Typical Random 3-SAT Formulae and the Satisfiability Threshold

O. Dubois[1], Y. Boufkhad[2] and J. Mandler[1]

[1]LIP6, Box 169, CNRS-Université Paris 6, 4 place Jussieu, 75252 Paris Cedex 05, France.
[2]CRIL, Université d'Artois, rue de l'Université SP16, F-62307 Lens Cedex, France.
Olivier.Dubois@lip6.fr, boufkhad@cril.univ-artois.fr, Jacques.Mandler@lip6.fr.

3-SAT formulae are defined as finite sets of clauses, each clause being a disjunction of 3 literals over a set of boolean variables. Experiments on solving random 3-SAT formulae have provided strong evidence of a *phase transition phenomenon*. Explicitly, almost every random 3-SAT formula is satisfiable when its ratio: number of variables to number of clauses, denoted by c, is less than a value of about 4.25 and becomes unsatisfiable when this ratio exceeds 4.25. More formally, a critical value ω of the ratio c, the so-called satisfiability threshold, is thought to exist such that, for a sufficiently large number n of variables and for any real ε, the probability of satisfiability of a random 3-SAT formula, decreases from almost 1 to almost 0 as c increases from $\omega - \varepsilon$ to $\omega + \varepsilon$.

Calculating the value of ω, besides being a combinatorial challenge, has appeared as a way to understand better how the space of solutions of random formulae is structured. This is of practical importance for the design of efficient algorithms for solving SAT formulae.

Up to now, estimates of ω from above and below have been produced from a 'semantic' approach focusing on how a random 3-SAT formula can be satisfied by binary assignments (possibly of a particular kind) to its variables. Thus, two lower bounds of ω were calculated : 1.63 and 3.003 [2]. And starting from an easy upper bound, 5.19, three others were successively obtained : 4.76 [3], 4.643 [1] and 4.601 [4]. However this semantic approach may be subject to diminishing returns in view of the complexity of the calculations for the latest bounds.

In this paper, in order to estimate ω better, a new 'syntactic' approach is proposed. This is suggested, e.g., simply by practical experience with a computer-based random generator and solver of formulae. The generator is completely ignorant of semantics, yet for large n, it produces (within a realistic timeframe) either only satisfiable formulae, for values of c a little below 4.25, or only unsatisfiable formulae, for values of c a little above 4.25. Only a certain class of formulae is likely to come out of the generator, the 'typical' formulae. If we could describe (to a sufficient degree) their syntactic structure, the convergence to 0 or to $+\infty$ of the expectation of the number of solutions, computed only for these typical formulae, should give a very good indication as to the value of the threshold.

It turns out that for one of the usual probabilistic spaces of random 3-SAT formulae (all usual spaces are equivalent with respect to the threshold), such a partial syntactic characterization can be given fairly simply in terms of the numbers of occurrences and signatures of the variables. The basic result, obtained by probabilistic techniques borrowed from large deviation theory, is as follows.

For any integers $0 \leq p \leq x$, define $\kappa(x,p) = \left(\frac{\lambda}{2}\right)^x \frac{e^{-\lambda}}{p!(x-p)!}$ where $\lambda = 3c$. Let the random variable $\Upsilon_n(x,p)$ be the proportion of variables of a random formula, having x occurrences, among them p with a positive signature. Then, for any $\delta > 0$, $\lim_{n\to\infty} \mathbf{Pr}\left(|\Upsilon_n(x,p) - \kappa(x,p)| > \delta\right) = 0$.

This result is unfortunately still too far from a complete syntactic characterization of typical random formulae to give, by itself, an improvement on current bounds. However, we show that by combining it with the method of negatively prime solutions (NPSs), or, symmetrically, PPSs [1], the best current upper bound for ω can be improved significantly, yielding 4.506. The main steps required are as follows.

First, independently of the $\kappa(x,p)$'s, we define an *asymptotic distribution* of the numbers of occurrences and signatures per variable for 3-SAT formulae as a family Ξ of real numbers $\xi(x,p)$, $0 \le p \le x$, such that : $\sum_{x=0}^{+\infty}\sum_{p=0}^{x}\xi(x,p) = 1$ and $\sum_{x=0}^{+\infty}\sum_{p=0}^{x} x\xi(x,p) = 3c = \lambda$. To any such Ξ, any real ε and any integer x_{max}, we associate the set $\mathcal{F}(\Xi,\varepsilon,x_{max},n,c)$ of 3-SAT formulae such that for $0 \le p \le x \le x_{max}$ the number of variables having x occurrences, p of which are positive, lies between $(\xi(x,p)-\varepsilon)n$ and $(\xi(x,p)+\varepsilon)n$.
For any distribution Ξ as above, we estimate the expected number of PPSs of a random formula drawn from $\mathcal{F}(\Xi,\varepsilon,x_{max},n,c)$ uniformly. More precisely, we show that for n sufficiently large it is dominated by $[(1+\eta)H(\Xi,x_{max},c)]^n$, where η is a function of ε and x_{max} and where $H(\Xi,x_{max},c)$, given below, is expressed in terms of two quantities: ϕ, the normalized number of literals equal to 1 in the random formula for an appropriate assignment to the n variables, and β_1, the proportion of clauses satisfied by a single literal, ϕ and β_1 being solutions of a system of two equations also given below.

With the notations : $\beta_2 = 3(1-\phi) - 2\beta_1$, $\beta_3 = \beta_1 - 2 + 3\phi$, $U = \frac{9(1-\phi)\beta_3}{(3\phi-\beta_1)\beta_2}$, $V = 1 + \frac{\beta_2^2}{3(3\phi-\beta_1)\beta_3}$ and $\xi = \sum_{0\le p\le x\le x_{max}}(x-p)\xi(x,p)$, we have :

$$H(\Xi,x_{max},c) = C(\lambda)g(\phi,\beta_1)f(\phi,\beta_1)$$

where $C(\lambda)$ is an explicit constant depending only on λ and on the distribution Ξ,

$$g(\phi,\beta_1) = \prod_{0\le p\le x\le x_{max}}\left[U^{x-2p}(V^p-1)+V^{x-p}\right]^{\xi_{x,p}} \times \frac{U^{\lambda\phi-\xi}}{(V-1)^{\beta_1 c}},$$

and :

$$f(\phi,\beta_1) = \left\{\frac{[3(1-\phi)]^{3(1-\phi)}(3\phi-\beta_1)^{3\phi-\beta_1}}{\beta_2^{\beta_2}(3\beta_3)^{\beta_3}}\right\}^c.$$

The two equations satisfied by ϕ and β_1 are :

$$\lambda\phi = \xi - \sum_{1\le p\le x\le x_{max}}(x-2p)\xi(x,p)\times\left[1-\frac{1}{1+(U/V)^{x-2p}(1-V^{-p})}\right]$$

and :

$$\beta_1 c = \frac{V-1}{V}\sum_{0\le p\le x\le x_{max}}\xi(x,p)\frac{pU^{x-2p}V^p+(x-p)V^{x-p}}{U^{x-2p}(V^p-1)+V^{x-p}}.$$

On the basis of the asymptotic distribution $\kappa(x,p)$ corresponding to typical random 3-SAT formulae, we show that there exists a distribution Ξ_0 such that if the expected number of PPSs of the formulae in $\mathcal{F}(\Xi,\varepsilon,x_{max},n,c)$, tends to 0, then so does the probability of satisfiability of a random 3-SAT formula, and that for this particular distribution Ξ_0, an appropriate pair (x_{max},ε) and $c = 4.506$, the number $(1+\eta)H(\Xi_0,x_{max},c)$ is smaller than 1. It follows that for $c \ge 4.506$, the probability of satisfiability is asymptotically 0, hence the satisfiability threshold is less than 4.506.

References

[1] O. Dubois and Y. Boufkhad. A general upper bound for the satisfiability threshold of random r-sat formulae. *Journal of Algorithms*, 24:395–420, 1997.

[2] A. Frieze and S. Suen. Analysis of two simple heuristics on a random instance of k-SAT. *Journal of Algorithms*, 20(2):312–355, 1996.

[3] A. Kamath, R. Motwani, K. Palem, and P. Spirakis. Tail bounds for occupancy and the satisfiability threshold conjecture. *Random Structures and Algorithms*, 7(1):59–80, 1995.

[4] L. M. Kirousis, E. Kranakis, D. Krizanc, and Y. C. Stamatiou. Approximating the unsatisfiability threshold of random formulas. *Random Structures and Algorithms*, 12(3):253–269, 1998.