$$(k-q)\text{card}(V_q)-t \qquad (26)$$

Therefore, we have not to consider in the $(q+1)^{th}$ term of the expansion (20) (section II.3), a number of elements equal to (26).

Nevertheless, the most interesting concerns the reason of the bounding formula (24) (section II.3) :

$$S_{2p-1}-S_{2p}=\sum_{\{j_1,...,j_{2p}\}} \text{card}(E_{j1} .. ..E_{j2p}) \qquad (27)$$

In these conditions, the lower is the value of (27), the more accurate is the both bounded formula (24) (section II.3). (27) will be analyzed in section IV from statistical point of view, in the context of a random model of generation of independent pinpoint cylinders or subsets of $\Omega=\{0,1\}^n$.

III.1.3. Simplification of the computing complexity by using a classification method.

We are concerned here with a methodology of hierarchical classification which builds on a set of unit data a classification tree. The latter is obtained by means of an ascendant construction, by successive agglomerations of the most similar classes. The combinatorial and statistical nature of the mentionned approach [Lerman 1981, 1991a] enables to detect in the hierarchy of classifications 'significant' classes and subclasses. On the other hand, the combinatorial structure of a given unit data, may be very general. In our case, each unit data will correspond to a pinpoint cylinder and the set to be classified is the set

$$G=\{E_j/1\leq j\leq k\} \qquad (28)$$

of the pinpoint cylinders respectively associated to the clauses (see sections II.1 and II.2).

Additionally and mainly, this method introduces a most original notion of 'statistics' for measuring statistical relationships and proximities, namely, the 'likelihood' concept. Thus, we set up the 'likelihood' notion as part of the 'resemblance' notion. This principle also underlies the 'information theory' formalism, in which the higher the amount of information quantity, the more unlikely is the event concerned. In our case the events correspond to the observed relations between the pinpoint cylinders considered in (28).

Let us begin by clarifying the usefullness for our problem of an ascendant hierarchical classification method. First recall that in order to obtain such a classification on a set of k elements [see (28)], there exists algorithms of which the computational complexity -in terms of number of comparisons- remains lower than $k^2\log_2 k$ [Bruynooghe 1989]. The interest of the hierarchical classification will be considered at two levels, the latter finer than the former. The purpose of the former is to partition G [see (28) above] into classes such that two pinpoint cylinders, respectively belonging to different classes, have an empty intersection. More precisely, let us designate by

$$\{F_g/1\leq g\leq h\} \qquad (29)$$

such a partition, where we denote by

$$F_g=\{E^g{}_1,\ldots,E^g{}_i,\ldots,E^g{}_{k(g)}\} \qquad (30)$$

the g^{th} class, of which the cardinality is k(g), $1\leq g\leq h$, where we have

$$k=k(1)+k(2)+\ldots+k(h) \qquad (31)$$

Then, the partition (29) is such that the following condition is satisfied

$$[\forall(1\leqslant g<g'\leqslant h),\ (1\leqslant i\leqslant k(g),1\leqslant i'\leqslant k(g'))],E^g{}_i \cap E^{g'}{}_{i'}=\phi \quad (32)$$

In these conditions and with respect to an evaluation based on the inclusion and exclusion formula, we can see that the maximum of the computational complexity is reduced to

$$2^{\max\{k(g)/1\leqslant g\leqslant h\}}\ ; \qquad (33)$$

but, $\max\{k(g)/1\leqslant g\leqslant h\}$ can be equal to k.

A more and significantly intersting aspect concerns the application of the above mentionned hierarchical classification method, based on the likelihood concept, to each class $F_g, 1\leqslant g\leqslant h$, [see (30) above]. For a given g, $1\leqslant g\leqslant h$, the purpose is to decompose F_g into dependent classes which are -more or less- relatively independent from statistical point of view. The process is recursive, because each of the latter classes will be decomposed at its turn ; but, with more dependence between the subclasses ; and so on...

We explore the tree structure on a given $F_g, 1\leqslant g\leqslant h$, [see (30) above] in a descendant process. The 'significant' nodes provided by our method [Lerman & Ghazzali 1991] enable to recognize dependent classes into F_g. Let us designate by

$$\{H_g(t)/1\leqslant t\leqslant u\} \qquad (34)$$

'natural' dependence classes discovered inside of F_g. The complete statistical inependence between the different $H_g(t)$ does mean that there exists a partition

$$\{I_t/1\leqslant t\leqslant u\} \qquad (35)$$

of $I=\{1,2,...,i,...,n\}$, such that, if we denote

$$I_t=\{i_{t1},i_{t2},...,i_{tl(t)}\}, \qquad (36)$$

where $n=\Sigma\{l(t)/1<t<u\}$, each boolean variable X_i, is -for i belonging to I_t- instanciated at least once in the different pinpoint cylinders of $H_g(t)$. But, X_i, for i does not belonging to I_t, is never instanciated in $H_g(t)$. Thus, a couple (E,E') of pinpoint cylinders, belonging to the cartesian product $H_g(t)\mathrm{x}H_g(t')$, are statistically independent for t=t' ; that is to say

$$\mathrm{card}(E \cap E')=\mathrm{card}(E)\mathrm{x}\mathrm{card}(E')/\mathrm{card}(\Omega) = 2^{n-r-r'}, (37)$$

where r (resp. r') is the number in variables instanciated in E (resp. E').

If we indicate by $h_g(t)$ the number of elements of $H_g(t)$, we have

$$k(g)=\Sigma\{h_g(t)/1\leq t\leq u\} \qquad (38)$$

Then, the exact evaluation of

$$\mathrm{card}(\cup\{E/E \in H_g(t)\}) \qquad (39)$$

by means of the inclusion exclusion formula, has a computing complexity of $2^{hg\ (t)}$ order ; but where each elementary calculation is linear with respect to $l(t)$ [see (36) above].

Now, let us designate by A_t the argument of the cardinal function in (39). At this stage, our purpose is to evaluate

$$\mathrm{card}(\cup\{A_t/1\leqslant t\leqslant u\}), \qquad (40)$$

by taking into account the previous evaluations of

$$\{\mathrm{card}(A_t)/1\leqslant t\leqslant u\} \qquad (41)$$

Once again, the inclusion and exclusion formula is applied, but at the level of the u components defined in (40). For the latter level, the computing complexity is of 2^u order. Each elementary calculation concerns the evaluation of an expression of the following form :

$$\mathrm{card}(A_{t1} \cap A_{t2} \cap \ldots \cap A_{tq}), \qquad (42)$$

where $\{t_1,t_2,\ldots,t_q\}$ is a subset of $\{1,2,\ldots,u\}$. In order to evaluate (42), establish in each A_{tp} the statistical distribution D_p of the number of instanciated variables per pinpoint cylinder. The sequence of distributions

$$\{D_p/1\leqslant p\leqslant q\} \qquad (43)$$

enables to analytically evaluate (42) by taking into account the independence relation (37).

Even the statistical independence relation (37) is not always strictly satisfied, the previous calculation under the independence hypothesis, will give an approximate value of $\mathrm{card}\{E_j/1\leqslant j\leqslant k\}$.

We are going now to be more explicit on the manner to obtain the classifications above considered. The general data that is considered first consists of a set of subsets of a given finite set Ω. For the latter, the above notation (28) is preserved ; but we can forget at first, the particular structure of Ω and E_j, $1\leqslant j\leqslant k$.

Relative to the comparison of two parts of Ω E_g and E_h, $1 \leq g < h \leq k$, let us introduce a "raw" proximity index which represents the cardinality of the intersection between E_g and E_h :

$$s(g,h)=\mathrm{card}(Eg \cap E_h) \qquad (44)$$

In order to obtain the above partition (29), let us define on the set of subsets of G [cf. (28)], the following similarity index :

$$s(C,C')=\max\{s(g,h)=\mathrm{card}\{(E_g \cap E_h)/(E_g,E_h) \in C\mathrm{x}C'\}, \qquad (45)$$

where C(resp. C') is a set of elements of G. The latter index will be used in the following ascendant construction of a hierarchical classification, only in case where C and C' are disjoint, in terms of subsets of G.

The starting state of the ascendant construction of a hierarchical classification, is defined by the finest partition ; where each class comprises exactly one element. Thus, in our case the partition of the zero level, can be written as follows :

$$\pi_o(G)=\{\{E_j\}/1 \leq j \leq k\} \qquad (46)$$

Then, at each step, passing from a given level to the following one, the class pairs {C,C'}, for which the similarity index s(C,C') [see expression (45) above] is maximum, are joined. By stopping the classification tree building, just when the index becomes negative, we necessarily end at the announced partition (29), of which the classes have been denoted by F_g, $1 \leq g \leq h$.

At this stage, on a given F_g, $1 \leq g \leq h$, it is of interest to apply "Likelihood Linkage Analysis (L.L.A.)" classification method (Lerman 1991a), in order to set up a hierarchical system of dependence classes (see above). For this purpose, we are going to recall the first aspect -of the mentionned classification method- which concerns the elaboration of a similarity measure between the elements of the set to be classified. Let us designate the latter by

$$F=\{E_j/1 \leq j \leq k'\}, \qquad (47)$$

so that F represents a subset of G [cf. (28) above].

We have already introduced the raw similarity index s(g,h) [cf. (44) above] between two elements E_g and E_h of G, $1 \leq g < h \leq k$. In the context of our problem, where G is composed of pinpoint cylinders from the cube $\Omega=\{0,1\}^n$, we have

$$s(g,h)=2^{n-q-r+p}, \qquad (48)$$

where the expressions of E_g and E_h are given in (2) and (3) (see section III.1.1).

The raw index s(g,h) is statistically normalized with respect to a probabilistic hypothesis of no relation or independence. The easiest expression of this hypothesis does not take into account the specifity of the geometrical structures to be compared. As a matter of fact, to the couple (E_g,E_h) of pinpoint cylinders from Ω, we associate a couple (X,Y) of independent random subsets of a set U ; such that, the cardinality of the triplet (X,Y ; U), respects in a probabilistic sense the cardinality of that one $(E_g,E_h ; \Omega)$:

$$[\mathrm{card}(X),\mathrm{card}(Y);\mathrm{card}(U)]\sim[\mathrm{card}(E_g),\mathrm{card}(E_h);\mathrm{card}(\Omega)] \quad (49)$$

In fact, there are three fundamental forms of the random model of the hypothesis of no relation or independence [Lerman 1981]. Let us introduce for a given form, the random raw index

$$S(g,h)=\mathrm{card}(X \cap Y) \quad (50)$$

that we have to calculate the mathematical expectation E[S(g,h)] and the variance var[S(g,h)]. For the three random models the distributions of S(g,h) are respectively, the Hypergeometric distribution, the Binomial distribution and the Poisson distribution. In all cases, we have

$$E[S(g,h)]=\frac{\mathrm{card}(E_g)\mathrm{x}\mathrm{card}(E_h)}{\mathrm{card}(\Omega)}=2^{n-q-r}, \quad (51)$$

if we consider the expressions (2) and (3) above for E_g and E_h (see section III.1.1).

The Poisson distribution leads to the easiest expression of var[S(g,h)]:

$$\mathrm{var}[S(g,h)=E[S(g,h)]=2^{n-q-r}, \quad (52)$$

Restricting oneself to the latter random model, the standardized association coefficient between E_g and E_h is expressed by

$$Q(g,h)=\frac{s(g,h)-E[S(g,h)]}{(\mathrm{var}[S(g,h)])^{1/2}} \quad (53)$$

More precisely, we have here

$$Q(g,h)=(2^{n-q-r})^{1/2}\ (2^p-1) \quad (54)$$

We can notice that $Q(g,h)$ is an increasing function of p ; but a decreasing function of (q+r). That is to say ; for a given p, the greater is $Q(g,h)$, the more filled is the space $\Omega=\{0,1\}^n$, by the union $E_g \ldots E_h$. The expression (54) weights in a certain manner the increase with respect to p and the decrease with respect to (q+r).

The preceding random models do not take into account the particular pinpoint cylinder structure of E_g and E_h in the logical cube $\{0,1\}^n$. More accurate random model will associate to a given pinpoint cylinder E, having q instanciated variables, a random pinpoint cylinder X*, where the randomness concerns the instanciated components. For a given component i, $1 \leq i \leq n$, the probability of an instanciation is set equal to $\chi(\chi=q/n)$. Generally, q is small relative to n ; then, a perfect approximation of the probability distribution of the random number L of instanciated components is given by the Poisson distribution of parameter q. On the other hand, for $L=l$ and for the specified components, the 2^l possible instanciations are considered as equally probable.

In these conditions, to the couple (E_g,E_h) of pinpoint cylinders, a couple (X*,Y*) of independent pinpoint cylinders is associated under a hypothesis of no relation, where Y* is associated to E_h in an analogous way as X* is associated to E_g ; the elements χ, L and l being respectively replaced by $\rho=r/n$, M and m. Then, one can determine the probability law of the random variable -representing a random similarity index- that we denote here by :

$$T(g,h)=\mathrm{card}(X^* \cap Y^*) \qquad (55)$$

This probability distribution will be clearly establised in section IV ; but, in a very different context. We will see that

$$E[T(g,h)]=2^{n-(q+r)}$$

$$\mathrm{var}[T(g,h)]=2^{[n-(q+r)]}\ (e^{n\chi\rho}-1) \qquad (56)$$

It is interesting to note the following properties :

$$E[T(g,h)]=E[S(g,h)] \ ; \text{ but, } \mathrm{var}[T(g,h)] \neq \mathrm{var}[S(g,h)] . (57)$$

Therefore, the statistically normalized index can be written here :

$$R(g,h)=\frac{s(g,h)-E[T(g,h)]}{(\mathrm{var}[T(g,h)])^{1/2}}$$

$$=\frac{2^p-1}{(e^{n\chi\rho})-1)^{1/2}} \qquad (58)$$

As for the non constrained form of the random model concerning the statistical independence hypothesis, we obtain the same phenomenon of increase with respect to p and of decrease with respect to a symmetrical function of q and r ; but the latter decrease is differently weighted in case of the coefficient R(g,h) [cf. (58)] than in case of the coefficient Q(g,h) [cf. (53)].

In the expressions (53) and (58), we necessarily assume a non empty intersection between E_g and E_h. If the pinpoint cylinders E_g and E_h are exclusive [s(g,h)=0], the coefficients Q(g,h) and R(g,h) take the following strictly negative values :

$$Q_o(g,h)=-\ (2^{n-q-r})^{1/2}$$

$$R_o(g,h)=-1/(e^{n\rho\lambda}-1)^{1/2} \qquad (59)$$

Otherwise, the values of Q(g,h) and R(g,h) are positive or null. Let us notice that the zero value corresponds exactly to the situation where card($E_g \cap E_h$) is equal to its mean value, according the probabilistic independence hypothesis. The latter case is that one where p=0.

In the framework of the L.L.A. method [Lerman 1991a], the table $\{Q(g,h)/1<g<h<k'\}$ (resp.$\{R(g,h)/1\leq g<h\leq k'\}$) leads to a probabilistic scale for measuring the associations between the elements of the set F [cf. (47) above]. The latter scale is established with respect to a global hypothesis of independence, where to F is associated a family

$$F^*=\{E^*_j/1\leq j\leq k'\} \qquad (60)$$

of independent subsets of Ω, such that, the structures of the different E^*_j, $1\leq j\leq k'$, are considered with more or less constraints ("free" subsets or pinpoint cylinders of Ω). On the other hand, the E^*_j, $1\leq j\leq k'$, respect -strictly or in a probabilistic sense- the cardinalities of the E_j, $1\leq j\leq k'$.

III.2. Algorithms of satisfiability recognition

III.2.1. Using the inclusion and exclusion formula

The recognition algorithm is directly deduced here from the both side bounded formula (24) (section II.3). To be more accurate, we are going to distinguish in the formula (20) (section II.3), the case where k=2q is even and that one where k=2q+1 is odd.

For k=2q, the last term of the second member of (20) is negative or null. The latter corresponds in the generic terms, to p=q. In these conditions, for p=1,2,...,q :

if $S(2p-1)<2^n$; then the satisfiability is ensured ;

if $S(2p)\geq 2^n$; then the system is contradictory ;

if $S(2p)<2^n\leq S(2p-1)$; the decision is impossible and then, do p=p+1.

Now, for k=2q+1, the last term of (20) (section II.3) is positive or null. The latter corresponds -relatively to the generic terms- to p=q+1. Therefore, the decision process corresponds exactly to above.

It is clear that the decision process will end when p reaches its highest value. Then, the maximal order of the computational complexity is 2^k. But we may hope in the most current practical cases to conclude with a less complexity order ; specially, if the average of the number r_j of specified variables per pinpoint cylinder E_j, $1<j<k$, is not too low. The latter point will become clearer in section IV.

III.2.2. Cartesian filling of the cube $\{0,1\}^n$

Consider the representation of a given pinpoint cylinder E,associated to a clause C, by a vector with n components, of which the i^{th} component, $1\leq i\leq n$, is equal to 1,0 or ε, according that in C, the variable X_i is instanciated by 1,0 or is indeterminate [see (5) section II.1].

Relative to the system $\{E_j/1\leq j\leq k\}$ of pinpoint cylinders, we are going to define a specific order on the n components, depending on statistical considerations. Morevoer, for each component i, $1\leq i\leq n$, an order will be established between the two values 0 and 1.

For a given i, $1\leq i\leq n$, consider the two complementary hyperplans :

$$\{i,1\} \text{ and } \{i,0\} \qquad (61)$$

and let us define the integer numbers $k(i,1)$ and $k(i,0)$, where $k(i,1)$ [resp. $k(i,0)$] is the number of pinpoint cylinders of which the intersection with $\{i,1\}$ (resp. $\{i,0\}$) is non empty. Then we define

$$l(i)=\min\{k(i,1),k(i,0)\}, \qquad (62)$$

$$1\leq i\leq n.$$

We may suppose without any loss of generality that

$$l(1)\leq l(2)\leq\ldots\leq l(i)\leq\ldots\leq l(n) \qquad (63)$$

On the other hand, relative to the i^{th} component, let us denote by α the logical symbol 0 or 1, for which

$$k(i,\alpha)<k(i,\tilde{\alpha}), \qquad (64)$$

where $\tilde{\alpha}$ is the complementary symbol :

$$\tilde{\alpha}=0(\text{resp.}1) \iff \alpha=1(\text{resp.}0). \qquad (65)$$

In these conditions, we set for the i^{th} component

$$\alpha<\tilde{\alpha} \qquad (66)$$

Now, consider the logical cube $\{0,1\}^s$ defined by the s first components, $1<i<s$, according to (63) above. The latter cube represents the whole set of values of the vector $(X_1,\ldots,X_i,\ldots,X_s)$ of the s first boolean variables. The s relations (66), established for $i=1,2,\ldots,s$, lead to the definition of a lexicographgic total order on the whole set of boolean vectors, belonging to $\{0,1\}^s$. To be completely clear on this purpose, we set up for the relation (66)

$$0<1, \qquad (67)$$

in case where $k(i,\alpha)=k(i,\tilde{\alpha})$, for a given i.

This construction is made in order to give a priori a maximum of chance for a satisfiability conclusion by the following algorithm. Effectively, at a given step, the latter will explore a sequence of boolean vectors -of a given dimensionnality- ranked according to the preceding lexicographic order.

In fact, the algorithm is recursive. At the s^{th} stage the situation can be represented by a two entries crossing table, comprising 2^s rows and k columns. The r^{th} row is labelled by the r^{th} value of the boolean vector $(X_1,X_2,\ldots,X_s)$ according to the above lexicographic order, $1\leq r\leq 2^s$. The j^{th} element of the r^{th} row is 0 or 1 ; the 0 (resp. 1) value indicates that the j^{th} pinpoint cylinder E_j is disjoint (resp. has non empty intersection) from (resp. with) the pinpoint cylinder -comprising s instanciated variables- defined by the above boolean vector.

If a given row does not contain any element of which the value is 1, the satisfiability condition is trivially performed. If not, consider the variable X_{s+1} and the associated logical value α, acording to (66) above, for i=s+1. Then, for r=1 to $r=2^s$:

(i) divide the r^{th} row into tow rows, which are labelled by two boolean vectors of which the dimensionnality is (s+1). The first s components are those of the preceding boolean vector concerning the preceding r^{th} row. The latest component of the first (resp. second) boolean vector is equal to α(resp. $\tilde{\alpha}$). For the new rows, the elements which are equal to 1, are <u>necessarily</u> among those which are equal to 1, for the preceding r^{th} row ;

(ii) if for one of two new rows respectively associated to the r^{th} row, there is no element which is equal to 1 ; then, the satisfiability of the system is acquired. If not ;

(iii) continue by considering r=r+1.

If the latter process ends with $r=2^s$, without obtaining, from the subdivision process, a row filled up with zero elements ; then, we reach the $(s+1)^{th}$ stage. Finally and necessarily, a conclusion is obtained on the satisfiability of the system ; but the maximal order of the computing complexity is 2^n.

III.2.3. Algorithm of parallel covering by hyperplans

This algorithm is one of the most interesting. It may 'quickly' conclude to the satisfiability of the system. The general idea consists in recognition of a pinpoint cylinder of $\{0,1\}^n$, which is exclusive from the union $U\{Ej/1\leq j\leq k\}$ of the all pinpoint cylinders associated to the different clauses [cf. section II.1]. The method operates by a suitable grouping of the pinpoint cylinders into blocks, respectively included in coordinate subspaces or hyperplans of $\{0,1\}^n$.

Relative to the set $\{E_j/1\leq j\leq k\}$ of the given pinpoint cylinders, let us designate by $\{i_1,\alpha_{i_1}\}$ a coordinate hyperplan which contains a maximal number of E_j, $1\leq j\leq k$. If k(1) is this maximal number, we have an inclusion relation such that

$$E_{jh} \subset \{i_1,\alpha_{i1}\},$$

$$\text{for } 1\leq j_h\leq k(1) \qquad (68)$$

If k(1)=k, the satisfiability condition is acquired, because

$$(U\{E_j/1\leq j\leq k\}\cap\{i_1,\alpha_{i1}\}=\phi, \qquad (69)$$

where $\tilde{\alpha}_{i1} = 1-\alpha_{i1}$.

The extension of the inclusion (68) is given by the following sequence of inclusions

$$U\{E_{jt}/k(u-1)+1\leqslant t\leqslant k(u)\}\subset\{i_u,\alpha_{iu}\}, \qquad (70)$$

where $1\leqslant u\leqslant l$, where $k(0)=0$ and where the sequence of the integer numbers :

$$\{k(u)-k(u-1)/1\leqslant u\leqslant l\} \qquad (71)$$

is decreasing.

If $k(l)=k$, the satisfiability is acquired, because we have

$$(U\{E_{jt}/1\leqslant t\leqslant k\}) \bigcap \{(i_1,\ldots,i_l),(\tilde{\alpha}_{i1},\ldots,\tilde{\alpha}_{il})\}=\phi \quad (72)$$

But we may reach a maximal value l_1 of l, with a strict inequality

$$k(l_1)<k \qquad (73)$$

In this case, it rests the following pinpoint cylinders

$$\{E_j/j \notin \{j_1,\ldots,j_{k(11)}\} \qquad (74)$$

outside of

$$U\{\{i_u,\alpha_{iu}\}/1\leqslant u\leqslant l_1\} \qquad (75)$$

Then, each E_j of (74) has necessarily the following form

$$E_j=\{(i'_1,\ldots,i'_h),(\tilde{\alpha}_{i'1}\ ,\ldots,\alpha_{i'h}\ )\} \qquad (76)$$

where

$$\{i'_1,\ldots,i'_h\} \subset \{i_1,i_2,\ldots,i_{l1}\}$$

The complementary subset of (75) is the pinpoint cylinder

$$\{(i_1,\ldots,i_l\ ),(\tilde{\alpha}_{i1},\ldots,\tilde{\alpha}_{il1})\} \qquad (77)$$

which is included in each E_j of (74) [see (76) above].

Let us designate by R the set (74) of pinpoint cylinders. We are going now, step by step to reduce R to the only pinpoint cylinder (77). With respect to the latter, the complementary subset in R will be integrated to the first member of (70).

Then, and recursively from i_1 to i_{l1} , consider the $i^{th}{}_u$ component and the latest state of R after (u-1) steps, $1 \leqslant u \leqslant l_1$. Detect in R all the pinpoint cylinders for which the $i^{th}{}_u$ is not specified. Decompose each of the latter into two pinpoint cylinders with one more (the $i^{th}{}_u$) specified component. This, is equal to α_{iu}, for the former and to α_{iu}, for the latter. In these conditions, the first pinpoint cylinder *is included in the* u^{th} subset *of the union considered in the* first member of (70). But, the second pinpoint cylinder -of which the $i^{th}{}_u$ component is equal to $\tilde{\alpha}_{iu}$ - is kept in R, unless it does already exist in R. In the latter case, it is suppressed.

As announced, after $u=l_1$, R is necessarily reduced to the pinpoint cylinder (77). On the other hand, the number of elements of each subset of the union -considered in the first member of (70)- increases in a non predictible manner. In any case the union between the latest state of the first member of (70) and the pinpoint cylinder (77) is exactly equal to the union $U\{E_j/1 \leqslant j \leqslant k\}$ of the different elements of G [cf. (28) section III.1.3.]. On the other hand and clearly, the pinpoint cylinder (77) is disjoint from the first member of (70).

Consider the relation (70) as defined at its final state after the application of the preceding algorithm. We have :

$$U\{E'_{jt}/k'(u-1)+1 \leqslant t \leqslant k'(u)\} \subset \{i_u, \alpha_{iu}\}, \quad (78)$$

where $1 \leqslant u \leqslant l_1$, where $k'(0)=0$ and where $k'(l_1)$ is generally greater than k. On the other hand,

$$k'(u)-k'(u-1) \geqslant k(u)-k(u-1), \qquad (79)$$

for $1 \leqslant u \leqslant l_1$. Notice that the sequence

$$\{k'(u)-k'(u-1)/1 \leqslant u \leqslant l_1\} \qquad (80)$$

does not still necessarily preserve the decreasing property of (71).

We may now conclude by the following.

THEOREM. The system is satisfiable if and only if, one at least of the l_1 preceding inclusion relation (78) is strict.

Thus, by the preceding algorithm the SAT problem with n variables is replaced by l_1 SAT problems with (n-1) variables, where l_1 is strictly lower than n and where the number of clauses (or pinpoint cylinders) concerning the u^{th} problem is $[k'(u)-k'(u-1)]$, $1 \leqslant u \leqslant l_1$.

It is of importance to notice that the l_1 problems can be treated in parallel. Then, the computational complexity relative to the number of clauses k, is now relative to

$$\max\{k'(u)-k'(u-1)/1\leq u\leq l_1\} \qquad (81)$$

On the other hand, the maximal computing complexity of the divinding process is of $[k-k(l_1)]2^{l1-1}$ order. More precisely, a pinpoint cylinder E_j, as given in (76) above, car generate 2^{l1-h} pinpoint cylinders. Therefore, the general problem remains tractable if both l_1 and $[k-k(l_1)]$ are enough low.

III.3. Concluding remarks

We want here to emphasize the great interest of a pre-treatment by a hierarchical classification process based on the L.L.A. method (above mentionned) ; and that, whatever is the basic NP problem (evaluation of the number of solutions or recognition of the satisfiability of the system). We have already in section III.1.3. illustrate this interest, in case of applying the inclusion and exclusion formula, in order to evaluate the number of solutions.

For a given F_g [cf.(30) section III.1.3.], consider the partition (34) [cf. section III.1.3.] into independent classes. The subsystem concerned by a fixed class $H_g(t)$ depends on only $l(t)$ variables ; $l(t)$ being a portion of n, $1\leq t\leq u$. Let us designate by $N_g(t)$ the number of solutions of the latter subsystem. Then, the number of points in the union of the pinpoint cylinders belonging to $H_g(t)$, is equal to

$$\tilde{N}_g(t)=2^{l(t)}-N_g(t), \qquad (82)$$

$1<t<u$. Therefore, the number of points in the union of the elements of F_g, is given by

$$\tilde{N}_g=\prod\{\tilde{N}_g(t)/1\leq t\leq u\} \qquad (83)$$

On the other hand, the whole system (concerning F_g) is satisfiable if and only if each subsystem, associated to $H_g(t)$, $1\leq t\leq u$, is satisfiable.

For a given $H_g(t)$, $1\leq t\leq u$, a good strategy consists in following the hierarchical classification tree established on $H_g(t)$, in an ascendant way and as parallely as possible, according to an inclusion relation between classes.

Consider for example the algorithm defined in section III.2.2. which can be regarded as the most classical. Very intuitively speaking, the contradiction of the satisfiability condition has some tendency to occur since the first levels. Because, the lower is the level, the fewer are the number of variables specified [see remarks after expressions (54) and (58) in section III.1.3.].

IV. EVALUATION OF N AND RECOGNITION OF THE SATISFIABILITY IN THE CASE OF A RANDOM MODEL

IV.1. Introduction ; description of different random models

As announced in the general introduction (see section I) we are going here to consistently retake the statistical aspects considered in [Simon & Dubois 1989]. The latter concern the both mentionned problems (i) (evaluation) and (ii) (recognition), in the context of a random generation of clauses. In fact, in our set theoretic and geometrical representation (see section II.1), a system of independent random subsets or pinpoint cylinders from a logical cube, is considered. Then it is necessary to be more explicit on the different versions of the generation random model.

At this stage, the observed data are assumed to be a couple (Ω,E) where Ω is the logical cube $\{0,1\}^n$ and where E is a pinpoint cylinder of order r. Two families of models can be considered, where the latter is more constrained than the former. For the first family, the cartesian structure of the pinpoint cylinder is not taken into consideration. We only retain that E is an l subset of a m set Ω, where $l=2^{n-r}$ and $m=2^n$. But, for the second family, the geometrical structure of (Ω,E) is intimately taken into account. Let us designate by (O^*,X^*) [resp. (Ω^*,E^*)] the random couple associated to (Ω,E) in the framework of the first (resp. second) family. As a matter of fact, the second family will correspond in a specific sense to the first one. Thus, for each family, there exists three fundamental forms for the random model. In these conditions, let us denote by $({}_iO^*,{}_iX^*)$ [resp. $({}_i\Omega^*,{}_iX^*)$ [resp. $({}_i\Omega^*,{}_iE^*)$] the random couple associated to the i^{th} form for the first (resp. second) type of the random model, $1\leqslant i\leqslant 3$. On the other hand, if ${}_iO^*=O_o$ (resp. ${}_i\Omega^*=\Omega_o$) is given, ${}_iX^*$ (resp. ${}_iE^*$) is a random subset of O_o (resp. Ω_o). Let us now describe each of these random models, by beginning with the first family.

For i=1; ${}_1O^*=\Omega$ and ${}_1X^*$ is a random subset of Ω, of which the cardinality is l [$\mathrm{card}({}_1X^*)=l$]. Thus, ${}_1X^*$ is an element in the set $P_l(\Omega)$-provided by an uniform probability measure- of all parts of Ω, having the same cardinality l. In other words, by considering the simplex 2^Ω of the set of parts of Ω, the concerned model concentrates and distributes uniformely all the probability measure on the l^{th} level. In these conditions, if E_o is a fixed subset of Ω, we have

$$\Pr({}_1X^*=E_o)=\begin{cases}0 \text{ if } \mathrm{card}(E_o)\neq l\\ 1/\binom{m}{l} \text{ if } \mathrm{card}(E_o)=l\end{cases} \qquad (1)$$

This model respects strictly the cardinal characteristics $(m=2^n$, $l=2^{n-r})$ of (Ω,E).

For i=2, the probability measure is shared by the different levels of the simplex 2^{Ω}. This form of the random model comprises two steps. The former consists of choosing a level e which is defined by $P_e(\Omega)$ and the latter consists in choosing one element of this level.

For the level choice, consider the integer random variable L which labells a random level of 2^{Ω}. We set for Pr(L=e), the following binomial probability :

$$Pr(L=e)=\binom{m}{e}\lambda^{e}(1-\lambda)^{m-e}, \qquad (2)$$

where λ is the proportion $1/m=2^{-r}$ (in our case).

Now, for the random choice of an element at a given level e, the probability (2) is uniformely distributed on the set of the $\binom{m}{e}$ points. Each of which represents an e subset of Ω. Therefore, each point of the latter is provided by the probability $\lambda^{e}(1-\lambda)^{m-e}$. In these conditions, if E_o is a given part of Ω of which the cardinality is c, we have

$$Pr[{}_2X^{*}=E_o/card(E_o)=c]=\lambda^{c}(1-\lambda)^{m-c} \qquad (3)$$

For i=3, the random model comprises three steps. On the contrary of the two previous models where Ω is fixed, consider here that ${}_3O^{*}$ is a random set. But, the randomness concerns only the cardinality M of ${}_3O^{*}$. We assume that M is an integer random variable which follows a Poisson distribution of parameter m :

$$Pr(M=p)=\frac{m^{p}}{p!}\,e^{-m}, \qquad (4)$$

for all p belonging to the set of the integer numbers.

Conditionnally to a given value p_o of M, let us introduce a set O_o of which the cardinality is p_o and that we can denote by

$$O_o=\{1,2,\ldots,i,\ldots,p_o\} \qquad (5)$$

Then, the two following steps of the model are entirely analogous to those considered for the preceding model. More precisely, consider a vertex of the level e of the simplex 2^{O_o}. To the latter -which represents a e subset of O_o- we assign the probability $\lambda^{e}(1-\lambda)^{p_o-e}$, where $\lambda=1/m$, keeps exactly the same value as above. Therefore, the random choice of the e level will be done with the binomial probability

$$\binom{p}{e}\lambda^{e}(1-\lambda)^{p_o-e}, \qquad (6)$$

$0\leqslant e\leqslant p_o$.

Now let us make clear the random models of the second family where the cartesian structure of the data (Ω,E) is preserved. To this purpose, instead of considering the definition of the model at the level of an abstract set associated to Ω, consider the latter one at the level of the variable set. More precisely, each form of the random model can be decomposed into two steps. The former consists of determining a subset of a component set. If $\{i'_1,i'_2,...,i'_p\}$ denotes such a subset ; then, only the variables $(X_{i'1}, X_{i'2}, ..., X_{i'p})$ have to be instanciated. In these conditions, all the 2^p instanciations are considered with the same probability $1/2^p$, according to the second step of the random model. The random construction of the above $\{i'_1,i'_2,...,i'_p\}$ is associated to the subset $\{i_1,i_2,...,i_r\}$ of $\{1,2,...,i,...,n\}$ which defines the specified components of the pinpoint cylinder E. The three preceding models can be considered here, since $\{i_1,i_2,...,i_r\}$ is a free subset of the set $\{1,2,...,i,...,n\}$. Let us precise once more in the latter context the second random model, which will have to be considered below.

For this model, n is fixed. Denoting by $\{n\}$ the set $\{1,2,...,i,...,n\}$, the probability measure is distributed on the whole set of all subsets of $\{n\}$: $2^{\{n\}}$. The p^{th} level of the simplex $2^{\{n\}}$ -which is defined by the whole set of p-subsets of $\{n\}$- is provided by the binomial probability :

$$\binom{n}{p}\rho^p(1-\rho)^{n-p}, \qquad (7)$$

$0 \leq p \leq n$, where ρ is the ratio r/n. On the other hand, the latter probability (7) is uniformely distributed on the $\binom{n}{p}$ vertices of the level p ; each representing a p subset of $\{n\}$.

In order to study the two main problems -evaluation or recognition- the one or the other of the preceding random models will be considered. The proposed evaluation will be the assumed value of the mathematical expectation of the number of solutions of a SAT problem, under a probabilistic hypothesis of no relation or independence, between the random structures respectively associated to clauses. In the recognition aspect a random generation -one by one- of the latter structures is concieved. Then, a random integer variable K is introduced. It is defined by the number of elements -representing clauses- for which the system becomes contradictory. The probabilistic law of K and its mathematical expectation are studied. Depending on the problem to be handled, different approaches are considered. One of them is based on the 'inclusion and exclusion' formula. In the another one [personal communication suggested by F. Daudé (Ph D researcher)] a Markow process is associated to the above mentionned random generation of clauses. In the latest approach, to a given point ω of the logical cube Ω, is associated an event A^k_ω which expresses the non covering of ω by the union of k independent random pinpoint cylinders.

The distinction between the random number of solutions N* and the above random variable K, will explain the results obtained in the experimental verification [see section 3.3. of Simon & Dubois 1989] already mentionned in the general introduction (see the above section I).

IV.2. Average number of solutions of a SAT problem in the framework of statistical hypothesis of no relation

IV.2.1. Introduction

With the formalism introduced in section II.4, we have, relatively to a set $\{E_j/1\leq j\leq k\}$ of pinpoint cylinders provided from the logical cube $\{0,1\}^n$, to give an estimation of

$$\widetilde{N}=\mathrm{card}(\bigcup_{1\leq j\leq k} E_j), \tag{8}$$

where $N=2^n-N$, in the context of a random model of probabilistic mutual independence between the different pinpoint cylinders. More precisely, to the sequence $\{E_j/1\leq j\leq k\}$, we will associate the one or the other of the two following random sequences

$$\{{}_iX^*{}_j/1\leq j\leq k\} \tag{9}$$

and

$$\{{}_iE^*{}_j/1\leq j\leq k\} \tag{10}$$

where the free random subset ${}_iX^*{}_j$ (resp. the random pinpoint cylinder ${}_iE^*{}_j$) is associated to E_j, according to the form i of the random model (see the preceding section IV.2.1.), $1<i<3$, $1<j<k$. On the other hand, the ${}_iX^*{}_j$ of (9) [resp. the ${}_iE^*{}_j$ of (10)] are mutually independent. In these conditions, for a given i, $1<i<3$, we associate to N [cf. (8) above] one of the two following random variables

$${}_i\widetilde{N}'^{*}=\mathrm{card}(\bigcup_{1\leq j\leq k} {}_iX^*{}_j) \tag{11}$$

and

$${}_iN^*=\mathrm{card}(\bigcup_{1\leq j\leq k} {}_iE^*{}_j), \tag{12}$$

where the latter one is more accurate in its conception than the former one. But in fact, we will establish that the assumed value of the mathematical expectation of the random variable associated to $\widetilde{N}$ does not depend on the chosen random model. Therefore, if E denotes the mathematical expectation, we have a common value of

$$E({}_i\widetilde{N}'^{*})=E({}_i\widetilde{N}^{*}), \tag{13}$$

for all i=1,2, or 3. Then consider the two general expressions

$$\cup\{X^*{}_j/1\leq j\leq k\} \tag{14}$$

and

$$\cup\{E^*{}_j/1\leq j\leq k\}, \tag{15}$$

which respectively refer to (9) and (10) above. Notice that the case (15) where the random structures are pinpoint cylinders, is more difficult to study. But, as said above, it is more accurate. Then, the latter will be considered, whenever it is tractable. If we not have to specify between (14) and (15) consider the most general expression

$$U\{G^*_j/1\leqslant j\leqslant k\} \qquad (16)$$

The first approach for determining the mathematical expectation of the cardinality of the latter random set, is based on the 'inclusion and exclusion' formula. Then, by taking into account the linearity of the expectation, it suffices to be able to evaluate expressions of the following type

$$E[\mathrm{card}(G^*_{j1} \cap G^*_{j2} \cap \dots \cap G^*_{jh})] \qquad (17)$$

-where $\{j_1, j_2, \dots, j_h\}$ is a fixed h- subset of the subscript set $\{1,2,\dots,k\}$, $1\leqslant h\leqslant k$.

More deeply, we will begin by recalling the probability distributions of card$(G^* \cap H^*)$; and in particular, their respective means and variances (see section IV.2.2.). Next, and by recurrence, we will clearly establish the value assumed by (17), whatever is the nature of $\{G^*_j/1\leqslant j\leqslant k\}$, according to above. For this purpose, we begin by considering the most interesting case where the $G^*_j, 1\leqslant j\leqslant k$, are random pinpoint cylinders (see section IV.2.3.). In section IV.2.4. an approximation of the probability distribution of the random variable

$$\mathrm{card}(G^*_1 \cap G^*_2 \cap \dots \cap G^*_h) \qquad (18)$$

is proposed in the simplest random model.

In section IV.2.5. the above mentionned approach by Markov process is exploited. In this framework, the first random model (see i=1, section IV.1) is considered in order to determine $E(\tilde{N}^*)$. On the other hand, the exact -and by approximation- probability distribution of the random variable $\tilde{N}^*$, is apprehended.

The latest approach -above mentionned in section IV.1- is studied in section IV.2.6. The latter, where to each vertex ω of the logical cube $\{0,1\}^n$, is associated the event A^k_ω to not be covered by the union (16), will be reconsidered in section IV.3. In the latter the recognition problem will be studied from statistical point of view.

IV.2.2. Probability distribution of card($G^*\cap H^*$)

Let us begin by considering that (G^*,H^*) is a couple of independent random subsets, associated to a couple (G,H) of given subsets of Ω. Denote g for card(G), h for card(H) and m for card(Ω). On the other hand, to be more precise, we will designate by $({}_iG^*,{}_iH^*)$ the preceding random couple of independent subsets, if i is the number of the random model considered, $1\leq i\leq 3$ (see section IV.1. above). Thus, to the raw index :

$$s(G,H)=\mathrm{card}(G \cap H), \qquad (19)$$

we associate the random raw index

$$S_i(G,H)=\mathrm{card}({}_iG^* \cap {}_iH^*), \qquad (20)$$

$1<i<3$. In [Lerman 1981] the following results are established :

(i) The probability distribution of $S_1(G,H)$ is a hypergoemetric probability law, of which the parameters are (m,g,h). More clearly

$$\Pr[S_1(G,H)=s]=\frac{\binom{g}{s}\binom{m-g}{h-s}}{\binom{m}{h}}=\frac{\binom{h}{s}\binom{m-h}{g-s}}{\binom{m}{g}} \qquad (21)$$

$\max(0,h+g-m)\leq s\leq \min(g,h)$.

On the other hand, the mean and the variance of $S_1(G,H)$ can be put in the following form

$$E[S_1(G,H)]=m\gamma\eta \qquad (22)$$

and

$$\mathrm{var}[S_1(G,H)]=\frac{m^2}{m-1}\gamma\bar{\gamma}\eta\bar{\eta}, \qquad (23)$$

where $\gamma=\frac{g}{m}$ (resp.$\eta=\frac{h}{m}$) and $\bar{\gamma}=1-\gamma$ (resp. $\bar{\eta}=1-\eta$).

(ii) The probability distribution of $S_2(G,H)$ is binomial with the parameters $(m,\gamma\eta)$. We have

$$\Pr[S_2(G,H)=s]=\binom{m}{s}\pi^s\bar{\pi}^{m-s}, \qquad (24)$$

$0\leq s\leq m$, where $\pi=\gamma\eta$ and $\bar{\pi}=1-\pi$. This implies

$$E[S_2(G,H)]=m\pi=m\gamma\eta \qquad (25)$$

and

$$\mathrm{var}[S_2(G,H)]=m\pi(1-\pi)=m\gamma\eta(1-\gamma\eta) \qquad (26)$$

(iii) The probability distribution of $S_3(G,H)$ is a Poisson distribution with the parameter $m\pi=m\gamma\eta$. In these conditions, we have

$$\Pr[S_3(G,H)=s]=\frac{(m\pi)^s}{s!}e^{-m\pi}, \tag{27}$$

$s\geq 0$,

$$E[S_3(G,H)]=\mathrm{var}[S_3(G,H)]=m\pi. \tag{28}$$

Let us now consider the more interesting case where (G^*,H^*) is a couple of independent pinpoint cylinders associated to an observed couple (G,H) of pinpoint cylinders belonging to the logical cube $\{0,1\}^n$. Therefore, random models of the second family have to be taken into account. Here, we will content ourselves with considering the second version of the random model, above described [see below expression (6) of section IV.1]. In these conditions, let us designate by

$$[(i_1,\ldots,i_r),(j_1,\ldots,j_s)] \tag{29}$$

the couple of the component sets, respectively specified in the couple (G,H) of the given pinpoint cylinders. On the other hand, let us denote by (I^*_2,J^*_2) the two independent random subsets of $\{n\}=\{1,2,\ldots,i,\ldots,n\}$, according to the concerned random model. Then; and strictly, $\mathrm{card}(I^*_2 \cap J^*_2)$ follows a binomial distribution $B(n,\rho\sigma)$, where $\rho=r/n$ and $\sigma=s/n$. But, in the context of our problem ρ and σ are enough small, in order to admit the excellent approximation of the above binomial distribution by the Poisson distribution with the parameter $\mu=\rho\sigma$. Clearly, consider

$$\Pr[\mathrm{card}(I^*_2 \cap J^*_2)=p]=\frac{\mu^P}{P!}e^{-\mu}, \tag{30}$$

for all p integer, $p\geq 0$.

For simpicity notations, let us denote by P the random integer variable $\mathrm{card}(I^*_2 \cap J^*_2)$. On the other hand, the random pinpoint cylinder G^* (resp. H^*) is associated to I^*_2(resp.J^*_2) as described in section IV.1. Let us recall that if $\{i'_1,\ldots,i'_{r'}\}$ (resp.$\{j'_1,\ldots,j'_{s'}\}$) denotes a realization of I^*_2(resp. J^*_2) ; then, only the variables $(X_{i'_1},X_{i'_2},\ldots,X_{i'_{r'}})$ [resp. $(X_{j'_1},X_{j'_2},\ldots,X_{j'_{s'}})$] have to be instanciated for G^* (resp. H^*). On the other hand, all the $2^{r'}$ (resp. $2^{s'}$) instanciations are considered with the same probability $1/2^{r'}$ (resp. $1/2^{s'}$) for the random definition of G^*(resp. H^*).

Relative to this latter model, let us designate by T the random variable $\mathrm{card}(G^* \cap H^*)$. An assumed value O of P, leads necessary to the following value of T

$$2^{n-r-s}, \tag{31}$$

which is obtained only in this case. Therefore, we have

$$\Pr(T=2^{n-r-s})=e^{-\mu}. \qquad (32)$$

A given value p strictly positive of P can either lead to a zero value or to the value $2^{n-r-s+p}$, for T. The former (resp. the latter) case occurs in case of disjunction (resp. non empty intersection) between G* and H*. The first circumstance occurs with the probability $(1-2^{-p})$; and the second one, with the complementary probability 2^{-p}.

Finally, by taking into account the Poisson distribution of P [cf. (30) above].

$$\Pr(T=0)=\sum_{p\geqslant 1}\frac{\mu^p}{p!}\,e^{-\mu}\times(1-2^{-p})$$

$$=1-e^{-\mu/2}, \qquad (33)$$

$$\Pr(T=2^{n-(r+s-p)})=\frac{\mu^p}{p!}\,e^{-\mu}\times 2^{-p}$$

$$=e^{-\mu/2}\times\left[\frac{(\mu/2)^p}{p!}e^{-\mu/2}\right], \qquad (34)$$

for $p\geqslant 0$. Concerning the latter relation, notice that the case where p assumes the zero value [see (32) above], is integrated in (34).

The calculations of mathematical expectation and second absolute moment of the random variable T give

$$E(T)=2^{n-r-s} \qquad (35)$$

and

$$E(T^2)=2^{2(n-r-s)}\times e^{\mu}. \qquad (36)$$

We deduce

$$\operatorname{var}(T)=2^{2(n-r-s)}(e^{\mu}-1) \qquad (37)$$

By this way, the expression (56) of the section III.1.3. -concieved in a very different context- is justified. On the other hand, it is of importance to notice that the mean defined by the mathematical expectation

$$E[\operatorname{card}(G^* \cap H^*)] \qquad (38)$$

does not depend on the chosen random model ; since the common value $m\gamma\eta$ [see expressions (22), (25) and (28) above] can be written as follows

$$m\gamma\eta=2^{n-r-s} \qquad (39))$$

if $m=2^n$, $\gamma=2^{n-r}/2^n=2^{-r}$ and $\eta=2^{n-s}/2^n=2^{-s}$.

As a consequence, the assumed value of the mathematical expectation of expression (18) above, is invariant whatever is the chosen random model. Therefore, this invariance will be preserved for $E(\widetilde{N}^*)$.

IV.2.3. Evaluation by recurrence of $E[\mathrm{card}(E^*_1\cap\ldots\cap E^*_k)]$ and calculation of $E(N^*)$.

Let us recall that $\{E^*_j/1\leqslant j\leqslant k\}$ is a sequence of independent random pinpoint cylinders, associate to an observed one $\{E_j/1\leqslant j\leqslant k\}$, in the framework of a given random model, among those presented [cf. § IV.1. and § IV.2.1.]. Suppose that E_j is a pinpoint cylinder of order r_j (r_j variables are instanciated), $1\leqslant j\leqslant k$. Thus, the volume of E_j is 2^{n-r_j}, $1\leqslant j\leqslant k$. Now, we assume to have established that

$$E[\mathrm{card}(E^*_1 \cap E^*_2 \cap\ldots\cap E^*_i)]=2^{n-(r_1+r_2+\ldots+r_i)}, (40)$$

for $2\leqslant i\leqslant j-1$. Then, we are going to prove the latter relation for $i=j$. We may write

$$E[\mathrm{card}(E^*_1 \cap E^*_2 \cap\ldots\cap E^*_j)]=$$

$$\Sigma\ E[\mathrm{card}(E^*_1 \cap\ldots\cap E^*_{j-1} \cap E^*_j)/\mathrm{card}(E^* \cap\ldots\cap E^*_{j-1})=2^{n-s}]$$

$$\times\ \Pr[\mathrm{card}(E^*_1\cap\ldots\cap E^*_{j-1})=2^{n-s}]\ ;(41)$$

because, 2^{n-s} is necessarily the general form of an intersection cardinal of pinpoint cylinders [see (12) of section II.1]. Now, $E^*_1 \cap\ldots\cap E^*_{j-1}$ and E^*_j are two independent random pinpoint cylinders of which the orders are respectively s and r_j, for the above conditional mathematical expectation. By taking into account the results obtained in the preceding section IV.2.2., the value of the mathematical expectation which concerns the second member of (41) is:

$$2^{n-s-r_j},$$

therefore, the latter second member can be reduced to

$$2^{-r_j}\ E[\mathrm{card}(E^*_1\cap\ldots\cap E^*_j)]. \qquad (42)$$

By considering the recurrence hypothesis [see (40) above], we obtain the expected result, namely

$$E[\mathrm{card}(E^*_1 \cap\ldots.\cap E^*_j)]=2^{n-r_1-r_2-\ldots-r_j}\ . \qquad (43)$$

This result does not depend on the random model considered above. It enables us to determine $E(N^*)$ [see (13) above) by means of the inclusion and exclusion formula [see (20), section II.3]. We have

$$E(\tilde{N}^*)=\sum_{1\leqslant h\leqslant k}(-1)^{h+1}\sum_{\{j_1,\ldots,j_h\}}2^{n-(r_{j_1}+\ldots+r_{j_h})}, \qquad (44)$$

where r_{j_i}, $1\leqslant i\leqslant h$, is the order of the pinpoint cylinder E_j .

We may recognize from (44), by considering the expansion of the second member,

$$E(\tilde{N}^*)=2^n[1-\prod_{1\leqslant j\leqslant k}(1-2^{-r_j})] \qquad (45)$$

The latter corresponds exactly to the result proposed in [Simon & Dubois 1989], where N* is denotes by $\tilde{N}$; but in our notations the symbol ~ has been systematically devoted to the complementation operation.

The latter referenced authors are surprised to notice that

$$E(N^*)=[\prod_{1\leqslant j\leqslant k}(1-2^{-r_j})]2^n \qquad (46)$$

does not depend on the distribution of the variables X_i, on the different clauses. As a matter of fact, this result is clearly foreseeable ; because, the random model of no relation hypothesis, only specifies the number of variables instanciated per clause. But, the number of clauses where a given variable is instanciated, does not intervene. A random model constrained by the both latter aspects, can be represented by a random incidence table -comprising zéro- one boolean values -with k rows and n columns, of which the margins are fixed. A row (resp. column) margin is the number of elements of which the value is equal to 1 in the concerned row (resp. column). An assumed value equal to 1 (resp. 0) at the intersection of the j^{th} row and the i^{th} column, does mean that the variable X_i has (resp. has not) to be instanciated in the j^{th} clause.

The analysis of such random model is much more difficult than the previous ones. The latter may lead to a value of E(N*) which depends on both margins of the random incidence table ; and then, the distributions of the variables X_i, $1\leqslant i\leqslant n$, will be taken into account. However, one may wonder whether the expected results will give more relevant information concerning the problem handled here. For this one, it is mainly in question to study the reduction of the computational complexity, that we can reach, faced with satisfiability instances.

Relative to our estimation problems, henceforth, suppose -without loss of generality with respect the computational complexity- that all the r_j, $1\leqslant j\leqslant k$, are mutually equal to r. In these conditions the above relation (46) becomes

$$E(N^*)=(1-2^{-r})^k \times 2^n \qquad (47)$$

E(N*) is an increasing function with respect to r and a decreasing function with respect to k. On the contrary, concerning the average of the space $\{0,1\}^n$ filling

$$E(\tilde{N}^*)=[1-(1-2^{-r})^k] \times 2^n \qquad (48)$$

is a decreasing (resp. increasing) function with respect to r (resp. k).

Let us now return to the evaluation of the range [see (27) section III.1.2.] of the both sides bounding (24) of section II.3. We may calculate its mathematical expectation under probabilistic independence hypothesis, according to one among the above introduced random models [see section IV.1]. By considering the preceding relation (43), we have

$$E(S^*_{2p-1}-S^*_{2p})=\binom{k}{2p}2^{n-2pr} \qquad (49)$$

Denote by M_{2p} the first member of (49). Then, it is easy to see that

$$\frac{M_{2(p+1)}}{M_{2p}} = 2^{-2r} \times \frac{(k-2p)(k-2p-1)}{(2p+1)(2p+2)} \qquad (50)$$

We may establish that if p is greater than or equal to $k/2^{r+1}$, then the latter second member is strictly lower to unity. A "large" value of r assumes a "small" value of (50) above. More precisely for the both preceding points, the greater is r, the smaller is the minimum value of p -expressed as a portion of k- from which the mathematical expectation of the random interval $[S^*_{2p}\ S^*_{2p-1}]$ range, becomes a decreasing function of p.

IV.2.4. Probability distribution of card$({}_1X^*_1 \cap {}_1X^*_2 \cap \ldots \cap {}_1X^*_h)$

Now, consider the independence or no relation hypothesis which uses the first version of the random model [see section IV.1 and (9) of section IV.2.1]. Let us precise once more that $({}_1X^*_1, {}_1X^*_2, \ldots, {}_1X^*_h)$ -or more briefly- $(X^*_1, \ldots, X^*_h)$ is a sequence of independent randomm subsets of Ω, having the same cardinality, denoted by l. l can be put equal to the volume 2^{n-r} of a given pinpoint cylinder and let us designate by m the volume 2^n of the space $\Omega=\{0,1\}^n$. $P_l(\Omega)$, denoting the set of all parts of Ω, having the same cardinality l, $(X^*_1, \ldots, X^*_h)$ is a sequence of independent random elements of $P_l(\Omega)$; the latter, being provided by an uniform probability measure. Let us recall that card$[P_l(\Omega)]=\binom{m}{l}$.

Property 1. For p, integer included in the interval [max(0,hl-m),l], we have

$$\Pr\{\text{card}(X^*_1 \cap \ldots \cap X^*_h) \geqslant p\} \leqq \binom{l}{p}\left[\binom{l}{p}/\binom{m}{p}\right]^{h-1} (51)$$

The result given in (51) corresponds to an approximation by excess. We will be content with it ; because, an exact result seems inextricable to be set up. In order to obtain (51), choose arbitrarily a value X^0_1 for the random subset X^*_1. X^0_1 is an l-subset of Ω. In these conditions, the necessary and sufficient condition to have

$$\mathrm{card}(X^0_1 \cap X^*_2 \cap \ldots \cap X^*_h) \geqslant p \qquad (52)$$

is that there exists a p-subset of X^0_1, included in each random subset $X^*_2, X^*_3, \ldots, X^*_h$. If a such p-subset is specified, the probability of the latter mutual inclusion is

$$[\binom{l}{p}/\binom{m}{p}]^{h-1} \qquad (53)$$

However, the chosen part of X^0_1, of which the cardinality is p, has not to be specified. Strictly, we have to apply the inclusion and exclusion formula, relatively to the set $P_p(X^0_1)$ of all $\binom{l}{p}$ p-subsets of X^0_1. Because, several elements of the set $P_p(X^0_1)$ can be altogether included in each X^*_j, $2 \leqslant j \leqslant h$. By neglecting the inclusion probability of at least two elements of $P_p(X^0_1)$ in the context of the probability inclusion of at least one element, we obtain the announced result (51) ; which does not depend on the occured value X^0_1 of X^*_1.

Now, we are attempted to propose an approximation by default of the second member of (51) which exceeds stlightly the probability expressed in the first member. The lower is the value of p, the more accurate is our proposed approximation, which can be expressed as follows :

$$\binom{l}{p}\lambda^p \qquad (54)$$

where we denote by λ the quantity

$$(1/m)^{h-1} = 2^{-r(h-1)}.$$

In these conditions, denoting by Y the integer random variable defined by $\mathrm{card}(X^*_1 \cap \ldots \cap X^*_h)$ [see first member of (51), the following approximation

$$\Pr(Y \geqslant p) \simeq \binom{l}{p}\lambda^p \qquad (55)$$

will be as much as admitted, that by this way, a probability distribution is defined, for which the assumed value of the mathematical expectation E(Y), is exactly given by the correspondent value of the second member of (43). That is to say, with the adopted notations, it is necessary to obtain from (55) :

$$\left.\begin{aligned} &\Pr(Y=q) \geqslant 0, \text{ for } 0 \leqslant q \leqslant l \\ &\sum_{0 \leqslant q \leqslant l} \Pr(Y=q) = 1 \\ &E(Y) = l\lambda = 2^{n-hr} \end{aligned}\right\} \qquad (56)$$

Property 2. If h is greater than or equal to n/r, then the numbers $\{\pi_p=[\binom{l}{p}\lambda^p-\binom{l}{p+1}\lambda^{p+1}]/0\leqslant p\leqslant l\}$ are positive and determine a probability distribution of an integer random variable Y, of which the mathematical expectation is equal to $l\lambda$.

π_p is increasing with respect to p. The positiveness of π_0 is acquired iff $h\geqslant n/r$. The sum of the π_p quantities, $0\leqslant p\leqslant l$, can be written

$$\sum_{0\leqslant p\leqslant l}\pi_p = (1+\lambda)^l-[(1+\lambda)^l-1]=1 \qquad (57)$$

On the other hand, we have

$$E(Y)=\sum_{p\geqslant 1} Pr(Y\geqslant 1)=\sum_{p\geqslant 1}\pi_p \qquad (58)$$

By taking into account the expression (55), we obtain

$$E(Y)=(1+\lambda)^l-1$$

$$\simeq 1+l\times(\frac{l}{m})^{h-1}-1=l\times(\frac{l}{m})^{h-1}, \qquad (59)$$

because, generally, λ is enough small. The last result (59) can be translated by 2^{n-hr}, in the context of our problem where $m=2^n$ and $l=2^{n-r}$.

Remarks. The results that we have just obtained with the above properties 1 and 2, correspond to approximations. The latter become mainly valid for h enough large. Thus, concerning the property 2, we reach coherence since the value n/r for h. Let us mention that -in a very different context- the probability distribution of $card({}_1X^*{}_1 \cap {}_1X^*{}_2 \cap {}_1X^*{}_3)$ (h=3) has been very exactly determined [Lerman, research report Irisa, Rennes 1984].

Now, we will not be restricted by the value of h, if we consider a more "fuzzy" random model, where, instead of the random subset ${}_1X^*{}_j$, a sequence of l independent random points of the space Ω, is considered, $1\leqslant j\leqslant h$. In these conditions, the probability for a given point of Ω, to belong to the intersection between the h independent random sequences -respectively associated to the ${}_1X^*{}_j$, $1\leqslant j\leqslant h$,- is equal to (l/m). Then the random number Z of points of Ω, falling in the preceding intersection, follows a Poisson distribution, since m is enough large and $(l/m)^h$, enough small. The parameter of the preceding distribution is precisely $l\lambda$, where $\lambda=(l/m)^{h-1}$.

IV.2.5. An interpretation in terms of a Markov chain

Such interpretation has been communicated to us by F. Daudé (researcher preparing a PhD thesis in data classification). The basic random model concerns the first form (i=1) presented in section IV.1. For the latter, imagine a sequence $(X^*_1, X^*_2, \ldots, X^*_{h-1}, X^*_h)$ of independent random subsets of Ω, having a common cardinality l. We will be interested in the random amount of increase of the filling of Ω, between two consecutive random subsets X^*_{h-1} and X^*_h. Let us designate here by S_g the random cardinal

$$\mathrm{card}(X^*_1 \cup X^*_2 \cup \ldots \cup X^*_g), \qquad (60)$$

then, we have for $\Pr(S_h=x+u/S_{h-1}=x)$ the following hypergeometric probability

$$\Pr(S_h=x+u/S_{h-1}=x) = \frac{\binom{m-x}{u}\binom{x}{l-u}}{\binom{m}{l}} \qquad (61)$$

which leads to

$$E(S_h-S_{h-1}/S_{h-1}) = \frac{(m-S_{h-1})l}{m} \qquad (62)$$

then

$$E(S_h/S_{h-1}) = S_{h-1} - \frac{l}{m} S_{h-1} + l, \qquad (63)$$

from which we denote by reccurrence

$$\begin{aligned} E(S_h) &= m[1-(1-\frac{l}{m})^h] \\ &= 2^n[1-(1-2^{-r})^h], \qquad (64) \end{aligned}$$

for $m=2^n$ and $l=2^{n-r}$.

Notice that the last result corresponds exactly to the above relation (48).

The preceding approach enables us to give the exact form of the probability distribution of S_k. Remind that S_k has been denoted by ${}_1\tilde{N}'^*$ [see expression (11), section IV.2.1.]. We have $S_1=l$. Now, let us designate by T_h the difference (S_h-S_{h-1}). Then, the probability of a given configuration $(l, t_2, \ldots, t_k)$ for $(S_1, T_2, \ldots, T_k)$ can be put into the following form :

$$p(l, t_2, \ldots, t_k) = \Pr(T_2=t_2, T_3=t_3, \ldots, T_k=t_k)$$

$$= \frac{\binom{m-l}{t}\binom{t}{t_2, t_3, \ldots, t_k}\prod_{2\leq h\leq k}\binom{t_1+t_2+\cdots+t_{h-1}}{l-t_h}}{\binom{m}{l}^k}, \qquad (65)$$

where we have denoted by t the sum $t_2+\ldots+t_k$ and where t_1 is put equal to 1, by construction. On the other hand, $\left(t_2,t_3,\overset{t}{\ldots},t_k\right)$ indicates a multinomial coefficient. In these conditions, we have

$$\Pr(S_k=1+t)=\Sigma\{p(1,t_2,\ldots,t_k)/(t_2,\ldots,t_k)\in P_{k-1}(t)\}, \quad (66)$$

where $P_{k-1}(t)$ is the set of all vectors with (k-1) integer components, where each component is comprised between 0 and 1 and where the sum of the components is equal to t.

In order to obtain the latter result (65), consider the product of (k-1) hypergeometric probabilities, where the h^{th}, $2\leqslant h\leqslant k$, corresponds exactly to

$$\Pr(T_h=t_h/T_2=t_2,\ldots,T_{h-1}=t_{h-1}) = \frac{\binom{m-\ell-t_2-\cdots-t_{h-1}}{t_h}\binom{\ell+t_2+\cdots+t_{h-1}}{\ell-t_h}}{\binom{m}{\ell}} \quad (66)$$

From (64) one may directly obtain

$$E(T_h)=\left(1-\frac{\ell}{m}\right)^{h-1}\ell$$
$$=(1-2^{-r})^{h-1}\times 2^{n-r} \quad (67)$$

which represents a portion of 2^{n-r}, which decreases exponentially with respect to h.

In order to simplify the expression (65), it is interesting to consider the Poisson distribution of the above (66) hypergoemetric probability. For this purpose, set

$$\lambda_h = \frac{t_1+t_2+\cdots+t_h}{m}, \ 1\leqslant h\leqslant k, \quad (68)$$

where t_1 is equal to 1 by construction. On the other hand, introduce the parameters $\mu_h=1\lambda_h$ for $h=1,2,\ldots,k$; and finally, instead of the values t_h, $1\leqslant h\leqslant k$, consider the deduced values

$$s_h=1-t_h,\ 1\leqslant h\leqslant k\ ; \quad (69)$$

then, the above probability (65) can be approximated by the following probability

$$\frac{1}{(s_1+s_2+\cdots+s_k)!}(\mu_1+\mu_2+\cdots+\mu_{k-1})^{s_1+s_2+\cdots+s_k}e^{-(\mu_1+\mu_2+\cdots+\mu_{k-1})}.$$

(70)

Notice that (70) is not a Poisson probability, because the quantity $(\mu_1+\mu_2+\ldots+\mu_{k-1})$ is directly related to the value of $(s_1+s_2+\ldots+s_{k-1})$. But, we may establish that the assumed value of the mathematical expectation of the random variable associated to $(\mu_1+\mu_2+\ldots+\mu_{k-1})$ is

$$kl-m[1-\frac{l}{m})^k] \qquad (71)$$

which is in accordance with the above relation (48), for what it concerns the mathematical expectation of the random variable associated to $(s_1+s_2+\ldots+s_k)$. The latter corresponds to $kl-S_k$ (or $kl-\tilde{N}^*$). In these conditions, the question arises whether a Poisson distribution can approximate the probability law of N^*. We are going to try to answer this question in the next section.

IV.2.6. An interpretation in terms of Bernoulli variables

Consider a fixed point $\omega=(\omega_1,\ldots,\omega_l,\ldots,\omega_m)$ of the logical space $\Omega=\{0,1\}^n$ and let be E^* a random pinpoint cylinder of order r, obtained according to the first form of the random model. More precisely, the subset $\{i_1,i_2,\ldots,i_r\}$ of the instanciated components is an element chosen in the set $P_r([n])$ -provided by an uniform probability measure- of all r-subsets, of the set $\{1,2,\ldots,i,\ldots,n\}$ that we have denoted by $[n]$. On the other hand, the 2^r possible instanciations are considered with a common probability equal to 2^{-r}. Notice that ω is not reached by E^* iff at least one of the different random components specified in E^* has a different value from the correspondent one in ω. Therefore, by using the inclusion and exclusion formula, we have :

$$\Pr\{\omega\notin E^*\}=r \times \frac{1}{2}-\binom{r}{2}\times\left(\frac{1}{2}\right)^2+\ldots+(-1)^{r+1}\binom{r}{r}\times\left(\frac{1}{2}\right)^r$$
$$=1-\frac{1}{2^r}=1-\frac{2^{n-r}}{2^n}=1-\frac{l}{m} \qquad (72)$$

Notice that this probability is the same in the more relaxed case, where instead of random pinpoint cylinder E^*, we have a "free" random subset of Ω, of which the cardinality is $l=2^{n-r}$.

Now, consider as in (10) above, a sequence $\{E_j^*/1\leq j\leq k\}$ of independent pinpoint cylinders, with a common order r. Then, define the event

$$A^k(\omega)=\{\omega\notin\bigcup_{1\leq j\leq k}E_j^*\} \qquad (73)$$

Taking into account the independence between the E_j^*, $1\leq j\leq k$, we have from (72)

$$\Pr[A^k(\omega)]=(1-2^{-r})^k=(1-\frac{l}{m})^k \qquad (74)$$

Of course, the same result holds for a sequence $\{X_j^*/1\leq j\leq k\}$ -as in (9) above- of independent "free" subsets, with a common cardinality $l=2^{n-r}$.

Let us now introduce the random boolean variables, respectively associated to the events $A^k(\omega),\omega\in\Omega$. We will denote by $\alpha(k,\omega)$ the random boolean (or Bernoulli) variable, which indicates the event $A^k(\omega)$ $\alpha(k,\omega)=1$(resp.0) iff. $A^k(\omega)$ occurs (resp. does not occur) $\omega\in\Omega$. Then random number of solutions -that we can designate by $N^*(k)$ (see section IV.2.1.) can be put in the following form

$$N^*(k)=\Sigma\{\alpha(k,\omega)/\omega\in\Omega\} \qquad (75)$$

By considering the more flexible random model where a sequence as (9) (section IV.2.1.) is considered, we may introduce the corresponding event to (73) :

$$B^k(\omega)=\{\omega\notin \bigcup_{1\leq j\leq k} X_j^*\} \ ; \qquad (76)$$

and then, by denoting $\beta(k,\omega)$ the indicatory function of $B^k(\omega)$, we have - according to expression (11) (section IV.2.1.)-

$$N'^*(k)=\Sigma\{\beta(k,\omega)/\omega\in\Omega\} \qquad (77)$$

The result obtained in (74) above, leads -by means of (75) and (77)- very directly to

$$\begin{aligned} E[N^*(k)]&=E[N'^*(k)]=2^n(1-2^{-r})^k \\ &=m(1-\frac{l}{m})^k \qquad (78)\end{aligned}$$

Now, consider an intersection of e events such $A^k(\omega)$ [resp.$B^k(\omega)$] respectively associated to e distinct points $\omega_1,\omega_2,\ldots,\omega_e$ of Ω. Let us designate by $A^k(\omega_1,\omega_2,\ldots,\omega_e)$ [resp. $B^k(\omega_1,\ldots,\omega_e)$] the new event:

$$A^k(\omega_1,\ldots,\omega_e)=\cap\{A^k(\omega_d)/1\leq d\leq e\} \qquad (79)$$

$$[\text{resp.}B^k(\omega_1,\ldots,\omega_e)=\cap\{B^k(\omega_d)/1\leq d\leq e\} \qquad (80)]$$

The probability of the latter $B^k(\omega_1,\ldots,\omega_e)$ event does not depend on the specificity of the elements of the subset $\{\omega_d/1\leq d\leq e\}$. More precisely, we have for the first form of the random model,

$$\begin{aligned} Pr[B^k(\omega_1,\ldots,\omega_e)]&=[\frac{\binom{m-e}{l}}{\binom{m}{l}}]^k \\ &=\left[\frac{(m-l)(m-1-1)\cdots(m-l-e+1)}{m(m-1)\cdots(m-e+1)}\right]^k \qquad (81)\end{aligned}$$

Consider the function q(e) determined by the ratio $\binom{m-e}{l}/\binom{m}{l}$. The interval definition of q(e) is [1,(m-1)]. On the latter, q(e) is a descreasing function and we have

$$\left.\begin{array}{l} q(1)=\dfrac{m-l}{m} = 1-2^{r} \\ q(m-1)=1/\binom{m}{l} \end{array}\right\} \qquad (82)$$

If e is small enough, a good approximation of q(e) by excess, is given by q^{e}, where q represents q(1).

On the contrary of the $B^{k}(\omega_1,\dots,\omega_e)$ event case, the probability of the event $A^{k}(\omega_1,\dots,\omega_e)$ does depend on the configuration of the subset $\{\omega_d/1\leqslant d\leqslant e\}$ of vertices of Ω. As a matter of fact, we have seen that the second absolute moment of $\widetilde{N}^{*}(2)$ is different from that one, of $\widetilde{N}'^{*}(2)$ [see expression (37) by comparison with expressions (23),(26) and (28), section IV.2.2.]. But relative to the case where the randomness concerns pinpoint cylinders, consider instead of a given $(\omega_1,\dots,\omega_e)$, a random sequence $(\omega_1{}^{*},\dots,\omega_e{}^{*})_{*}$ of independent points from $\Omega=\{0,1\}^{n}$. For a fixed subsequence $(\omega_{d1}{}^{*},\dots,\omega_{ds}{}^{*})$, the probability for each of its elements to be covered by a random pinpoint cylinder E* of order r, can be expressed as follows

$$\Pr[\cap\{\{\omega_{dq}{}^{*} \in E^{*}\}/1\leqslant q\leqslant s\}]$$

$$= \sum_{E\in\Omega(r)}\Pr(E^{*}=E)\mathrm{x}\Pr[\cap\{\{\omega_{dq}{}^{*} \in E\}/1\leqslant q\leqslant s\}\}], \quad (83)$$

where we have denoted by $\Omega(r)$ the set of all pinpoint cylinders of order r. The cardinality of $\Omega(r)$ is $\binom{n}{r}\,2^{r}$ and then, the value of the second member of (83) is equal to

$$(2^{-r})^{s} \qquad (84)$$

Therefore, by inclusion and exclusion formula, we deduce the probability C_e, for at least one $\omega_d{}^{*}, 1\leqslant d\leqslant e$, to be covered by E* :

$$C_e=e2^{-r}-\binom{e}{2}(2^{-r})^{2}+\dots+(-1)^{e+1}\binom{e}{e}(2^{-r})^{e} \;;(85)$$

and then, the complementary probability D_e can be written

$$D_e=1-C_e=(1-2^{-r})^{e} \qquad (86)$$

hence, we retrieve the above expression q^{e}, where q=q(1) [see above].

The latter probability D_e can be interpreted as the mean of $A^{k}(\omega_1,\dots,\omega_e)$ over the cartesian power Ω^{e} space. In these conditions, it is of interest to study the probability distribution of the random integer variable N'*(k) [see (77) above]. The latter corresponds to the complementary with respect 2^{n}, of the random variable S_k considered in section IV.2.5. But the obtained relations (65), (66) and even (70) make intraciable the probability law.

The above relation (81) makes clear that the events of the family

$$\{B^k(\omega)/\omega \in \Omega\} \qquad (87)$$

are interchangeable that is to say, if $(\omega_1,\omega_2,\ldots,\omega_e)$ is a sequence of mutual distinct elements of Ω, the probability of the above event $B^k(\omega_1,\ldots,\omega_e)$, does only depend on the number e. This circumstance enables to establish the following relation

$$\Pr[N'*(k)=u]=\sum_{0\leqslant j\leqslant m-u}(-1)^j\binom{m}{u+j}\binom{u+j}{u}p(u+j) \quad (88)$$

where $p(u+j)$ represents in our context, $[q(u+j)]^k$ [see above, after relation (81).

The interchangeability property for the above family (87) permits to envisage the known approximation of the probability law of N'*(k), by Poisson distribution [Y.S. Show & H. Teicher 1978]. But, in order to establish the validity of such approximation, the following sufficient condition is required

$$0\leqslant m^{[s]}p(s)\leqslant\lambda_0{}^s,\ 1\leqslant s\leqslant m, \qquad (89)$$

where λ_0 is a finite positive bounded number and where we have denoted by $m^{[s]}$ the s^{th} factorial power of m ; namely, $m(m-1)\ldots(m-s+1)$. In our context the expression $m^{[s]}p(s)$ becomes

$$(m-1)^{[s]}\left[\frac{(m-1)^{[s]}}{m^{[s]}}\right]^{k-1} \qquad (90)$$

But, there is no garantee for a finite limiting behavior of (90). Consider s=1 and obtain the corresponding following expression for (90)

$$m(1-\frac{1}{m})^k \qquad (91)$$

which precisely represents the mathematical expectation of N'*(k). The assumed value of the latter tends to infinity, for a fixed k and for n -> ∞.

It is easy to deduce from (76) that

$$E[\beta(k,\omega)\beta(k,\omega')]=\left[\frac{(m-1)^{[2]}}{m^{[2]}}\right]^k \qquad (92)$$

and then, by considering (77) and by denoting $\gamma=1/m$,

$$\mathrm{var}[N'*(k)]=m(1-\gamma)^k+m(m-1)(1-\gamma)^{2k}-m^2(1-\gamma)^{2k}$$

$$=m(1-\gamma)^k[1-(1-\gamma)^k] \qquad (93)$$

It remains interesting to study -at least experimentally- the probability distribution of the standardized random variable

$$\frac{N'^{*}(k)-E(N'^{*}(k))}{(\mathrm{var}[N'^{*}(k)])^{1/2}}=\frac{N'^{*}(k)-m(1-\gamma)^{k}}{\{m(1-\gamma)^{k}[1-(1-\gamma)^{k}]\}^{1/2}} \qquad (94)$$

Our problem [see the interpretation given in section IV.2.5.] can be related to the classical occupancy problem [see Feller, vol. 1, chap. IV, 1964], for which, each subset X_j^*, contains exactly one element.

IV.3. Average number of clauses to reach the non satisfiability of a random system

IV.3.1. Introduction of a relevant random variable

Consider the first form of the probabilistic hypothesis of no relation [see (9) with i=1 of section IV.2.1.] that we have clearly dealt with, above and particularily, in sections IV.2.5. and IV.2.6. We have established by different approaches that $E[\tilde{N}'^{*}(k)]=m[1-(1-\frac{1}{m})^{k}]$ [see (78) above]. Now, following Daudé's suggestion, let us be interested in the assumed value k_o of the number of clauses k, for which the difference between the volume $m=2^n$ of the whole space Ω and the cardinality of $U\{X_j^*/1\leq j\leq k\}$, is strictly less than unity. Thus, k_o is the lowest value of k for which

$$E(\tilde{N}_k^*)=m[1-(1-\frac{1}{m})^k]>m-1 \qquad (95)$$

By denoting ι_n the logarithm function, we have

$$k > \frac{\ln m}{\ln m-\ln(m-1)}, \qquad (96)$$

$$\frac{k}{n} > \frac{-\ln 2}{\ln(1-2^{-r})} \simeq 2^r\, 2n\ 2\ ; \qquad (97)$$

where the higher is r, the more accurate is the latter approximation.

Thus, we retrieve the proposed value of k_o [see Simon & Dubois 1989]

$$k_o=n\times[\frac{-\ln 2}{\ln(1-2^{-r})}]\simeq n\cdot 2^r\ln 2, \qquad (98)$$

It is a mistake to beleive that the assumed value of k_o provides the average of the random number defined by the minimum number of random subsets $X_j^*$of which the union covers the whole space Ω. Remin d that the latter situation corresponds -in our representation- to the non satisfiability reaching. Therefore, we have to clearly distinguish between the preceding random variable K -that we will formally precise below- and the value k_o of the subscript k from which the average value $E[\widetilde{N}'*(k)]$ of the space filling tends to cover the entire space Ω. This distinction permits to understand why the assumed value of the mathematical expectation E(K) of the relevant random variable K, is perceptibly lower than k_o. The latter circumstance explains the results obtained in the table 1 of section 3.3. of [Simon & Dubois 1989] concerning the "experimental verification".

Notice that our problem is in fact a generalization of the classical problem of "waiting times in sampling with replacement" [see Feller, vol.1, chap. IX, 1964], where l has to be considered equal to unity and this corresponds to r=n. In these conditions, according to the second member of (98), we obtain

$$k_o(n,r=n)=2^n \ln 2^n = n2^n \ln 2 \qquad (99)$$

Rigorously speaking the same error of interpretation remains in the latest mentionned reference. However, the approximation of E(K) by k_o is good in the considered case, because l=1. More generally, and according to the results obtained in the above mentionned table, the approximation by excess of E(K) by k_o, is as especially good as l is small ; that is to say, as r/n is large ($l=2^{n-r}$, $r\leq n$).

Let us now present formally the random variable K in the exact framework of a non finite sequence $\{E_j^*/j\geq 1\}$ of independent pinpoint cylinders, having the same order r. An observation of such random variable is defined by

$$k=\min\{h/\bigcup \{E_j/1\leq j\leq h\}=\Omega, \qquad (100)$$

where E_j is an occurrence of E_j^*, $1\leq j\leq h$.

IV.3.2. Average number of random pinpoint cylinders covering the entire space.

The above (74) and (86) obtained relations enable us to take into consideration the fundamental interest of the more flexible case, where we have a non finite sequence $\{X_j*/j\geq 1\}$ of independent random subsets of Ω, with the same cardinality $l=2^{n-r}$. Reconsider here the random variable K that we have just introduced above [see around expression (100). The event $\{K>k\}$ expresses that at least one of the previous events $B_k(\omega)$ [see (76)] occurs. By using the inclusion and exclusion formula we may write

$$\Pr\{K>k\} = mq_1{}^k - (\;)q_2{}^k + \ldots + (-1)^{s+1}(\;)q_s{}^k + \ldots$$
$$\ldots + (-1)^{m-l+1}\binom{m}{m-l} q_{m-1}{}^k \qquad (101)$$

The general term of the second member expansion can be written

$$\binom{m}{s} q_s{}^k = \binom{m-l}{s} q_s{}^{k-1} \qquad (102)$$

By this way, the above relation (101) becomes

$$\Pr\{k>k\} = \binom{m-l}{1} q_1{}^{k-1} - \binom{m-l}{2} q_2{}^{k-1} + \ldots + (-1)^{s+1}\binom{m-l}{s} q_s{}^{k-1}$$
$$+ \ldots + (-1)^{m-l+1}\binom{m-l}{m-l} q_{m-1}{}^{k-1} \qquad (103)$$

The latter probability is necessarily equal to unity if $k<m/l=2^r$. We have effectively verified this property for m=8, l=2 and k=2.

By admitting the above considered approximation [see around expressions (81) and (82) above] of q_s by q^s (where $q=1-(l-m)$), we have

$$\Pr\{K\leqslant k\} \simeq (1-q^{k-1})^{m-l} \qquad (104)$$

As expected, the latter tends to unity as n tends to infinity.

By exploiting the expression

$$E(K) = \Sigma\{\Pr\{K\geqslant h\}/h\geqslant 1\},$$

we have to sum the expression in the right member of (103) from k=0, in order to obtain

$$E(K) = \binom{m-l}{1}\frac{1}{q_1(1-q_1)} - \binom{m-l}{2}\frac{1}{q_2(1-q_2)} + \ldots$$
$$\ldots + (-1)^{m+1}\binom{m-l}{s}\frac{1}{q_s(1-q_s)} + \ldots$$
$$\ldots + (-1)^{m-l-1}\binom{m-l}{m-l}\frac{1}{q_{m-l}(1-q_{m-l})} \qquad (105)$$

We also have from (101)

$$E(K) = \sum_{1\leqslant s\leqslant m-1} (-1)^{s+1}\binom{m}{s}\frac{1}{1-q_s} \qquad (106)$$

And, by taking into account the above approximation of q_s by q^s, the relation (105) can be written

$$E(K) = \sum_{1\leqslant s\leqslant m-l} (-1)^{s+1}\binom{m}{s}\frac{1}{1-q^s} \qquad (107)$$

In sprite of their computational complexity it is of interest to experiment the relations (104) (105) and (106). For this purpose, the binomial coefficient

$$\binom{m}{s} = \prod_{0 \leqslant i \leqslant s-1} \left(\frac{m-i}{s-i}\right)$$

will be determined from its logarithm.

IV.3.3. A recurrence formula for the probability distribution of K

Relative to the above considered sequence $\{X_j^*/j \geqslant 1\}$ of independent random subsets of Ω, with the same cardinality 1, let us designate by Pm(k-1,<u>u</u>) the probability of exact covering of a <u>specified</u> subset of Ω, of which the cardinality is u, by $U\{X_j*/1 \leqslant j \leqslant k\}$. If k=2, this probability is null for u different from 1. Then, consider $k \geqslant 3$. We have

$$1 \leqslant u \leqslant \min[(k-1)1,m] \qquad (108)$$

Now, if we denote by $Q_m(k-1,u)$ the probability of exact covering of a <u>non specified</u> u-subset of Ω, we have

$$Q_m(k-1,u) = \binom{m}{u} P_m(k-1,\underline{u}), \qquad (109)$$

where $\binom{m}{u}$ is the number of u-subsets. On the other hand, we have

$$P_m(k-1,\underline{u}) = [\binom{u}{l} / \binom{m}{l}]^{k-1} Q_u(k-1,u) \qquad (110)$$

In the preceding relation, $P_m(k-1,\underline{u})$ is multiplicatively decomposed into two probabilities ; where the former concerns the probability of the inclusion event

$$U\{X_j^*/1 \leqslant j \leqslant k-1\} \subset \underline{u}$$

and where the latter is defined by a conditional probability for <u>u</u> to be filled up.

We have the following recurrence formula

$$Q_m(k,m) = \Sigma\{ \frac{\binom{u}{l-m+w}}{\binom{m}{l}} Q_m(k-1,u)/m-1 \leqslant u \leqslant m\} \qquad (111)$$

which is obtained by considering how must be built the occurrence of X_k^*, if we know that the occurrence of $U\{X_j^*/1 \leqslant j \leqslant k-1\}$ covers exactly u elements of Ω.

By deducing $Q_m(k-1,u)$ from relation (109) followed by relation (110), we may write

$$Q_m(k,m)=\Sigma\{\binom{l}{v} * \left[\frac{\binom{m-v}{l}}{\binom{m}{l}}\right]^{k-1} Q_u(k-1,u)/o\leq v\leq l\},\ (112)$$

where u and v are related by u+v=m.

To start the recursion, note that

$$P_m(2,\underline{u}) = \frac{\binom{u}{l}}{\binom{m}{l}} * \frac{\binom{l}{2l-u}}{\binom{m}{l}} \qquad (113)$$

and then

$$P_u(2,\underline{u}) = \frac{\binom{l}{2l-u}}{\binom{u}{l}} \qquad (114)$$

On the other hand, by (108), we have

$$Q_u(2,u)=P_u(2,\underline{u}) \qquad (115)$$

Let us now retake the second member of (112). It is easy to show that the ratio between the two binomial coefficients (under the symbol []) can be approximated by

$$(1-\frac{l}{m})^v=(1-2^{-r})^v$$

Thus, the above reccurrence formula (112) becomes

$$Q_m(k,m)\simeq\Sigma\{(\ )\beta^{v(k-1)}Q_{m-v}(k-1,m-v)/o\leq v\leq l\},\ (116)$$

where

$$\beta=1-\frac{l}{m}=1-2^{-r}$$

It is also here of interest to tabulate the probabilities $Q_m(k,m)$ by using the relations (112)and (116) which can be compared for moderate values of m (e.g. $m=2^{10)}$ and $l=2^7$).

IV.3.4. Realizing that $k_o(r,n)$ is greater than E(K)

Let us recall that $k_o(n,r)$ has been defined by the formula (99) above. This section concerns an illustration which enables to answer the purpose stated in the above title. In this order, suppose that at a given step, exactly (m-i) points of Ω are reached. The latter step being considered as the initial state of the filling system, we are going to compare

(a) the necessary number of steps for which the assumed value of the mathematical expectation of the complementary filling, covers the last i points ;

(b) the mathematical expectation of the necessary number of steps in order to attain the last i points.

In order to make the calculations, the easiest possible, we are going to consider the three cases corresponding to i=1,2 or 3.

<u>i=1</u>

After (k+1) trials of an 1 random subset of Ω, the assumed value of the mathematical expectation of the complementary filling is given by

$$p(1+q+...+q^j+...+q^k) \qquad (117)$$

where $p=1/m=2^{-r}$ and $q=1-(1/m)=1-2^{-r}$. The latter expression (117) which is obtained by recursion, can be put in the following form

$$1-q^{k+1} \; ; \qquad (118)$$

which converges to unity, only when k tends to infinity.

Now, consider

$$\Pr\{K\geq k\}=\left[\frac{\binom{m-1}{l}}{\binom{m}{l}}\right]^{k-1}=q^{k-1} \qquad (119)$$

We obtain

$$E(K)=\sum_{k\geq 1} q^{k-1}=\frac{1}{1-q}=\frac{m}{l}=2^r, \qquad (120)$$

which is clearly finite and does not depend on m.

<u>i=2</u>

The same recursion principle of calculation enables to establish that the expression of the mathematical expectation of the complementary filling, after (k+1) trials, is given by

$$2p(1+q+...+q^k)=2(1-q^{k+1}), \qquad (121)$$

which tends to the 2 value, as k tends to infinity.

Now, by the inclusion and exclusion formula, we obtain :

$$\Pr\{K\geq k\}=\Pr\{K>k-1\}=2\left[\frac{\binom{m-1}{l}}{\binom{m}{l}}\right]^{k-1}-\left[\frac{\binom{m-2}{l}}{\binom{m}{l}}\right]^{k-1}$$

$$\simeq 2q^{k-1}-q^{2(k-1)} \quad . \qquad (122)$$

And then

$$E(K)=\sum_{K\geqslant 1}\Pr\{K\geqslant k\} \cong 2\times\frac{1}{1-q}-\frac{1}{1-q^2}$$

$$=2\times\frac{(3\times 2^r-2)}{(2\times 2^r-1)} \quad ; \qquad (123)$$

and that, for r=2, gives a value of E(K) comprising between 5 and 6.

<u>i=3</u>

The result corresponding to (121) becomes

$$3p(1+q+...+q^k)=3(1-q^{k+1}) \qquad (124)$$

and that one, corresponding to (122), gives

$$\Pr\{K\geqslant k\}=\Pr(K>k-1\}=3q^{k-1}-\binom{3}{2}q^{2(k-1)}+\binom{3}{3}q^{3(k-1)} \quad ;(125)$$

and then

$$E(K)=\frac{1}{2^{-r}}\left(3-3\times\frac{1}{2-2^{-r}}+\frac{1}{3-2^{-r+1}-2^{-r}+2^{-2r}}\right)(126)$$

which is equal to 6.87, for r=2.

V. CONCLUSION

At the term of this long study concerning satisfiability instances with k clauses and n boolean variables, by using a very clear and synthetic representation, different types of problems and situations have been distinguished ; namely, the problem of evaluation of the *number of solutions and that one of recognition of the solution* existence [see (i) and (ii) of section II.4]. On the other hand, the situation where the data is a real observed system of clauses, has received a separate treatment from the case where a random system of clauses is considered. Relative to the former case, we have tried to "do the best" in order to reduce computational complexity.

Without an exact formal proof ; but by the wide variety of the algorithmic approaches (see section III) and by the statistical analysis of the random case, the following main result -concerning either problems (i) or (ii) mentionned above- becomes of evidence :

Relative to the couple (k,n) of parameters -where k is the number of clauses and n, the number of variables- there does not exist a resolution algorithm, which cannot become exponential with respect to either a portion αn of n ($0<\alpha<1$) or (non exclusive) a portion βk of k ($0<\beta<1$).

As a matter of fact, in the case of using the inclusion and exclusion formula, each elementary calculation is quasi-linear with respect to the number of variables n ; but, the exponentiation can be reachable with respect to k. On the other hand, as we have seen, for the interesting algorithm of section III.2.3., the price for deleting one dimension, can be an exponential expansion for a subset of pinpoint cylinders, with respect to a part of the variables.

We should realize that a situation provided by the occurrence of a random system of clauses (represented by pinpoint cylinders) according to a random model of independence (or no relation) hypothesis (see section IV.1.), is the least favorable in terms of computational complexity for exact determination by an algorithm. The latter may either concern the number of solutions or the existence of a solution. Moreover, the assumed value of the mathematical expectation of the number of solutions is a portion of 2^n. Otherwise, detecting statistical independencies between classes of clauses, can provide simplification in determining an approximate value of the number of solutions, by means of mathematical expressions.

In practice, general independence for a real observed system of clauses is absolutely unrealistic. As a matter of fact, there are almost always statistical associations and exclusions between the representative pinpoint cylinders. By proposing a hierarchical classification on the set of the pinpoint cylinders associated to the system, into dependence and subdependency classes, an approximate organization of the preceding relations, is given. Such organization is provided by our method of hierarchical classification where the dependence classes are set up by means of "significant nodes" in the classification tree, automatically detected [Lerman 1981, 1991a], [Lerman & Ghazzali 1991]. Relative to the inclusion relation between two dependence classes, the smaller is the class, the higher is the dependence degree. We have seen (see sections III.1.3 and III.3) how to use the hierarchical classification scheme in order to notably reduce the computational complexity of the different algorithms proposed. This reduction may entail to only be able to determine an approximate value of the expected result. Notice also that algorithms described in sections III.1.1., III.2.2. and III.2.3., include statistical aspects of compactness in their definition. Therefore, our analysis leads to the following second main result :

Compaction algorithmic techniques reduce notably the exponent of the exponential computational complexity of satisfiability problems in most real cases. Among theses techniques, clustering algorithms producing classification schemes, play an important role.

Faced with a real problem of satisfiability instances, after provided it to be irreductible (see just above section III), we begin by applying the hierarchical classification algorithms considered in section III.1.3. Afterwards, we have to establish different possible strategies related to the different algorithms which have been proposed (see section III). These algorithmic strategies can be led in parallel.

ACKNOWLEDGEMENT

We are indebted to Olivier DUBOIS for his remarks and comments after having studied very minutely our first research report [Lerman 91_b]. One of his remarks led us to give a correct and definitive version of the algorithm presented in section III.2.3.

REFERENCES

[1] M. Bruynooghe (1989) ; "Nouveaux algorithmes en classification automatique applicables aux très grands ensembles de données, rencontrées en traitement d'images et en reconnaissance des formes", Thèse de Doctorat d'Etat, Université de Paris VI, 23 janvier 1989.

[2] S.A. Cook (1971) ; "The complexity of the proving procedures" Proceed. 3^r Ann. ACM Symp. on theory of computation, ACM N.Y., pp. 151-158.

[3] S.A. Cook (1983) ; "An overview of computational complexity", Commun. ACM 26, pp. 401-408.

[4] M.R. Garey and D.S. Johnson (1979) ; "Computers and Intractability", Freeman & Co. 1979.

[5] I.C. Lerman (1981) ; "Classification et Analyse Ordinale des Données", Dunod, Paris, 1981.

[6] I.C. Lerman (1991_a); "Foundations of the Likelihood Linkage Analysis (LLA) classification method", Applied Stochastic Models and Data Analysis, vol. 7, number 1, march 1991, J. Wiley, pp.63-76.

[7] I.C. Lerman (1991_b) ; "Nombre de solutions et satisfiabilité d'un problème SAT ; une approche ensembliste, combinatoire et statistique", Publication interne Irisa n° 600, Septembre 1991, 88 pages. Rapport de recherche Inria 1545, octobre 1991.

[8] I.C. Lerman and N. Ghazzali (1991) ; "What do we retain from a classification tree ? an experiment in image coding" in "Symbolic-Numeric data analysis and learning", edited by E. Diday and Y. Lechevallier, INRIA, Nova Science Publishers 1991, proceedings of the conference of Versailles, September 18-20, pp. 27-42.

[9] Y.S. Show and H; Teicher (1978) ; "Probability Theory ; Independence, Interchangeability and Martingals", Springer Verlag (1978).

[10] J.C. Simon and O. Dubois (1989) ; "Number of solutions of satisfiability instances - Applications to knowledge bases", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 3, n° 1, 1989, pp. 53-65.

LISTE DES DERNIERES PUBLICATIONS INTERNES IRISA

PI 632 L-STABLE PARALLEL ONE-BLOCK METHODS FOR ORDINARY DIFFERENTIAL EQUATIONS
Philippe CHARTIER, Bernard PHILIPPE
Janvier 1992, 28 pages.

PI 633 ON EFFICIENT CHARACTERIZING SOLUTIONS OF LINEAR DIOPHANTINE
EQUATIONS AND ITS APPLICATION TO DATA DEPENDENCE ANALYSIS
Christine EISENBEIS, Olivier TEMAM, Harry WIJSHOFF
Janvier 1992, 22 pages.

PI 634 UN NOYAU DE SYSTEME REPARTI POUR LES APPLICATIONS GEREES
PAR UN TEMPS VIRTUEL
Philippe INGELS, Carlos MAZIERO, Michel RAYNAL
Janvier 1992, 20 pages.

PI 635 A NOTE ON CHERNIKOVA'S ALGORITHM
Hervé LE VERGE
Février 1992, 28 pages.

PI 636 ENSEIGNER LA TYPOGRAPHIE NUMERIQUE
Jacques ANDRE, Roger D. HERSCH
Février 1992, 26 pages.

PI 637 TRADE-OFFS BETWEEN SHARED VIRTUAL MEMORY AND MESSAGE
PASSING ON AN IPSC/2 HYPERCUBE
Thierry PRIOL, Zakaria LAHJOMRI
Février 1992, 26 pages.

PI 638 RUPTURES ET CONTINUITES DANS UN CHANGEMENT DE SYSTEME
TECHNIQUE
Alan MARSHALL
Mars 1992, 510 pages.

PI 639 EFFICIENT LINEAR SYSTOLIC ARRAY FOR THE KNAPSACK PROBLEM
Rumen ANDONOV, Patrice QUINTON
Mars 1992, 18 pages.

PI 640 TOWARDS THE RECONSTRUCTION OF POSET
Dieter KRATSCH, Jean-Xavier RAMPON
Mars 1992, 22 pages.

PI 641 MADMACS : A TOOL FOR THE LAYOUT OF REGULAR ARRAYS
Eric GAUTRIN, Laurent PERRAUDEAU
Mars 1992, 12 pages.

PI 642 ARCHE : UN LANGAGE PARALLELE A OBJETS FORTEMENT TYPES
Marc BENVENISTE, Valérie ISSARNY
Mars 1992, 132 pages.

PI 643 CARTESIAN AND SATISTICAL APPROACHES OF THE SATISFIABILITY
PROBLEM
Israël-César LERMAN
Mars 1992, 58 pages.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique.

ISSN 0249-6399