Random k-SAT: the limiting probability for satisfiability for moderately growing k

Amin Coja-Oghlan*
Alan Frieze†
Department of Mathematical Sciences,
Carnegie Mellon University,
Pittsburgh PA 15213, USA.
acoghlan@inf.ed.ac.uk,alan@random.math.cmu.edu

Submitted: Sep 11, 2007; Accepted: Jan 17, 2008; Published: Feb 4, 2008
Mathematics Subject Classification: 05C88

Abstract

We consider a random instance $I_m = I_{m,n,k}$ of k-SAT with n variables and m clauses, where $k = k(n)$ satisfies $k - \log_2 n \to \infty$. Let $m = 2^k(n \ln 2 + c)$ for an absolute constant c. We prove that

$$\lim_{n\to\infty} \mathbf{Pr}(I_m \text{ is satisfiable}) = e^{-e^{-c}}.$$

1 Introduction

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

*Supported by DFG COJ 646.
†Supported in part by NSF grant CCF-0502793
[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 $x, \bar{x}$.

Random k-SAT has been well studied, to say the least, see the references in [6]. 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

$$\lim_{n\to\infty} \mathbf{Pr}(I_{m,n,2} \text{ is satisfiable}) = \begin{cases} 1 & m \leq (1-\epsilon)n \\ 0 & m \geq (1+\epsilon)n \end{cases}$$

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 [5]. There has been considerable work on trying to find estimates for this threshold in the case $k = 3$, see the references in [6]. Currently the best lower bound for the threshold is 3.52, due to Hajiaghayi and Sorkin [7] and Kaporis, Kirousis, and Lalas [8]. Upper bounds have been pursued with the same vigour. Currently the best upper bound for the threshold is 4.506 due to Dubois, Boufkhad and Mandler [4].

Building upon Achlioptas and Moore [1], Achlioptas and Peres [3] made a considerable breakthrough for $k \geq 4$. Using a sophisticated secnd moment argument, they showed that if $m \leq (2^k \ln 2 - t_k)n$ then **whp** a random instance of k-SAT $I_{m,n,k}$ is satifiable, where $t_k = O(k)$. Since a simple first moment argument shows that $I_{m,n,k}$ is unsatisfiable if $m > (2^k \ln 2 + o(1))n$, they have obtained an asymptotically tight estimate of the threshold for satisfiability when k is a large constant.

An earlier paper by Frieze and Wormald [6] showed the following: Suppose $\omega = k - \log_2 n \to \infty$. Let

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

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_m be a random instance of k-SAT with n variables and m clauses. Then

$$\lim_{n\to\infty} \mathbf{Pr}(I_m \text{ is satisfiable}) = \begin{cases} 1 & m \leq (1-\epsilon)m_0 \\ 0 & m \geq (1+\epsilon)m_0. \end{cases} \tag{2}$$

The aim of this short note is to tighten (2) and prove the following.

Theorem 1. *Suppose $\omega = k - \log_2 n \to \infty$ but $\omega = o(\ln n)$. Let $m = 2^k(n \ln 2 + c)$ for an absolute constant c. Then*

$$\lim_{n\to\infty} \mathbf{Pr}(I_m \textit{ is satisfiable}) = 1 - e^{-e^{-c}}.$$

Theorems such as this are common in random graphs and usually indicate that the threshold for a certain property $\mathcal{P}_1$ depends on the occurrence of some much simpler property $\mathcal{P}_2$, a classic example being the case where $\mathcal{P}_1$ is Hamiltonicity and $\mathcal{P}_2$ is minimum degree at least two. Here there does not seem to be a good candidate for $\mathcal{P}_2$.

2 Proof of Theorem 1

Let $X_m = X(I_m)$ denote the number of satisfying assignments for instance I_m. Suppose that $k = \log_2 n + \omega$. Let $m_0 \sim 2^k n \ln 2$ be as in (1) and $m_1 = m_0 - 2^k\gamma$, where $\gamma = \ln \omega$. The following results can be deduced from the calculations in [6]: If σ_1, σ_2 are two assignments to the variables V, then $h(\sigma_1, \sigma_2)$ is the number of indices i for which $\sigma_1(i) \neq \sigma_2(i)$ (i.e., the Hamming distance of σ_1 and σ_2).

P1 $X_{m_1} \sim \mathbf{E}(X_{m_1}) \sim 2^n(1-2^{-k})^{m_1} = e^\gamma$ **whp**.

P2 Let Z_t denote the number of pairs of satisfying assignments σ_1, σ_2 for which $h(\sigma_1, \sigma_2) = t$. Then **whp** $Z_t = 0$ for $0 < t < 0.49n$.

Because these properties are not explicitly spelled out in [6], in Section 3 we indicate briefly how they can be demonstrated using the arguments in this reference. We defer their verification until Section 3 and now show how they can be used to prove Theorem 1.

We generate our instance I_m by first generating I_{m_1} and then adding the $m - m_1$ random clauses $J = \{C_1, C_2, \ldots, C_{m-m_1}\}$. Suppose that in this case I_{m_1} has satisfying assignments $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$, where by **P1** we can assume that $r \sim e^\gamma$. Now add the random clauses J and let $Y = |\{i : \ \sigma_i \text{ satisfies } J\}|$. We show that for any fixed positive integer t,

$$\mathbf{E}(Y_{(t)}) \sim e^{-ct}, \tag{3}$$

where $Y_{(t)} = \prod_{j=0}^{t-1}(Y-j)$ signifies the t'th falling factorial. Thus by standard results, Y is asymptotically Poisson with mean e^{-c} and Theorem 1 follows.

Proof of (3)**:** Since each of the clauses $C_1, \ldots, C_{m-m_1}$ is chosen independently of all others, we have

$$\mathbf{E}(Y_{(t)}) = r_{(t)}\mathbf{Pr}(\sigma_1, \ldots, \sigma_t \text{ satisfy } J) = r_{(t)}\mathbf{Pr}(\sigma_1, \ldots, \sigma_t \text{ satisfy } C_1)^{m-m_1}. \tag{4}$$

Now

$$\mathbf{Pr}(\sigma_1, \ldots, \sigma_t \text{ satisfy } C_1) \quad = \quad 1 - \mathbf{Pr}(\exists 1 \leq i \leq t : \ \sigma_i \text{ does not satisfy } C_1),$$

and

$$\mathbf{Pr}(\exists 1 \leq i \leq t : \ \sigma_i \text{ does not satisfy } C_1) \quad \leq \quad t\mathbf{Pr}(\sigma_1 \text{ does not satisfy } C_1) = \frac{t}{2^k}.$$

On the other hand, by inclusion/exclusion

$$\mathbf{Pr}(\exists 1 \leq i \leq t : \ \sigma_i \text{ does not satisfy } C_1) \geq t\mathbf{Pr}(\sigma_1 \text{ does not satisfy } C_1) - \sum_{1 \leq i < j \leq t} \mathbf{Pr}(\sigma_i, \sigma_j \text{ do not satisfy } C_1).$$

We then write

$$\begin{aligned}
&\mathbf{Pr}(\sigma_i, \sigma_j \text{ do not satisfy } C_1) \\
&= \mathbf{Pr}(\sigma_i, \sigma_j \text{ do not satisfy } C_1 \mid \mathbf{P2})\mathbf{Pr}(\mathbf{P2}) + \mathbf{Pr}(\sigma_i, \sigma_j \text{ do not satisfy } C_1 \mid \neg\mathbf{P2})\mathbf{Pr}(\neg\mathbf{P2}) \\
&= \left(\frac{n-\tau}{2n}\right)^k + o(1) \le \frac{1}{3^k}
\end{aligned}$$

Finally, going back to (4), we obtain

$$r_{(t)}\left(1 - \frac{t}{2^k}\right)^{m-m_1} \le \mathbf{E}(Y_{(t)}) \le r_{(t)}\left(1 - \frac{t}{2^k} + \frac{t^2}{3^k}\right)^{m-m_1}.$$

Since $t^2(m - m_1) = O(m - m_1) = O(\omega 2^k) = o(3^k)$, we get

$$\mathbf{E}(Y_{(t)}) \sim r_{(t)}\left(1 - \frac{t}{2^k}\right)^{m-m_1} \sim e^{t\gamma}\left(1 - 2^{-k}\right)^{t(m-m_1)} \sim e^{-ct},$$

thereby proving (3). □

3 Verification of P1 and P2

P1: Let us first compute the expected number $\mathbf{E}(X_{m_1})$ of satisfying assignments of I_{m_1}. For any fixed assignment the probability that a single random clause over k distinct variables is satisfied equals $1 - 2^{-k}$ (because there are 2^k ways to assign values to the k variables occurring in the clause, out of which $2^k - 1$ cause the clause to be satisfied). Since the m_1 clauses are chosen independently, and as there are 2^n assignments in total, we conclude that $E(X_{m_1}) \sim 2^n(1 - 2^{-k})^{m_1}$. Furthermore, in [6, Section 2] it is shown that $\mathbf{E}(X_{m_1}^2) \sim \mathbf{E}(X_{m_1})^2$ and so **P1** follows from the Chebyshev inequality.

P2: If σ_1, σ_2 are two assigments at Hamming distance $h(\sigma_1, \sigma_2) = t$, then the probability that either σ_1 or σ_2 does not satisfy a random clause C_1 is $2^{1-k} - 2^{-k}(1 - t/n)^k$. For the probability that *one* assignment σ_i does not satisfy C_1 is 2^{-k} $(i = 1, 2)$. Moreover, if both σ_1 and σ_2 violate C_1, then C_1 is false under σ_1, which occurs with probability 2^{-k}, and in addition σ_1 and σ_2 assign the same values to all the variables in C_1, which happens with probability $(1 - t/n)^k$. Consequently, the expected number of *satisfying* assignment pairs σ_1, σ_2 at Hamming distance t in I_{m_1} is

$$F(t) = \mathbf{E}(Z_t) = 2^n\binom{n}{t}(1 - 2^{1-k} + 2^{-k}(1 - t/n)^k)^{m_1}$$

(cf. [6, eq. (5)]). Setting $\rho = m_1/n = 2^k(\ln 2 - \gamma/n) + O(1/n)$, $\tau = t/n$ and taking logarithms, we obtain

$$\begin{aligned}
f(\tau) &= n^{-1}\ln F(t) \\
&\le \ln 2 - \tau\ln\tau - (1-\tau)\ln(1-\tau) + \rho\ln(1 - 2^{1-k} + 2^{-k}(1-\tau)^k) + O(\tau/n) \\
&\le \ln 2 - \tau\ln\tau - (1-\tau)\ln(1-\tau) - 2^{-k}\rho(2 - (1-\tau)^k) + O(\tau/n) \\
&= \ln 2 - \tau\ln\tau - (1-\tau)\ln(1-\tau) - (\ln 2 - \gamma/n)(2 - (1-\tau)^k) + O((\tau + 2^{-k})/n). \qquad (5)
\end{aligned}$$

To show that $\sum_{1\le t\le 0.49n} F(t) = o(1)$, we consider three cases:

Case 1: $n^{-1} \le \tau \le \ln^{-1.1} n$**.** Since $(1-\tau)^k = 1 - k\tau + O(k^2\tau^2)$, $-(1-\tau)\ln(1-\tau) \le \tau$, and $k\ln 2 = \ln n + \omega \ln 2$, we obtain via (5),

$$\begin{aligned} f(\tau) &\le \tau(1-\ln\tau) - k\tau\ln 2(1 - O(k\tau)) + 2\gamma/n \\ &\le \tau\left(1 + \ln n - (\ln n + \omega\ln 2) + o(1)\right) \\ &\le -\tau\omega/2. \end{aligned}$$

Consequently,

$$\sum_{1\le t\le n\ln^{-1.1} n} F(t) = \sum_{1\le t\le n\ln^{-1.1} n} \exp(nf(t/n)) \le \sum_{1\le t\le n\ln^{-1.1} n} \exp(-\omega t/2) = o(1). \quad (6)$$

Case 2: $\ln^{-1.1} n < \tau \le k^{-1}\ln\ln n$**.** We have, for large n,

$$-\tau\ln\tau - (1-\tau)\ln(1-\tau) \le \tau(1-\ln\tau) \le \frac{(1+\ln k)\ln\ln n}{k} \le k^{-\frac{1}{2}} \le \ln^{-\frac{1}{2}} n.$$

On the other hand, for large n,

$$(1-\tau)^k \le \exp(-k\tau) \le \exp(-k\ln^{-1.1} n) \le 1 - \ln^{-0.1} n.$$

Thus, from (5),

$$f(\tau) \le \ln 2 + \ln^{-\frac{1}{2}} n - \ln 2 - \frac{\ln 2}{\ln^{0.1} n} \le -\frac{1}{2}\ln^{-0.1} n.$$

Hence, if $n\ln^{-1.1} n < t \le nk^{-1}\ln\ln n$, then $F(t) \le \exp(-\frac{1}{2}n\ln^{-0.1} n)$, which implies

$$\sum_{n\ln^{-1.1} n < t \le nk^{-1}\ln\ln n} F(t) = o(1). \quad (7)$$

Case 3: $k^{-1}\ln\ln n < \tau \le 0.49$**.** Since $\tau \gg k^{-1}$, we have $(1-\tau)^k = o(1)$, whence

$$(\ln 2 - \gamma/n)(2 - (1-\tau)^k) \sim 2\ln 2.$$

Furthermore, as the entropy function $\tau \mapsto -\tau\ln\tau - (1-\tau)\ln(1-\tau)$ is increasing on $[0, \frac{1}{2}]$, we have

$$\ln 2 - \tau\ln\tau - (1-\tau)\ln(1-\tau) \le \ln 2 - 0.49\ln(0.49) - 0.51\ln(0.51) < 1.9998\ln 2.$$

Hence, $f(\tau) \le -0.0001$. Therefore, $F(t) \le \exp(-0.0001n)$, and thus

$$\sum_{nk^{-1}\ln\ln n < \tau \le 0.49n} F(t) = o(1). \quad (8)$$

Combining (6)–(8), we conclude that $\sum_{1\le t\le 0.49n} F(t) = o(1)$. Thus, **whp** $Z_t = 0$ for all $1 \le t \le 0.49$.

4 Conclusion

It is instructive to compare the k-SAT problem with $k > \log_2 n + \omega$, which we have studied in the present paper, with the case of constant k. We have shown that for $k > \log_2 n + \omega$ in the regime $m/n - 2^k n \ln 2 = \Theta(2^k)$ the number of satisfying assignments is asymptotically Poisson. The basic reason is that the mutual Hamming distance of any two satisfying assignments is about $n/2$ (cf. property **P2**). Hence, the set of all satisfying assignments consists of isolated points in the Hamming cube, which are mutually far apart. By contrast, in the case of constant k in the near-threshold regime the set of satisfying assignments seems to consist of larger "cluster regions" (cf. Achlioptas and Ricci-Tersenghi [2] and Krzakala, Montanari, Ricci-Tersenghi, G. Semerjian, and L. Zdeborova [9]).

In Theorem 1 we assume that $\omega = k - \log_2 n = o(\ln n)$. While this assumption eases some of the computations, the result (and the proof technique) can be extended to larger values of k. Nevertheless, the case $k < \log_2 n$ appears to us to be a more interesting problem.

References

[1] D. Achlioptas and C. Moore: Random k-SAT: two moments suffice to cross a sharp threshold. *SIAM Journal on Computing* 36 (2006) 740–762.

[2] D. Achlioptas and F. Ricci-Tersenghi: On the solution-space geometry of random constraint satisfaction problems. *Proceedings of the 38th Annual ACM Symposium on Theory of Computing* (2006) 130–139.

[3] D. Achlioptas and Y. Peres, The threshold for random k-SAT is $2^k - \log 2 - O(k)$, *Journal of the American Mathematical Society*, 17 (2004), 947-973.

[4] 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.

[5] 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.

[6] A.M. Frieze and N. Wormald, Random k-SAT: A tight threshold for moderately growing k, Combinatorica 25 (2005) 297-305.

[7] M.T. Hajiaghayi and G.B. Sorkin, The satisfiability threshold of random 3-SAT is at least 3.52. IBM Research Report RC22942 (2003)

[8] A.C. Kaporis, L.M. Kirousis, and E.G. Lalas: Selecting complementary pairs of literals. *Electronic Notes in Discrete Mathematics* 16 (2003)

[9] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborova, Gibbs states and the set of solutions of random constraint satisfaction problems. Preprint (arXiv:cond-mat/0612365).