Random k-SAT: A tight threshold for moderately growing k

Alan Frieze[*]
Department of Mathematical Sciences,
Carnegie Mellon University,
Pittsburgh PA 15213, USA.
e-mail alan@random.math.cmu.edu

Nicholas C. Wormald[†].
Department of Mathematics and Statistics,
University of Melbourne
VIC 3010,
Australia.
e-mail nick@ms.unimelb.edu.au

February 12, 2002

Abstract

We consider a random instance I of k-SAT with n variables and m clauses, where $k = k(n)$ satisfies $k - \log_2 n \to \infty$. Let $m_0 = 2^k n \ln 2$ and let $\epsilon = \epsilon(n) > 0$ be such that $\epsilon n \to \infty$. We prove that

$$\lim_{n\to\infty} \mathbf{Pr}(I \text{ is satisfiable}) = \begin{cases} 1 & m \leq (1-\epsilon)m_0 \\ 0 & m \geq (1+\epsilon)m_0 \end{cases}$$

1 Introduction

An instance of k-SAT is defined by a set of variables, $V = \{x_1, x_2, \ldots, x_n\}$ and a set of clauses $C_1, C_2, \ldots, C_m$. We will let clause C_i be a *sequence* $(\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k})$ where each *literal* $\lambda_{i,l}$ is a member of $L = V \cup \bar{V}$ where $\bar{V} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$. In our random model, each $\lambda_{i,l}$ is chosen independently and uniformly from L. [1]

[*] Supported in part by NSF grant CCR-9818411

[†] Research supported in part by the Australian Research Council and in part by Carneegie Mellon University Funds

[1] We are aware that this allows clauses to have repeated literals or instances of $x, \bar{x}$. The focus of the paper is on $k = O(\ln n)$, although the main result is valid for larger k. Thus most clauses will not have repeated clauses or contain a pair $x, \bar{x}$. For moderate size k we could repeat the calculations for randomly chosen clauses without repeats or instances of $x, \bar{x}$. We doubt that this would change the nature of our main result, Theorem 1, but it would complicate its derivation. Of course, for $k > n$ we would be forced to repeat literals or introduce instances of $x, \bar{x}$ into each clause.

Random k-SAT has been well studied, to say the least. If $k = 2$ then it is known that there is a *satisfiability threshold* at around $m = n$. More precisely, if $\epsilon > 0$ is fixed and I is a random instance of 2-SAT then

$$\lim_{n\to\infty} \mathbf{Pr}(I \text{ is satisfiable}) = \begin{cases} 1 & m \leq (1-\epsilon)n \\ 0 & m \geq (1+\epsilon)n \end{cases}$$

This was proved in Chvatál and Reed [7] and sharpened by Goerdt [13], Fernandez de la Vega [9], Verhoeven [16] and Frieze and Sorkin [11]. The tightest results presently known are due to Bollobás, Borgs, Chayes, Kim and Wilson [3]. Thus random 2-SAT is now pretty much understood.

For $k \geq 3$ the story is very different. It is now known that a threshold for satisfiability exists in some (not completely satisfactory) sense, Friedgut [10]. There has been considerable work on trying to find estimates for this threshold in the case $k = 3$ – Chao and Franco [5, 6], Broder, Frieze and Upfal [4], Frieze and Suen [12], Achlioptas [1], Achlioptas and Sorkin [2], the last mentioned paper giving a lower bound of 3.26. Upper bounds have been pursued with the same vigour – Kirousis, Kramakis, Krizanc and Stamatiou [15], Janson, Stamatiou and Vamvakari [14], Dubois, Boufkhad and Mandler [8], the last-mentioned paper giving an upper bound of 4.506.

For larger values of k, even less is known. It was shown in [7] that if $m < \frac{2^k}{4k}n$ and k is constant then a random instance of k-SAT is satisfiable with probability tending to 1 and that if $m > 2^k n \ln 2$ then it is unsatisfiable with probability tending to 1 as $n \to \infty$. This is where it stands for such k. While the focus has been on constant k (in particular $k = 2, 3$) it is also worth considering $k \to \infty$. Sometimes allowing parameters to grow simplifies the problem and this is the case here. We prove the following *sharp* threshold:

Theorem 1. *Suppose $\omega = k - \log_2 n \to \infty$. Let*

$$m_0 = -\frac{n \ln 2}{\ln(1-2^{-k})} = (2^k + O(1))n \ln 2. \tag{1}$$

so that $2^n \left(1 - \frac{1}{2^k}\right)^{m_0} = 1$ and let $\epsilon = \epsilon(n) > 0$ be such that $\epsilon n \to \infty$. Let I be a random instance of k-SAT with n variables and m clauses. Then

$$\lim_{n\to\infty} \mathbf{Pr}(I \textit{ is satisfiable}) = \begin{cases} 1 & m \leq (1-\epsilon)m_0 \\ 0 & m \geq (1+\epsilon)m_0. \end{cases}$$

This sheds considerable light on the likely threshold for k fixed but large and we conjecture that the threshold here is $c_k n$ where $c_k \sim 2^k \ln 2$ (where $\sim$ is interpreted as $k \to \infty$ arbitrarily slowly). We also conjecture that the upper bound on the width of the scaling window implied by this theorem, $2^k \omega'$ for any $\omega' \to \infty$, is tight. The theorem says nothing about algorithms for finding satisfying assignments below the threshold or for proving unsatisfiability above the threshold. Are there polynomial time algorithms which work with high probability in this context?

2 Proof of Theorem 1

Our method of proof is quite straightforward. Let $X = X(I)$ denote the number of satisfying assignments for I. When $m \geq (1+\epsilon)m_0$ we show that $\mathbf{E}(X) \to 0$ and when $m \leq (1-\epsilon)m_0$ we use the second moment method to show that $\mathbf{Pr}(X > 0) \to 1$.

The upper bound: There are 2^n possible assignments of truth values to V. Let A_T denote the "all-true" assignment in which $x_j = T$ for $j = 1, 2, \ldots, n$. Assume that $m \geq (1+\epsilon)m_0$. Then

$$\begin{aligned} \mathbf{E}(X) &= 2^n \mathbf{Pr}(A_T \text{ satisfies } I) = 2^n \left(1 - \frac{1}{2^k}\right)^m = \left(1 - \frac{1}{2^k}\right)^{m-m_0} & (2) \\ &\leq \exp\left\{-\frac{m-m_0}{2^k}\right\} = 2^{-\epsilon n(1+o(1))} \to 0. \end{aligned}$$

The lower bound: Now assume that $m = (1-\epsilon)m_0$ where $\epsilon n \to \infty$ arbitrarily slowly. In particular, for concreteness, take

$$m = m_0(1 - O(\ln n/n)). \tag{3}$$

It is sufficient to consider this case, since the result for larger ϵ will follow by monotonicity.

First observe that

$$\mathbf{E}(X) = 2^{\epsilon n(1+o(1))} \to \infty.$$

We use the inequality

$$\mathbf{Pr}(X > 0) \geq \frac{\mathbf{E}(X)^2}{\mathbf{E}(X^2)}. \tag{4}$$

For this we need to estimate $\mathbf{E}(X^2)$. We find (as explained below),

$$\mathbf{E}(X^2) = 2^n \sum_{t=0}^{n} \binom{n}{t} \left(1 - \frac{2}{2^k} + \left(\frac{t}{2n}\right)^k\right)^m \tag{5}$$

and so by (2)

$$\begin{aligned} \frac{\mathbf{E}(X^2)}{\mathbf{E}(X)^2} &= 2^{-n} \sum_{t=0}^{n} \binom{n}{t} \left(\frac{1 - \frac{2}{2^k} + \left(\frac{t}{2n}\right)^k}{\left(1-\frac{1}{2^k}\right)^2}\right)^m & (6) \\ &= 2^{-n} \sum_{t=0}^{n} \binom{n}{t} \left(1 + \frac{\left(\frac{t}{2n}\right)^k - \frac{1}{2^{2k}}}{\left(1-\frac{1}{2^k}\right)^2}\right)^m. & (7) \end{aligned}$$

Explanation of (5): We let t denote the number of j for which $x_j = T$ in some assignment A and then consider the probability that both A_T and A are satisfying assignments. For a fixed j, if we choose clause j at random, the probability that at least one of A, A_T does not satisfy C_j is precisely $\frac{2}{2^k} - \left(\frac{t}{2n}\right)^k$. Finally, multiply by 2^n for the same reason as in (2).

Let u_t denote the tth term of the sum in (7). Then using Stirling's formula in the form $s! = (s/e)^s\sqrt{2\pi s}e^{\sigma/(12s)}$ where $|\sigma| \leq 1$ we obtain

$$\ln u_t \leq n \ln n - t \ln t - (n-t)\ln(n-t) + m\left(\frac{t}{2n}\right)^k + O\left(\frac{m}{2^{2k}}\right).$$

We put $t = \tau n$ and focus on the function

$$f(\tau) = -\tau \ln \tau - (1-\tau)\ln(1-\tau) + \alpha\tau^k \tag{8}$$

where

$$\alpha = m/(2^k n) = \ln 2 + O(\ln n/n) \tag{9}$$

by (1) and (3). Then

$$u_t \le e^{nf(t/n)}(1+o(1)) \tag{10}$$

uniformly for t in the range $[0,n]$. For various ranges of t, we will bound u_t from above either directly or using f.

Differentiating (8) with respect to τ we get

$$f'(\tau) = \ln\frac{1-\tau}{\tau} + \alpha k\tau^{k-1}. \tag{11}$$

We then parameterise $\tau = \frac{1+\beta}{2}$ and search for zeros of

$$g(\beta) = f'\left(\frac{1+\beta}{2}\right) = \ln\left(\frac{1-\beta}{1+\beta}\right) + \frac{\alpha k}{2^{k-1}}(1+\beta)^{k-1}.$$

Differentiating this with respect to β,

$$g'(\beta) = -\frac{2}{1-\beta^2} + \frac{\alpha k(k-1)}{2^{k-1}}(1+\beta)^{k-2}. \tag{12}$$

Note also that

$$g'(\beta) = \frac{\alpha k}{2^{k-1}} - \left(2 - \frac{\alpha k(k-1)}{2^{k-1}}\right)\beta + O\left(\beta^2\right) \qquad \beta \to 0. \tag{13}$$

It follows from (12) that f is strictly concave in the range $[0,\tau_2]$, $\tau_2 = \frac{1+\beta_2}{2}$, $\beta_2 = 1 - \frac{5\ln k}{k}$, since then $(1+\beta)^{k-2} < 2^k/k^2$ (k sufficiently large). Within this interval there is by (13) a unique maximum occurring at $\tau_0 = \frac{1+\beta_0}{2}$ where

$$\beta_0 = \frac{\alpha k}{2^k} + O\left(\frac{k^3}{2^{2k}}\right).$$

Having established the location of this maximum, we proceed by showing that "near" $t = \frac{1+\beta_0}{2}n$, u_t behaves like the corresponding binomial coefficient, (14), (15). Other values of u_t for t in the interval $[0,\tau_2]$ will be shown to be negligible by computing the values of f "near" $\frac{1+\beta_0}{2}n$ and using the concavity of f. We then have only to show then that the contributions of u_t, $t \ge \tau_2 n$, are also negligible.

From the definition of u_t as the term in (7) we see that for $t = \frac{1+\beta}{2}n$, $|\beta| \le n^{-1/2}\ln n$,

$$u_t = \binom{n}{t}\left(1 + O\left(\frac{km\ln n}{n^{1/2}2^{2k}}\right)\right) = \binom{n}{t}(1+o(1)) \tag{14}$$

when $k = O(\ln n)$, whilst for $k >> \ln n$

$$u_t = \binom{n}{t}\left(1 + O\left(\left(\frac{1+\beta}{4}\right)^k\right)\right)^m = \binom{n}{t}\exp\left(O\left(m\left(\frac{1+\beta}{4}\right)^k\right)\right) = \binom{n}{t}(1+o(1)). \tag{15}$$

Furthermore, if $\beta_1 = \pm n^{-1/2}\ln n$ then for some $\tilde{\beta}$ between 0 and β_1,

$$\begin{aligned} f\left(\frac{1+\beta_1}{2}\right) &= f\left(\frac{1}{2}\right) + \frac{1}{2}g'(0)\beta_1^2 + \frac{1}{6}g''(\tilde{\beta})\beta_1^3 \\ &= f\left(\frac{1}{2}\right) - \beta_1^2 + O\left(\left(\frac{k^2}{2^k}\right)\beta_1^2 + \beta_1^3\right) \\ &\le f\left(\frac{1}{2}\right) - \frac{(\ln n)^2}{2n} = \log 2 - \frac{(\ln n)^2}{2n} + o(1/n), \end{aligned}$$

where we used (12) for the second step and (8) for the last.

Thus, by the concavity of f on the interval $[0, \tau_2]$ and by (10), $u_t \leq e^{nf((1+\beta_1)/2)} = o(2^n/n)$ for $t \leq \tau_2 n$ such that $|t - n/2| \geq n^{1/2} \ln n$. So, using (14), with $t_2 = \lfloor \tau_2 n \rfloor$,

$$\sum_{t=0}^{t_2} u_t \leq (1 + o(1)) \sum_{t=0}^{t_2} \binom{n}{t} + o(2^n/n)(t_2 + 1) \leq (1 + o(1))2^n. \tag{16}$$

Now let $t_3 = \lfloor \left(1 - \frac{1}{k}\right) n \rfloor$ and let $t = (1 - \theta)n \in [t_2 + 1, t_3]$. Then, from (7),

$$\begin{aligned} u_t &\leq \binom{n}{t} \left(1 + \frac{(1-\theta)^k - 1}{(2^k - 1)(1 - 2^{-k})}\right)^m \\ &\leq \exp\left(n\left(\theta \ln\left(\frac{e}{\theta}\right) + \left(\frac{m(1-\theta)^k}{2^k n}\right)\left(1 + O(2^{-k})\right)\right)\right) \\ &\leq \exp\left(n\left(\theta \ln\left(\frac{e}{\theta}\right) + (1-\theta)^k \ln 2 \left(1 + O(2^{-k})\right)\right)\right) \\ &\leq 2^n \exp(-n(1 - e^{-1} + o(1)) \ln 2) \end{aligned}$$

where the second-last step uses (1) and the last step uses $\theta \ln\left(\frac{e}{\theta}\right) = o(1)$ and $(1 - \theta)^k \leq (1 - 1/k)^k \leq e^{-1}$. Thus

$$\sum_{t=t_2+1}^{t_3} u_t = o(2^n). \tag{17}$$

Now for $t \geq t_3 + 1$, $t = (1 - \theta)n$, we have $\theta < 1/k$, and (11) gives

$$f'(1-\theta) = \ln\theta - \ln(1-\theta) + \alpha k(1-\theta)^{k-1} \geq \ln\theta - \ln(1-\theta) + \alpha k/e$$

since $\ln(1 - 1/k) > -1/(k - 1)$. So, clearly $f'(1 - \theta) \geq \alpha k/50$ for $\theta \geq e^{-\alpha k/3}$. Putting $t_4 = \min\{n(1 - e^{-\alpha k/3}), n - 1\}$ it follows that

$$f(t/n) \leq f(t_4/n) - \frac{\alpha k}{50}(t_4 - t)/n \qquad t_3 \leq t \leq t_4.$$

Consequently, since $k \to \infty$ and $\alpha \sim \ln 2$, (10) implies that

$$\sum_{t=t_3+1}^{\lfloor t_4 \rfloor} u_t \leq (1 + o(1))e^{nf(t_4/n)}. \tag{18}$$

Before proceeding, we note that

$$nf(1) = n\alpha = m/2^k = (1 - \epsilon)m_0/2^k = (1 + O(2^{-k}))n(1 - \epsilon)\ln 2$$

and so

$$e^{nf(1)} = o(2^n). \tag{19}$$

Similarly for n sufficiently large

$$f(1 - 1/n) \leq \frac{\ln n}{n} + \alpha \exp(-k/n) \leq \frac{\ln n}{n} + \alpha\left(1 - \frac{\log_2 n}{n} + O\left(\frac{\ln^2 n}{n^2}\right)\right) = \alpha + O\left(\frac{\ln^2 n}{n^2}\right)$$

by (9). Hence as before

$$e^{nf(1-1/n)} = o(2^n). \tag{20}$$

Case 1: $t_4 = n-1$.
In this case we use (10), (18) and (19) to obtain

$$\sum_{t=t_3+1}^{n} u_t = o(2^n). \tag{21}$$

Case 2: $t_4 < n-1$.
Then $e^{\alpha k/3} < n$. For $\theta \leq e^{-\alpha k/3}$ we see that

$$f'(1-\theta) = \ln\theta + \alpha k + O(k^2 e^{-\alpha k/3}).$$

Consequently,

$$\theta \geq \frac{1}{n} \text{ implies } f'(1-\theta) \geq \ln\left(\frac{2^k}{n}\right) + O(k^2 e^{-\alpha k/3}) \to \infty.$$

So (10) and (18) imply

$$\sum_{t=t_3}^{n} u_t = (1+o(1))e^{nf(1-1/n)} = o(2^n) \tag{22}$$

by (20). The proof of the lower bound now follows from (4), (7), (16), (17), (21) and (22). □

References

[1] D. Achlioptas, Setting two variables at a time yields a new lower bound for random 3-SAT, *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (2000) 28–37.

[2] D. Achlioptas and G. Sorkin, Optimal myopic algorithms for random 3-SAT, *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computing* (2000) 590–600.

[3] B. Bollobás, Christian Borgs, Jennifer Chayes, J.H. Kim, and D.B. Wilson, The scaling window of the 2-SAT transition, *Random Structures and Algorithms* (2001) 201–256.

[4] A.Z. Broder, A.M. Frieze and E. Upfal, On the satisfiability and maximum satisfiability of random 3-CNF formulas, *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, (1993) 322–330.

[5] M.T. Chao and J. Franco, Probabilistic analysis of two heuristics for the 3-satisfiability problem, *SIAM Jornal on Computing* 15 (1986) 1106–1118.

[6] M.T. Chao and J. Franco, Probabilistic analysis of a generalization of the unit-clause literal selection heuristics for the k satisfiable problem, *Information Science* 51 (1990) 289–314.

[7] V. Chvatál and B. Reed, Mick gets some (the odds are on his side), *Proceedings of the 33rd Annual IEEE Symposium on the Foundations of Computer Science* (1992) 620–627.

[8] O. Dubois, Y. Boufkhad and J. Mandler, Typical random 3-SAT formulae and the satisfiability threshold, *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms* (2000) 126–127.

[9] W. Fernandez de la Vega, On random 2-SAT, manuscript 1992.

[10] E. Friedgut, Sharp thresholds of graph properties, and the k-sat problem. With an appendix by Jean Bourgain. *Journal of the American Mathematical Society* 12 (1999) 1017–1054.

[11] A.M. Frieze and G. Sorkin, A note on random 2-SAT with prescribed literal degrees, to appear in SODA 2002.

[12] A.M. Frieze and S. Suen, Analysis of Two Simple Heuristics on a Random Instance of k-SAT, *Journal of Algorithms* 20 (1996) 312–355.

[13] A. Goerdt, A threshold for unsatisfiability, *Journal of Computer and System Sciences* 33 (1996) 469–486.

[14] S. Janson, Y. Stamatiou and M. Vamvakari, Bounding the unsatisfiability threshold of random 3-SAT, *Random Structures Algorithms* 17 (2000) 103–116.

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

[16] Y. Verhoeven, Random 2-SAT and unsatisfiability, *Information Processing Letters* 72 (1999) 119–123.

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.