A general upper bound for the satisfiability threshold of random r-SAT formulæ

Olivier Dubois
LAFORIA, CNRS et Université Paris 6, France

October 23, 1995

[summary by Danièle Gardy]

Abstract

It is well known that the general problem of checking the satisfiability of a set of clauses is NP-complete. Experimentations have shown that there is a threshold on the ratio "number of clauses/number of variables" that separates the set of clauses for which a solution can be (easily) found from those for which it is impossible to find a solution. The subject of this talk is the r-SAT problem, in which the clauses have a constant number r of literals. This summary is based on [2].

1. The problem

A literal is either a boolean variable x_i or its negation $\bar{x}_i$. A *clause* is a disjunction of literals over a set of boolean variables; for example $x_1 \vee x_2 \vee \bar{x}_3 \vee x_4 \vee \bar{x}_5$ is a clause on the literals $x_1, \ldots, \bar{x}_5$. A formula is a finite set of clauses, or equivalently a conjunction of clauses. The *satisfiability problem* is to determine whether there exists a truth assignment (each literal is assigned a value *true* or *false*) satisfying a given formula. This famous problem is *NP*-complete as soon as the number n of literals is at least equal to 3; it was the first problem to be proved so [3, 4].

If we cannot find an algorithm that is guaranteed to work in polynomial time (worst-case complexity), what about the average complexity? This natural question leads to the notion of random clauses. The first point is to define a model of random clauses, i.e. a probability law on the set of all possible clauses on n literals. Two approaches have been attempted (in both, clauses are chosen independently of each other):

(1) *Constant density:* The literal x_i is present in a clause with probability p_i, its negation $\bar{x}_i$ is present with probability q_i, and the probability that neither x_i nor $\bar{x}_i$ are present is equal to $1 - p_i - q_i$.
(2) *Constant length:* The problem is restricted to all clauses of given length r; there are $C = 2^r\binom{n}{r}$ such clauses, and the probability distribution on this set is uniform: Each clause is chosen with a probability $1/C$.

We choose m clauses amongst C, with replacement. The first model leads to clauses of variable length; an easy analysis shows that, when the number m of clauses and the number n of variables are polynomially related, almost every formula is satisfiable.

The model under active study is the second one, the so-called r-SAT problem. Simulations have shown the importance of the ratio c_r = *Number of clauses/Number of variables:* If c_r is smaller than some threshold value, the probability of finding an assignment of the variables that satisfies the formula is close to 1 for $n, m \to +\infty$; if c_r is larger than this threshold, the probability of finding an assignment that satisfies the set of clauses is close to 0 for $n, m \to +\infty$. This threshold is an increasing function of r; experiments lead to believe that the value for $r = 3$ is $\rho = 4.25...$ Moreover, the backtracking algorithms used to solve r-SAT behave differently according to the ratio c_r. Experimentally, the difficulty of either finding an assignment satisfying a formula or proving that a formula is unsatisfiable is exponentially greater when c_r is close to the threshold than when it is either lower or greater.

The theoretical proof of the existence of a threshold value for the ratio $c_r = n/m$ lags behind. For 3-SAT, the best lower bound presently is 3.003, and the best upper bound is 4.64... (a result established precisely by Dubois and Boufkhad, and presented in this talk). There remains a gap between 3.003 and 4.64..., around the observed threshold 4.25.

2. The result

The main result is as follows:

> A random r-SAT formula ($r \geq 3$) is unsatisfiable with probability asymptotically close to 1, when $n \to +\infty$, as soon as $c_r := m/n$ is at least equal to some specified value $c_{r,min}$.

This lower bound $c_{r,min}$ is defined in terms of the solution x_0 of a transcendental equation, and can be computed numerically with the help of a Computer Algebra System. For $r = 3$, we get $x_0 = 1.924714266...$, which gives the bound $c_r \leq 4.642476157...$

For $r \geq 4$, the bound obtained by Dubois and Boufkhad improves on the general upper bound $c_r \leq -\log 2/\log(1 - 2^{-r})$. For example, with $r = 4$, some minutes of experiment with Maple give $x_0 = 2.69945696....$ and $c_4 \leq 10.2168796...$, which is a slight improvement on the known bound $c_r \leq -\log 2/\log(1 - 2^{-r}) = 10.74005367...$ For $r = 5$, we obtained $x_0 = 3.429641...$ and $c_r \leq 21.32022...$, which is still slightly better than the known bound $c_r \leq -\log 2/\log(1 - 2^{-r}) = 21.83230235...$ For $r = 10$, the known bound gives $c_r \leq 709.436...$, and Dubois's method gives $x_0 = 6.92993239...$ and $c_r \leq 708.935...$ These computations also show that the gain becomes marginal for large r. However, experiments seem to indicate that the difference between the bound $-\log 2/\log(1 - 2^{-r})$ and the threshold is slowly varying, and that the accuracy of the bound of Dubois and Boufkhad actually increases.

3. The proof

The proof relies on the existence of a special type of solutions, called *negatively prime solutions* (NPS), which are defined below, and to which is applied the method of the first moment. The idea behind this method is simple. To show that some problem has no solution, define X as the number of solutions of a random instance and show that the expectation $E[X]$ can be made as close to 0 as desired. This argument, applied to the r-SAT problem, leads to the following reasoning:

- Show that every satisfiable formula has at least one NPS (easy). The average number of NPS of a satisfiable formula is then at least 1.
- Compute the expectation $E[\text{NPS}]$ of the number of NPS on the set of random formulæ with n variables and m clauses.
- If $E[\text{NPS}] = 0$ then a random formula has no negatively prime solution, hence no solution.
- Then we should compute $E[\text{NPS}]$ and study its asymptotic behaviour as $n, m \to +\infty$ with $n/m = c_r$.

3.1. Negatively prime solutions. A solution of a formula F is defined as a set of n literals, each variable appearing either as x_i or as $\bar{x}_i$, such that the assignment of *true* to these literals satisfies F. A *negatively prime solution* is a solution such that, if we substitute x_i for a negative literal $\bar{x}_i$, the resulting set is no longer a solution of F.

It is easy to see that each solution of F either is a NPS, or leads to a NPS (by inverting negative literals as long as possible). Thus the number of solutions of F is greater than or equal to the number of negatively prime solutions; the same holds for expectations, and the method of the first moment, when applied to the number of NPS, will give a better bound than when applied to the number of solutions, as for example in [1, 5].

It is possible to define a *positively prime solution* (PPS) in a similar way (an assignment minimal for the substitution of $\bar{x}_i$ to x_i); as $E[\mathrm{NPS}] = E[\mathrm{PPS}]$, the bound obtained is exactly the same.

3.2. The expectation E[NPS]. Dubois and Boufkhad show that

$$E[\mathrm{NPS}] = \sum_{0 \le i \le j \le n} 2^{i-rm} \binom{n}{i} \binom{m}{j} i!\, S_{j,i} \left(\frac{r}{n}\right)^j (2^r - 1 - r)^{m-j}.$$

In this formula, $S_{j,i}$ is a Stirling number of second kind: $S_{j,i}$ is the number of ways to partition a set of j elements into i nonempty subsets.

In passing, they also remark that *for any set of literals* $\{l_i, i = 1, \ldots, n\}$ *($l_i = x_i$ or $l_i = \bar{x}_i$), there exists at least one formula that has this set as a NPS.*

The next step is to get an upper bound on E[NPS], using a bound on Stirling numbers due to Temme [6]:

$$E[\mathrm{NPS}] \le \left(\frac{2^r - r - 1}{2^r}\right)^{nc_r} + c_r\sqrt{2\pi}\, n^{5/2} e^{1/12n} A^n (1 + o(1)),$$

with A defined as the maximum of some function. The first term of the r.h.s. is $o(1)$ when $n \to +\infty$; the behaviour of the second term (and of the upper bound) is given by A^n. Then a concavity argument is used to prove that $A < 1$ for c_r greater than a value $c_{r,min}$ that can be precisely defined. This shows that, for $m/n > c_{r,min}$, $E[\mathrm{NPS}] \to 0$, i.e. a random formula cannot be satisfied.

This approach does not give any information for $m/n < c_{r,min}$; however a closer analysis (done by the authors, but not presented in [2]) shows that $E[\mathrm{NPS}] \ge Q(n)A^n$, with a polynomial factor $Q(n)$, and the same exponential basis A; hence E[NPS] is of exponential order A^n.

Bibliography

[1] de la Vega (W. F.) and El Maftouhi (A.). – On random 3-sat. *Combinatorics, Probability and Computing*, 1995, pp. 189–195.

[2] Dubois (O.) and Boufkhad (Y.). – *A general upper bound for the satisfiability threshold of random r-SAT formulæ.* – Technical report, LAFORIA, CNRS and University of Paris 6, 1996.

[3] Garey (M. R.) and Johnson (D. S.). – *Computers and intractability : A guide to the theory of NP-Completeness.* – Freeman, San Francisco, 1979.

[4] Johnson (D. S.). – *A catalog of complexity classes*, Chapter 2, pp. 67–161. – Elsevier, 1990.

[5] Kamath (A.), Motwani (R.), Palem (K.), and Spirakis (P.). – Tail bounds for occupancy and the satisfiability threshold conjecture. In *35th Annual Symposium on Foundations of Computer Science*, pp. 592–603. – 1994.

[6] Temme (N. M.). – Asymptotic estimates of Stirling numbers. *Studies in Applied Mathematics*, vol. LXXXIX, n° 3, 1993, pp. 233–244.