Modern Management based on Big Data II and Machine Learning and Intelligent Systems III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210281

About Strong Dependence of the Complexity of Analysis of the Random 3-CNF Formulas on the Ratio of Number of Clauses to the Number of Variables

Sergey I. UVAROV [a, 1]
[a] *Institute of control Science RAS, Moscow, Russia*

Abstract. The results of a computational experiment on the assessment of the complexity of proving the unsatisfiability of random 3-CNF logical formulas are presented. The dependence of the complexity of this proving on the *R*-ratio of the number of clauses to the number of variables is demonstrated. The computational experiment was carried out for the range of the *N*-number of variables from 256 to 512. An exponential dependence of the median complexity of proving the unsatisfiability of formulas on the number of variables was revealed for each of *R* value: 4.3, 4.6, 5.0, 5.5, 6.0. A formula is constructed that approximates the results of the experiment. According to this formula the exponential component of the median complexity of the analysis of random 3-CNF is estimated as 2 to the power N / (8.4R-17.8).

Keywords. satisfiability (SAT) problem, conjunctive normal form (CNF), clause, variable, literal, complexity.

1. Introduction

The ultimate goal of the research is the question of whether the complexity of the SAT-problem is polynomial or exponential. In this paper, attention is focused on the relationship between the complexity of proving the unsatisfiability of random 3-CNF formulas and the *R*-ratio of the *M*-number of clauses to the *N*-number of variables.

Focusing on the proof of unsatisfiability is methodically justified by the fact that it makes necessary to consider all branches of the formula analysis tree. The number of branches in this tree analysis serves as a measure of complexity. At the same time, the process of proving the unsatisfiability of logical formulas is important for artificial intelligence systems. This is the base of automated theorem proving (also known as ATP or automated deduction). The complexity of automated deduction in general CNF formulas restrict the number of variables involved in deduction. Such limitation on the number of variables arouses interest in estimating the complexity of analyzing CNF formulas. and to the development of the fastest possible computational algorithms.

As an instances for a computational experiment are selected widely used random 3-CNF formulas [1-7]. In the course of the study, random 3-CNF formulas (with relatively small value of *R*) demonstrated consistently high complexity (of the analysis).

[1] Corresponding author, Institute of control Science RAS, Moscow, Russia; E-mail: uvarov53@gmail.com.

Collected representative statistics show an exponential grows of complexity with an increase in the number N of variables from 256 to 512, and fixed values of R (4.3, 4.6, 5.0, 5.5, 6.0).

The use of the median characteristics of the experimental data made it possible to obtain predictable dependencies illustrated graphically.

The research was conducted using SAT-solver based on well-known Davis–Putnam–Logemann–Loveland (DPLL) algorithm [1,2] with *backtracking,* using l*ocal search* [1, 2] and *learning clauses* [1-3] procedures.

We find strong dependence of the complexity of the 3-CNF problem on R. In the computational experiment the denominator of the exponent depends linearly on R.

The median complexity of analysis of random 3-CNF formulas from N variables is estimated as 2 to the power of $N/(\beta R\text{-}\delta)$.

Perhaps in the future it will be possible to prove that all of SAT-solvers of this kind (based on, backtracking, local search and learning clauses) are characterized by an exponential complexity estimate for arbitrary N.

2. Random 3-CNF

We use a well-known and very convenient method [1,2] for generating random 3-CNF formulas. 3-CNF formula is a conjunction of a set of M clauses. Each clause is a disjunction of three terms (literals) of different boolean variables. A term is a logical variable v itself or its negation $\neg v$. The total number of such three-term clauses for N variables is $8\binom{N}{3} \sim O(N^3)$, and each clause is numbered.

When constructing a formula from this general set of numbered clauses, the required number M of clauses is selected randomly. 3-CNF is satisfiable if there is any logical variables assignment that gives true value to the formula.

Such a simple method for generating of random formulas has a remarkable property that limits the area of simultaneous existence of satisfiable and unsatisfiable formulas. It is proved [3,4], that with an increase in the number of variables at $3.52 > R$, unsatisfiable formulas are generated negligibly rarely. At $R > 4.51$ negligibly rarely are generated satisfiable formulas. Statistical studies have shown that at $4.3 < R$ the share of unsatisfiable formulas prevails over the share of satisfiable ones.

It is generally accepted that for randomly generated 3-CNF formulas with certain R ratio of clauses to variables it is difficult to analyze their satisfiability or unsatisfiability.

3. Experimental Results

Table 1 shows the results of a computational experiment to assess the complexity of proving the unsatisfiability of formulas.

In the cells of the table there are μE (N, R) median values of the complexity of proving the unsatisfiability of formulas expressed via the number of branches constructed. Further, in parentheses, the deviation of the μE (N, R) median value from the value specified by the obtained approximation formula is given, expressed as a percentage.

Table 1. The median value of the number of branches in the proof of the unsatisfiability of a random 3-CNF formula depending on N and R.

N	R*=4.3	R=4.6	R=5.0	R=5.5	R=6.0
256	408·10^2 (2%)	8960 (4%)	3010 (1%)		
288	152·10^3 (12%)	260·10^2 (4%)	7670 (1%)		
320	493·10^3 (7%)	799·10^2 (2%)	191·10^2 (1%)	5020 (3%)	
352	154·10^4 (26%)	217·10^3 (5%)	450·10^2 (7%)	110·10^2 (2%)	
384	493·10^4 (7%)	777·10^3 (3%)	145·10^3 (19%)	274·10^2 (13%)	7830 (2%)
416	195·10^5 (8%)	207·10^4 (8%)	310·10^3 (3%)	503·10^2 (2%)	151·10^2 (2%)
448	565·10^5 (7%)	564·10^4 (2%)	719·10^3 (5%)	112·10^3 (1%)	291·10^2 (4%)
480	193·10^6 (7%)	151·10^5 (8%)	179·10^4 (13%)	245·10^3 (1%)	593·10^2 (2%)
512		422·10^5 (12%)	458·10^4 (3%)	546·10^3 (3%)	118·10^3 (2%)

Columns 3 through 6 show the results of an experiment with randomly generated formulas where the number of clauses is a priori set by the value $M = NR$.

It should be noted that for $N \geq 256$ and $R \geq 4.6$, no satisfiable formula has been obtained.

The second column represents random unsatisfiable formulas such that in their construction the number of clauses involved gradually increases until an unsatisfiable formula is obtained. Here, half of the random formulas are unsatisfiable with $R < 4.3$.

In this column, the value $R^* = 4.3$ is the median value of the ratio of the number of clauses to the number of variables at which the proportion of unsatisfiable formulas becomes prevalent. R^* corresponds to the «phase transition» from satisfiability to unsatisfiablity [1, 2] for random 3-CNF formulas.

The experimental results presented in Table 1 are graphically illustrated in Figure 1. Each point of the graph corresponds to the natural logarithm of the number $\mu E(N, R)$ presented in the cell of Table 1.

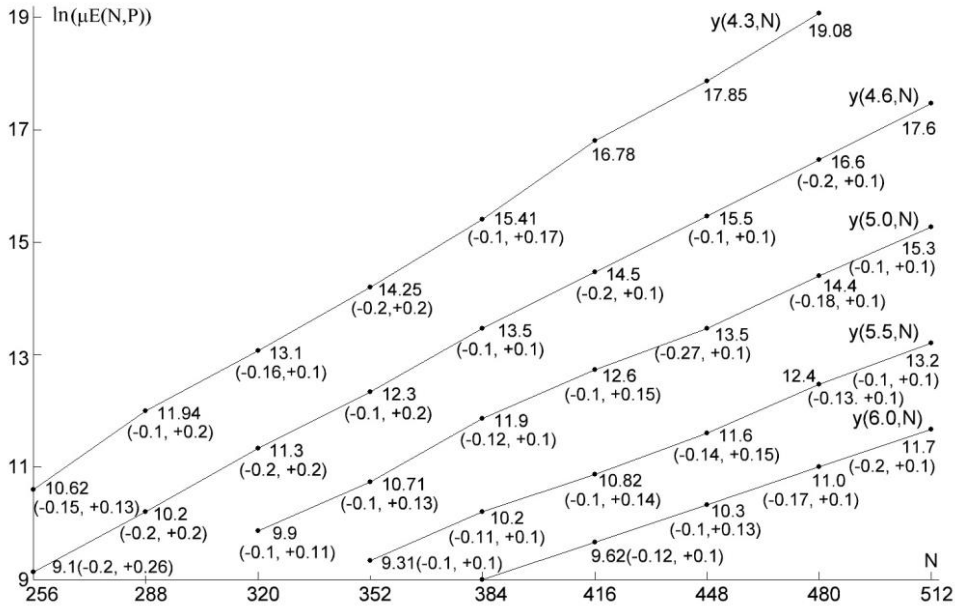

Figure 1. The logarithmic dependence $\mu E(N, R)$.

In proving the unsatisfiability of each of the formulas, the value $Ej(N, R)$, $j = 1 \div K$ was calculated, which represents the number of branches in the analysis tree.

The number $\mu E\ (N, R)$ is the median value among the complexities $Ej\ (N, R)$, $j = 1 \div K$, which are associated with this cell of Table 1.

If K=64 or more random formulas were constructed and analyzed for a cell of Table 1, the confidence interval for the probability 0.95 is given. Confidence interval is shown (in brackets) next to the value of the natural logarithm of $\mu E\ (N, R)$ in Figure 1, The absence of a confidence interval means that the median complexity is obtained for K=16 random formulas.

4. Some Generalizations

The approximation by the least squares method of the experimental data presented in Figure 1 is expressed in the form:

$$y(4.3, N) = N/26.18 + b(4.3),$$
$$y(4.6, N) = N/30.09 + b(4.6),$$
$$y(5.0, N) = N/34.93 + b(5.0),$$
$$y(5.5, N) = N/41.44 + b(5.5),$$
$$y(6.0, N) = N/46.65 + b(6.0).$$

The coefficients $b(4.3)$, $b(4.6)$, $b(5.0)$, $b(5.5)$, and $b(6.0)$ are not of great interest in this context. It is convenient to introduce $B(R) = e^{b(R)}$. In a more visual form, the approximation of the exponential component of the estimation of complexity of the 3-CNF problem is expressed as:

$$\mu E(N, 4.3) \approx B(4.3) \cdot 2^{N/18.15} = B(4.3) \cdot 2^{N/D(4.3)},$$
$$\mu E(N, 4.6) \approx B(4.6) \cdot 2^{N/20.86} = B(4.6) \cdot 2^{N/D(4.6)},$$
$$\mu E(N, 5.0) \approx B(5.0) \cdot 2^{N/24.20} = B(5.0) \cdot 2^{N/D(5.0)},$$
$$\mu E(N, 5.5) \approx B(5.5) \cdot 2^{N/28.72} = B(5.5) \cdot 2^{N/D(5.5)},$$
$$\mu E(N, 6.0) \approx B(6.0) \cdot 2^{N/32.35} = B(6.0) \cdot 2^{N/D(6.0)}.$$

Where $B(4.3)$=2.267, $B(4.6)$=1.896, $B(5,0)$=2.020, $B(5.5)$=2.283 and $B(6.0)$=2.039.

Figure 2 illustrates the dependence of D the denominator in the exponent on R. Visually, all five points $D(4.3)$, $D(4.6)$, $D(5.0)$, $D(5.5)$ and $D(6.0)$ lie almost in a straight line. This suggests a linear relationship between D and R.

Representing $D(R)$ in the form $D = \beta R - \delta$ and one more apply the least squares method, we obtain $\beta = 8.4$, $\delta = 17.8$.

The median complexity (expressed in the number of branches of the analysis tree) of the proof of the unsatisfiability of 3-CNF formulas which are constructed using random numbers, obtained in the course of the computational experiment can be represented as:

$$\mu E(N, R) \approx B(R) \cdot 2^{N/(8.4R - 17.8)}.$$

In an early publication [6] devoted to the study of random 3-CNF formulas, the median value of the denominator of exponent at $R^{\sim} = 4.3$ was estimated as $D^{\sim}(4.3) = 17$, and for $R^{\sim} = 10$ the value $D^{\sim}(10) = 57$.

This estimate can be interpreted so that the algorithm used in [6] for solving random 3-CNF problem is characterized by $D^{\sim} = 7R$-13. Accordingly, $\beta^{\sim}$=7, $\delta^{\sim}$ =13.

The values of the β and δ coefficients depend on the used procedures of *local search* and *learning clauses* and can be used to compare SAT-solvers.

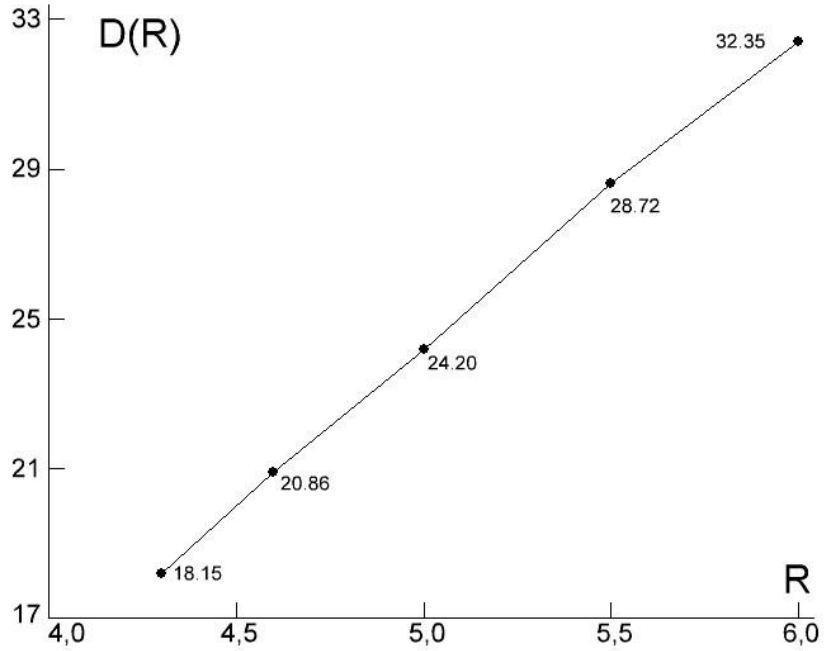

Figure 2. Graphical representation of the dependence of the denominator *D* of the exponent on *R*.

5. Conclusions

The study confirmed the well-known hypothesis that the most difficult to analyze random 3-CNF formulas are concentrated in the area of coexistence of both satisfiable and unsatisfiable examples. This area is defined by the range $3.52 \leq R \leq 4.51$.

In the experiment carried out, the interval of *R* values where both satisfiable and unsatisfiable formulas are simultaneously present ranges from 4.01 to 4.56. The interval width is 0.55.

It is very likely that the study of formulas in this area is most fruitful for answering the question about the polynomial or exponential nature of the SAT problem.

The random formulas obtained by the described method [1, 2], [4-6] turned out to be rather complicated. Indeed, for formulas with $N = 480$ associated with the second column of Table 1, the proof of the unsatisfiability requires the analysis of a tree containing more than 193,000,000 branches.

For a fairly advanced algorithm for proving the unsatisfiability of random 3-CNF formulas. in the range of the *N* number of variables from 256 to 512, the exponential nature of the dependence of the complexity of such a proof on *N* is revealed. This result correlates well with the numerous previous published results [1,2,6] obtained for CNF formulas (with fewer N number of variables).

Demonstrated strong dependence of the analysis complexity of formulas on *R*. This points the way for further improvement of the methods for producing formulas,

which are difficult to analyze. For example, in [7], a generator is proposed that balances the load on literals of variables by generating 3-CNF (partially random) formulas. For such generating the fraction of unsatisfiable formulas becomes predominant at $R \approx 3.5$.

As a result of a computational experiment, a linear dependence of the denominator of the exponent on R- the ratio of the number of clauses to the N number of variables.

In the research the exponential component of median complexity of unsatisfiability analysis of random 3-CNF formulas approximates by a function of N and R namely 2 to the power of $N/(8.4R-17.8)$, in the range $N=256\div512$, $R=4.3\div6.0$.

In the future, it is planned to try to prove that when using SAT-solvers built on the basis of *backtracking* with l*ocal search* and *learning clauses* procedures the complexity of SAT-problem is exponential.

References

[1] Biere A, Heule M, Maaren H, Walsh T. Handbook of Satisfiability. IOS Press; 2009. p. 1-966.
[2] Gomes C, Rautz H, Sabharwal A, Selman B. Satisfiability Solvers. in Handbook of Knowledge Representation, Elsevier B.V.; 2008. p. 88-134.
[3] Beame P, Kautz H, Sabharwal A. Towards understanding and harnessing the potential of clause learning. J. Artif. Intell. Res. 22(2004), p. 319-351.
[4] Kaporis Alexis C, Kirousis Lefteris M, Laias Efthimios G. The probabilistic analysis of a greedy satisability algorithm. Random Struct. Algor.; 2006. 28(40) p..444-480.
[5] Mohammad TH, Gregory S. The satisfiability threshold of random 3-SAT is at least 3.52. 2003/10/13. arXiv preprint math /0310193.
[6] Crawford J, Auton I. Experimental results on the crossover point in satisfiability problems. Proceeding of AAAI-93, Washington, DC. 1993. p. 21-27.
[7] Uvarov SI. An Improved generator for 3-CNF formulas. Automat. Rem. Contr.; 2020, 81(1), p. 95-103.