On the Maximum Satisfiability of Random Formulas

Dimitris Achlioptas*
Microsoft Research, Redmond, Washington
optas@microsoft.com

Assaf Naor
Microsoft Research, Redmond, Washington
anaor@microsoft.com

Yuval Peres†
Departments of Statistics and Mathematics, University of California, Berkeley
peres@stat.berkeley.edu

Abstract

Maximum satisfiability is a canonical NP-complete problem that appears empirically hard for random instances. At the same time, it is rapidly becoming a canonical problem for statistical physics. In both of these realms, evaluating new ideas relies crucially on knowing the maximum number of clauses one can typically satisfy in a random k-CNF formula. In this paper we give asymptotically tight estimates for this quantity.

*Let us say that a k-CNF formula is p-*satisfiable *if there exists a truth assignment satisfying $1-2^{-k}+p2^{-k}$ fraction of all clauses (every k-CNF is 0-satisfiable). Let $F_k(n,m)$ denote a random k-CNF formula on n variables formed by selecting uniformly, independently and with replacement m out of all $(2n)^k$ possible k-clauses. Finally, let $\tau(p) = 2^k \ln 2/(p + (1-p)\ln(1-p))$.*

It is easy to prove that for every $k \geq 2$ and $p \in (0,1]$, if $r \geq \tau(p)$ then the probability that $F_k(n, m = rn)$ is p-satisfiable tends to 0 as n tends to infinity. We prove that there exists a sequence $\delta_k \to 0$ such that if $r \leq (1-\delta_k)\tau(p)$ then the probability that $F_k(n, m = rn)$ is p-satisfiable tends to 1 as n tends to infinity. The sequence δ_k tends to 0 exponentially fast in k. Indeed, even for moderate values of k, e.g. $k = 10$, our result gives very tight bounds for the fraction of satisfiable clauses in a random k-CNF. In particular, for $k > 2$ it improves upon all previously known such bound.

1. Introduction

Given a Boolean CNF formula F, the Satisfiability problem is to determine whether there exists a truth assignment that satisfies F. When F has exactly k literals in each clause, Satisfiability is known as k-SAT and it is NP-complete [Coo71] for all $k \geq 3$. A natural generalization of Satisfiability is determining whether there exists a truth assignment that satisfies a given number of clauses in F. For k-CNF this problem is known as Max k-SAT and it is NP-complete for all $k \geq 2$ (see [GJ79]).

Optimization problems with random inputs are pervasive in operations research (e.g., TSP for random point sets), statistical physics (determining ground states of spin glasses) and computer science. Historically, the motivation for studying random instances has been the desire to understand the behavior of "typical" instances. For problems defined on formulas, random k-CNF is the most studied generative model (defined below). Indeed, such formulas have been a very popular benchmark for testing and tuning satisfiability algorithms. In particular, some of the most successful ideas in practice today come from insights gained by studying the performance of algorithms on random k-CNF [SLM92, SK93, GSCK00].

Let C_k denote the set of all $(2n)^k$ possible disjunctions of k literals from some canonical set of n Boolean variables. To form a random k-CNF formula, $F_k(n,m)$, on n variables with m clauses we select uniformly, independently and with replacement m clauses from C_k and take their conjunction[1]. We emphasize that throughout the paper k is arbitrarily large but fixed, while n tends to infinity. We will say that a sequence of events $\mathcal{E}_n$ occurs *with high probability* (w.h.p.) if $\lim_{n\to\infty} \Pr[\mathcal{E}_n] = 1$ and *with uniformly positive probability* if $\liminf_{n\to\infty} \Pr[\mathcal{E}_n] > 0$.

For every $k \geq 2$ and $p \in (0,1]$, define

$$r_k(p) \equiv \sup\{r : F_k(n,rn) \text{ is } p\text{-satisfiable w.h.p.}\}$$
$$r_k^*(p) \equiv \inf\{r : F_k(n,rn) \text{ is } \textit{not } p\text{-satisfiable w.h.p.}\}$$

* Part of this work was done while visiting UC Berkeley.
† Research supported by NSF Grant CCR-0121555, NSF Grant DMS-0104073 and a Miller Professorship at UC Berkeley.

1 Our results hold in all common models for random k-CNF, e.g. when clause replacement is not allowed. See the end of Section 2.

One of the most intriguing aspects of random k-CNF formulas is the *Satisfiability Threshold Conjecture*: for every $k \geq 3$, $r_k(1) = r_k^*(1)$. Much work has been done to bound $r_k(1)$ and $r_k^*(1)$. For general k, the best known bounds, from [AP03, DB97] respectively, are:

$$2^k \ln 2 - O(k) < r_k(1) \leq r_k^*(1) < 2^k \ln 2 - O(1) \ .$$

A natural starting point for considering Max k-SAT is the observation that for every k-CNF formula there exists a truth assignment satisfying at least $(1 - 2^{-k})$ of all clauses. Indeed, if a k-CNF formula has m clauses, the average over all 2^n truth assignments of the number of satisfied clauses is precisely $(1 - 2^{-k})m$. With this in mind, let us say that a k-CNF formula is *p-satisfiable*, where $p \in [0, 1]$, if there exists a truth assignment satisfying $1 - 2^{-k} + p2^{-k}$ of all clauses.

For $p = 1$, the bounds above show that $r_k(p)$ and $r_k^*(p)$ are equal, up to first order terms. For $p < 1$, on the other hand, the state of the art for general k was presented in an important recent paper by Coppersmith, Gamarnik, Hajiaghayi, and Sorkin [CGHS03]. Specifically, it was proved that there exists an absolute constant $c > 0$ such that for all k and all $p \in (0, p_0(k)]$,

$$\frac{2^{k+1} \ln 2}{c\, k\, p^2} \leq r_k(p) \leq r_k^*(p) \leq \frac{2^{k+1} \ln 2}{p^2} \ . \qquad (1)$$

The upper bound in (1) was proved via the first moment method, while the lower bound is algorithmic. For small k the two are reasonably close, but their ratio is $\Theta(k)$ and, thus, the ratio tends to infinity as k grows; this naturally raises the question which bound is closer to the truth. Our main result resolves this question by determining the values of $r_k(p)$ and $r_k^*(p)$ with relative error that tends to zero exponentially fast in k. For every $p \in (0, 1)$ denote

$$T_k(p) = \frac{2^k \ln 2}{p + (1 - p) \ln(1 - p)} \, , \qquad (2)$$

with $T_k(1) = 2^k \ln 2$ so that T_k is continuous on $(0, 1]$.

Theorem 1. *There exists a sequence $\delta_k = O(k2^{-k/2})$ such that for all $k \geq 2$ and every $p \in (0, 1]$,*

$$(1 - \delta_k)\, T_k(p) < r_k(p) \leq r_k^*(p) \leq T_k(p) \ . \qquad (3)$$

The upper bound for $r_k^*(p)$ in Theorem 1 can be established readily using the entropic-form Chernoff bound for the Binomial (see Section 4). Our main contribution is the lower bound for $r_k(p)$. Indeed, our proof yields an explicit lower bound for $r_k(p)$ for each $k \geq 2$. For $k = 2$, i.e. Max 2-SAT, the algorithm presented in [CGHS03] dominates our lower bound uniformly, i.e. for every density it satisfies a greater fraction of all clauses. Already for $k \geq 3$, though, our methods yield a better bound, as indicated by the following plots.

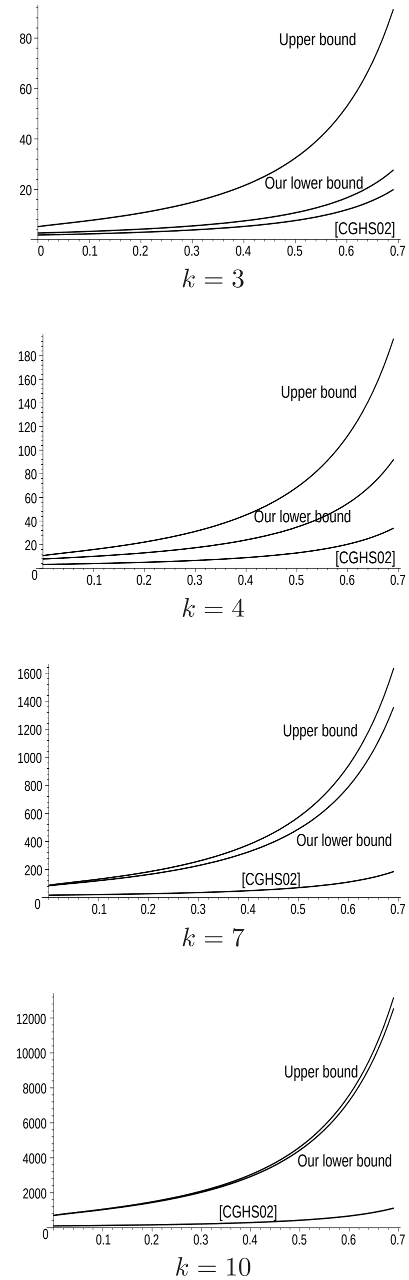

Figure 1. Bounds for $r_k(1-p), r_k^*(1-p)$.

Our proof of Theorem 1 is based on a delicate application of the second moment method to a random generating function in two variables. In general, it is notoriously difficult to obtain precise asymptotics from such random multivariable generating functions; for random Max k-SAT this is possible due to the surprising cancellation of four terms of equal magnitude in our analysis. This cancellation hints at the existence of some unexpected hidden structure in random Max k-SAT; characterizing this structure combinatorially (rather than just analytically) appears to us worthy of further study.

1.1. Background

For a random formula $F_k(n, m)$, denote by $s_k(n, m)$ the random variable equal to the maximum (over all truth assignments σ) of the number of clauses satisfied by σ. Perhaps the first rigorous study of random Max k-SAT appeared in the work of Frieze, Broder and Upfal [BFU93] where it was shown that $s_k(n, m)$ is sharply concentrated around its mean. Specifically,

Theorem 2 ([BFU93]).

$$\Pr\left[\left|s_k(n,m)-\mathbf{E}[s_k(n,m)]\right| > t\right] < 2\exp(-2t^2/m) \ .$$

Corollary 1 below will allow us to "boost" positive probability bounds for $r_k(p)$ to high probability bounds.

Corollary 1. *If $F_k(n, rn)$ is p_0-satisfiable with uniformly positive probability, then for every constant $p < p_0$, $F_k(n, rn)$ is p-satisfiable w.h.p.*

Proof. Let $S \equiv (1 - 2^{-k} + p_0 2^{-k})rn$. Since $F_k(n, rn)$ is p_0-satisfiable with uniformly positive probability we must have $\mathbf{E}[s_k(n, rn)] > S - n^{2/3}$. For otherwise, by Theorem 2, the probability of p_0-satisfiability would be $o(1)$. By the same token, $\Pr[s_k(n, rn) < S - 2n^{2/3}] = o(1)$, yielding the claim. □

Regarding the mean in Theorem 2, since $s_k(n, m) \geq (1-2^{-k})m$ always, it is natural to consider $\Phi_k(n, m) = \mathbf{E}[s_k(n, m)]-(1-2^{-k})m$, measuring the difference between the optimum truth assignment and an average one. In [CGHS03] it was shown that in $F_k(n, rn)$, for all k and sufficiently large r,

$$\liminf_{n\to\infty} \frac{\Phi_k(n,rn)}{n} \geq \frac{2}{k+1}\sqrt{\frac{kr}{\pi 2^k}} - O(1) \quad (4)$$

$$\limsup_{n\to\infty} \frac{\Phi_k(n,rn)}{n} \leq \sqrt{\frac{r(2^k-1)\ln 2}{2^{2k-1}}} \ . \quad (5)$$

This is equivalent to the assertion that for p sufficiently small,

$$r_k(p) \geq \frac{k2^{k+2}}{\pi(k+1)^2} \times p^{-2} - O(p^{-1}) \quad (6)$$

$$r_k^*(p) \leq 2(2^k-1)\ln 2 \times p^{-2} \ , \quad (7)$$

which is a more precise formulation of (1).

Since for $k = 2$, the threshold for satisfiability is known, namely $r_2(1) = r_2^*(1)$=1, in [CGHS03] very fine results were derived for $s_k(n, rn)$ when $r \approx 1$. In particular, when $r = 1 + \varepsilon$ one has $\mathbf{E}[s_2(n, m)] = (1 + \varepsilon - O(\varepsilon^3))n$, while for large $r > 1$ the bounds in (4),(5) can be improved to

$$\liminf_{n\to\infty} \frac{\Phi_2(n,rn)}{n} \geq \frac{\sqrt{r}\ (\sqrt{8}-1)}{3\sqrt{\pi}} - O(1)$$

$$\limsup_{n\to\infty} \frac{\Phi_2(n,rn)}{n} \leq \frac{\sqrt{r}\ \sqrt{3\ln 2}}{\sqrt{8}} \ .$$

Another intriguing aspect of random k-CNF formulas is their proof complexity. In a seminal paper, Chvátal and Szemerédi [CS88] proved that for all $k \geq 3$ and $r > 2^k \ln 2$ there exists $\varepsilon = \varepsilon(r)$ such that w.h.p. every resolution refutation of $F_k(n, rn)$ contains at least $(1 + \varepsilon)^n$ clauses. Since then there have been a number of extensions of this result [BP96, BKPS02] and it is widely believed that random k-CNF are hard for much stronger proof systems than resolution. Indeed, recently, Feige [Fei02] showed that a hypothesis asserting that proving unsatisfiability of random k-CNF with $r \gg 2^k \ln 2$ is hard, implies a number of strong inapproximability results. A closely related hypothesis is that approximating Max k-SAT for such formulas is also hard for all $k \geq 2$. We note that recent work by Fernandez De la Vega and Karpinski [FdlVK02] proves that Max 3-SAT can be approximated on $F_3(n, rn)$ within $9/8$, which is better than the trivial $8/7$ bound.

2. Outline

2.1. Understanding sources of correlation in MAX k-SAT

The following consequence of the Cauchy-Schwarz inequality underlies the second moment method.

Lemma 1. *For any non-negative random variable X,*

$$\Pr[X > 0] \geq \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]} \ . \quad (8)$$

For any fixed $p \in (0, 1]$ one can let X be the number of p-satisfying assignments of $F_k(n, rn)$ and use (1) to bound $\Pr[X > 0]$ from below. Unfortunately, it turns out that for every $r > 0$, there exists a constant $\beta = \beta(k, r) > 1$ such that $\mathbf{E}[X^2] > \beta^n \mathbf{E}[X]^2$. As a result, this straightforward approach only gives an exponentially small bound on the probability of p-satisfiability.

In [AP03], it was shown that when $p = 1$ a major factor contributing to the failure of the above approach is populism: satisfying assignments tend to lean towards the majority vote truth assignment. Roughly, this is because truth assignments that satisfy more literal occurrences than average i) have higher probability of being satisfying, ii) they tend to agree with each other (and the majority truth assignment) on more than half the variables. As a result, the successes of these assignments tend to be highly correlated and they dominate $\mathbf{E}[X^2]$.

To avoid this pitfall we would like to apply the second moment method to truth assignments that satisfy approximately half of all literal occurrences, as in [AP03]. We call such truth assignments *"balanced"*. However, when $p < 1$ focusing on balanced assignments is not enough to obtain a lower bound for $r_k(p)$ that asymptotically matches the upper bound.

To capture the behavior of balanced assignments for Max k-SAT we begin by defining two "fitness" gauges.

Given any k-CNF formula F on n variables and any truth assignment $\sigma \in \{0,1\}^n$ let

1. $H = H(\sigma, F)$ be the number of satisfied literal occurrences in F under σ, minus the number of unsatisfied literal occurrences in F under σ.
2. $U = U(\sigma, F)$ be the number of unsatisfied clauses in F under σ.

We would like to focus on truth assignments that are balanced and p-satisfying, up to fluctuations one would expect from a central limit theorem, i.e., truth assignments σ which for some constant $A > 0$ satisfy

$$|H(\sigma, F)| \le A\sqrt{m} \tag{9}$$

$$|U(\sigma, F) - (1-p)2^{-k}m| \le A\sqrt{m} \ . \tag{10}$$

To do this for a given p_0 let

$$u_0 = (1-p_0)2^{-k}$$

and fix $\gamma, \eta < 1$. For a random k-CNF formula F, consider the weighted sum

$$X = X(\gamma, \eta) = \sum_{\sigma} \gamma^{H(\sigma,F)} \eta^{U(\sigma,F) - u_0 m} \ .$$

Since $\gamma, \eta < 1$ we see that in X the truth assignments σ for which $H(\sigma, F) > 0$ or $U(\sigma, F) > u_0 m$ are suppressed exponentially, whereas the rest are rewarded exponentially. Decreasing $\gamma, \eta \in [0, 1)$ makes this phenomenon more and more acute, with the limiting case $\gamma, \eta = 0$ corresponding to a 0-1 weighting scheme (we adopt the convention $0^0 \equiv 1$).

Indeed, applying the second moment method to X with $\eta = 0$ corresponds to the method of [AP03] for the random k-SAT threshold, where only satisfying assignments receive non-zero weight $\gamma^{H(\sigma,F)}$. Here, we need to tune γ and η *simultaneously* (as p varies) to focus on truth assignments σ for which both (9) and (10) hold. Moreover, we will need to truncate the random variable X in order to reduce its variance.

2.2. Tuning parameters and truncation

When $\eta > 0$, attempting to apply the second moment method to X we encounter two major problems.

The first problem is that while $X > 0$ implies satisfiability when $\eta = 0$, having $X > 0$ when $\eta > 0$ does not imply p_0-satisfiability: in principle, X could be positive due to the contribution of assignments falsifying many more clauses than $u_0 m$. This necessitates restricting the sum defining X to truth assignments falsifying at most $u_0 m + O(\sqrt{m})$ clauses, i.e. truncating X.

The second, more severe, problem is that with or without this truncation, the ratio $\mathbf{E}[X^2]/\mathbf{E}[X]^2$ becomes exponentially large when r is, roughly, half the upper bound of Theorem 1. To get within $1 - o(1)$ of the upper bound we will, rather counterintuitively, remove from the sum defining X those "heroic" truth assignments falsifying *fewer* than $u_0 m$ clauses.

Specifically, for some fixed $A > 0$ let

$$\mathcal{S}^* = \Big\{\sigma \in \{0,1\}^n : H(\sigma, F) \ge 0 \text{ and } U(\sigma, F) \in [u_0 m, u_0 m + A\sqrt{m}]\Big\} \ .$$

Moreover, given u_0, let γ_0, η_0 be defined by

$$\begin{aligned} 1 - \eta_0 &= (1-\gamma_0^2)(1+\gamma_0^2)^{k-1} \\ u_0 &= \frac{\eta_0}{(1+\gamma_0^2)^k - (1-\eta_0)} \ . \end{aligned} \tag{11}$$

These two equations define γ_0, η_0 such that the main contribution to X comes from truth assignments for which (9) and (10) hold. The origin of (11) will become clear after we establish (25) and (26) in Section 5.

Now, let

$$X_* = \sum_{\sigma \in \mathcal{S}^*} \gamma_0^{H(\sigma,F)} \eta_0^{U(\sigma,F) - u_0 m} \ .$$

Note that, by definition, when $X_* > 0$ at least one truth assignment must falsify at most $u_0 m + A\sqrt{m}$ clauses. Thus, if for a given p_0 we can prove that there exists a constant $D > 0$ such that $\mathbf{E}[X_*^2] < D \times \mathbf{E}[X_*]^2$ then, by Corollary 1, it follows that $F_k(n, rn)$ is w.h.p. p-satisfiable for all $p < p_0$.

Bounding the second moment of X_* is decomposed to Lemmata 2–5 below. In all these lemmata $k \ge 2$ is a fixed integer and $r > 0$ is a constant.

For $\alpha \in [0, 1]$, let

$$\begin{aligned} f(\alpha, \gamma, \eta) &= \eta^{-2u_0} \Bigg[\left(\alpha \left(\frac{\gamma^2 + \gamma^{-2}}{2} \right) + 1 - \alpha \right)^k \\ &\quad -2(1-\eta) \left(\frac{\alpha\gamma^{-2} + (1-\alpha)}{2} \right)^k \\ &\quad +(1-\eta)^2 \left(\frac{\alpha\gamma^{-2}}{2} \right)^k \Bigg] \end{aligned} \tag{12}$$

and

$$g_r(\alpha, \gamma, \eta) = \frac{f(\alpha, \gamma, \eta)^r}{\alpha^\alpha (1-\alpha)^{1-\alpha}} \ . \tag{13}$$

Lemma 2. *For every $u_0, \gamma, \eta \in [0,1)$,*

$$\mathbf{E}[X(\gamma,\eta)]^2 = \Big(2\, g_r(1/2,\gamma,\eta)\Big)^n \ .$$

Lemma 3. *For every u_0, there exists $\theta = \theta(k,A) > 0$ such that as $n \to \infty$,*

$$\frac{\mathbf{E}[X_*]}{\mathbf{E}[X(\gamma_0,\eta_0)]} \to \theta \ .$$

Applying Lemma 2 with $\gamma = \gamma_0$ and $\eta = \eta_0$ gives us $\mathbf{E}[X(\gamma_0,\eta_0)]^2$. Lemma 3, now, asserts that for every value of u_0, $\mathbf{E}[X_*]$ is a constant fraction of $\mathbf{E}[X(\gamma_0,\eta_0)]$. Thus, combined with Lemma 2, it gives us $\mathbf{E}[X_*]^2$ up to a constant factor (which is all we need).

Lemma 4. *Let $\gamma(z), \eta(z)$ be arbitrary sequences such that $\gamma(z) \geq \gamma_0$ and $\eta(z) \geq \eta_0$ for every $0 \leq z \leq n$. Then, for every u_0,*

$$\mathbf{E}[X_*^2] \leq 2^n \sum_{z=0}^{n} \binom{n}{z} f(z/n, \gamma(z), \eta(z))^{rn} \ .$$

Lemma 4 expresses $\mathbf{E}[X_*^2]$ as a sum with $n+1$ terms, the z-th term capturing the contribution to $\mathbf{E}[X_*^2]$ of the $2^n\binom{n}{z}$ pairs of truth assignments with overlap z. The contribution of each such pair is then bounded by $f(z/n,\gamma,\eta)^{rn}$ where γ,η are allowed to *depend on* z, subject only to $\gamma \geq \gamma_0$ and $\eta \geq \eta_0$ respectively. In other words, Lemma 4 allows us to adapt γ and η to α, which is crucial for $p < 1$.

Lemma 5. *Let $\chi : [0,1] \to [\gamma_0, 1)$ and $\omega : [0,1] \to [\eta_0, 1)$ be arbitrary functions satisfying $\chi(1/2) = \gamma_0$, $\omega(1/2) = \eta_0$. Let $g_r(\alpha) = g_r(\alpha, \chi(\alpha), \omega(\alpha))$. Assume that there exists a constant $c = c_{\chi,\omega}(k,r) > 0$ such that for every $\alpha \neq 1/2$, $g_r(\alpha) \leq g_r(1/2) - c(\alpha - 1/2)^2$. Then there exists a constant $D = D_{\chi,\omega}(k,r) > 0$ such that for all sufficiently large n*

$$\mathbf{E}[X_*(\gamma_0,\eta_0)^2] < D \times \big[\mathbf{E}X_*(\gamma_0,\eta_0)\big]^2$$

Lemma 5 allows us to bound the sum in Lemma 4 by exploiting the fact that for any "smooth" choice of sequences $\gamma(z), \eta(z)$ in Lemma 4, the sum will be dominated by the contribution of the $\Theta(n^{1/2})$ terms around its maximum term. As a result, if χ, ω express our scheme for adapting γ, η to α, then the maximum of $g_r(\alpha, \chi(\alpha), \omega(\alpha))$ over $\alpha \in (0,1)$, characterizes the sum in Lemma 4 up to a constant factor.

So, all in all, we see that if for a given r and $u_0 = (1-p_0)2^{-k}$ there exist χ, ω such that for all $\alpha \neq 1/2$

$$g_r\Big(1/2, \gamma_0, \eta_0\Big) > g_r\Big(\alpha, \chi(\alpha), \omega(\alpha)\Big) \qquad (14)$$

then $F_k(n, rn)$ is w.h.p. p-satisfiable for all $p < p_0$.

Indeed, to prove Theorem 1 we will show that for every $p_0 \in (0,1]$ and for the stated $r = r(p)$, there exist functions χ, ω for which (14) holds. To simplify the asymptotic analysis, we use the crudest possible such functions, paying the price of this simplicity in the value of k_0 in Proposition 6 below. We note that by choosing more refined (and more cumbersome) functions χ, ω this value can be improved greatly.

Definition 1. *Let*

$$t_k(q) = \frac{2^k \ln 2}{1 - q + q\ln q}\Big(1 - 20k2^{-k\varphi(q)}\Big) \qquad (15)$$

where

$$\varphi(q) = \frac{(1-\sqrt{q})^2}{1 - q + q\ln q} \ .$$

Theorem 1 follows from the following Proposition.

Proposition 6. *For any $p_0 \in (0,1]$ let*

$$G_r(\alpha) = \begin{cases} g_r(\alpha, \gamma_0, \eta_0) & \textit{if } \alpha \in \left[\frac{3\ln k}{k}, 1 - \frac{3\ln k}{k}\right] \\ & \\ g_r(\alpha, \sqrt{\gamma_0}, \sqrt{\eta_0}) & otherwise. \end{cases}$$

For all $k \geq k_0$, if $r \leq t_k(1-p_0)$ then there exists a constant $c = c(k, r, p_0)$ such that for all $\alpha \in [0,1]$, $G_r(\alpha) \leq G_r(1/2) - c(\alpha - 1/2)^2$.

We also note that for any fixed k, one can partition $[0,1]$ to a large number of intervals and numerically find a good value of γ, η for each interval. We discuss this point further in Section 7.

Finally, we note that general large deviations considerations imply that for every k and p, the condition (14) is sharp for our method. That is, no better lower bound can be derived by considering balanced assignments and, in fact, by any argument that classifies assignments according to their number of satisfied literal occurrences in the formula.

In the following sections we prove Lemmata 2–5 and the upper bound of Theorem 1. The proof of Proposition 6 is quite technical and it is not presented here due to space limitations (but see [ANP03]). Before delving into the probabilistic calculations involved in proving Lemmata 2–5 a couple of remarks are in order.

Relationship to other k-CNF models: Recall that the m clauses of $F_k(n,m)$ are chosen independently with replacement among the $(2n)^k$ possibilities. Thus, the m clauses $\{c_i\}_{i=1}^m$ are i.i.d. random variables, each c_i being the conjunction of k i.i.d. random variables $\{\ell_{ij}\}_{j=1}^k$, each ℓ_{ij} being a uniformly random literal. This viewpoint of the formula as a sequence of km i.i.d. random literals will be very handy for our calculations.

Clearly, in this model some clauses might be improper, i.e. they might contain repeated and/or contradictory literals. At the same time, though, observe that the probability that any given clause is improper is smaller than k^2/n and, moreover, the proper clauses are uniformly selected among all such clauses. Therefore w.h.p. the number of improper clauses is $o(n)$ implying that if for a given r, $F_k(n, rn)$ is p-satisfiable w.h.p. then for $m = rn - o(n)$, the same is true in the model where we only select among proper clauses. The issue of selecting clauses without replacement is completely analogous as w.h.p. there are $o(n)$ clauses that contain the same k variables as some other clause.

Notation: In the ensuing probabilistic calculations it will be convenient to write $\sigma \not\models F$ to denote that the truth assignment σ violates the formula F where F can be a single literal, a clause, or an entire CNF formula.

3. The first moment and proof of Lemma 2

By linearity of expectation and since the $m = rn$ clauses $c_1, c_2, \ldots, c_m$ are chosen independently we have

$$\begin{aligned}
\eta^{u_0 m}\,\mathbf{E}[X(\gamma,\eta)] &= \mathbf{E}\left[\sum_\sigma \gamma^{H(\sigma,F)}\eta^{U(\sigma,F)}\right] \\
&= \sum_\sigma \mathbf{E}\left[\prod_{c_i} \gamma^{H(\sigma,c_i)}\eta^{U(\sigma,c_i)}\right] \\
&= \sum_\sigma \prod_{c_i} \mathbf{E}\left[\gamma^{H(\sigma,c_i)}\eta^{U(\sigma,c_i)}\right] \quad .
\end{aligned} \tag{16}$$

Observe now that since the clauses are identically distributed, by symmetry, it suffices to consider the expectation in (16) for a single random clause $c = \ell_1 \vee \cdots \vee \ell_k$ and a fixed truth assignment σ. Moreover, if we write $\gamma^H \eta^U$ as $\gamma^H + \gamma^H(\eta^U - 1)$ we see that the latter expression is non-zero only when $U > 0$, i.e. when c is violated by σ. So, since the literals $\ell_1, \ldots, \ell_k$ are i.i.d.

$$\begin{aligned}
&\mathbf{E}\left[\gamma^{H(\sigma,c)}\eta^{U(\sigma,c)}\right] = \\
&= \mathbf{E}\left[\gamma^{H(\sigma,c)} - \gamma^{H(\sigma,c)}\left(1-\eta^{U(\sigma,c)}\right)\right] \\
&= \mathbf{E}\left[\gamma^{H(\sigma,c)}\right] - \mathbf{E}\left[\gamma^{-k}(1-\eta)\,\mathbf{1}_{\sigma\not\models c}\right] \\
&= \mathbf{E}\left[\prod_{\ell_i}\gamma^{H(\sigma,\ell_i)}\right] - 2^{-k}\gamma^{-k}(1-\eta) \\
&= \prod_{\ell_i}\mathbf{E}\left[\gamma^{H(\sigma,\ell_i)}\right] - 2^{-k}\gamma^{-k}(1-\eta) \\
&= \left(\frac{\gamma+\gamma^{-1}}{2}\right)^k - (2\gamma)^{-k}(1-\eta) \\
&\equiv Z(\gamma,\eta) \quad .
\end{aligned} \tag{17}$$

Thus,

$$\mathbf{E}[X(\gamma,\eta)] = \eta^{-u_0 rn}\,2^n\,Z(\gamma,\eta)^{rn}\,. \tag{18}$$

Observe now that

$$\left(\eta^{-u_0}Z(\gamma,\eta)\right)^2 = f(1/2,\gamma,\eta) \quad .$$

Therefore,

$$\begin{aligned}
\mathbf{E}[X(\gamma,\eta)]^2 &= \left(\eta^{-u_0 rn}\,2^n\,Z(\gamma,\eta)^{rn}\right)^2 \\
&= \left[\left(\eta^{-u_0 r}\,2\,Z(\gamma,\eta)^r\right)^2\right]^n \\
&= \left[4\,f(1/2,\gamma,\eta)^r\right]^n \\
&= \left[2\,g_r(1/2,\gamma,\eta)\right]^n \quad .
\end{aligned}$$

4. The upper bound in Theorem 1

As mentioned earlier, this upper bound can be readily established by using the entropic-form Chernoff bound for the Binomial (see Lemma A.10 in [AS91] or Lemma 3.8 in [DM95]). Nevertheless, we feel that the following self-contained argument is perhaps more informative. Recall the definition of $T_k(\cdot)$ from (2).

Lemma 7. *For all $k \geq 2$ and $p \in (0,1]$, if $q = 1-p$ then*

$$r_k^*(p) \leq \frac{2^k \ln 2}{q\ln q - (2^k - q)\ln\left(\frac{2^k-1}{2^k-q}\right)} \leq T_k(p) \quad . \tag{19}$$

The right hand inequality of (19) follows from the inequality $\ln t \leq t - 1$ applied to $t = \frac{2^k-1}{2^k-q}$, so we just need to verify the left hand inequality. To do that, write $u_0 = 2^k q$. Let $\eta \in (0,1)$, and observe that if F is p-satisfiable, then $U(\sigma,F) \leq u_0 m$ for some σ, whence

$$X(1,\eta) = \sum_\sigma \eta^{U(\sigma,F)-u_0 m} \geq 1\,.$$

Using Markov's inequality and (17),(18) from Section 3 we have

$$\begin{aligned}
\mathbf{P}[X(1,\eta) \geq 1] &\leq \mathbf{E}[X(1,\eta)] \\
&= \left(2\eta^{-qr2^{-k}}\left(1-(1-\eta)2^{-k}\right)\right)^n \quad .
\end{aligned} \tag{20}$$

Thus, if the n-th root of the RHS of (20) is less than 1, the probability of p-satisfiability decays exponentially in n. Taking $\eta = q(2^k-1)/(2^k-q)$ yields the lemma.

5. Proof of Lemma 3

The proof is based on Cramer's classical "change of measure" technique in large deviation theory.

By linearity of expectation, it suffices to prove that there exists some $\theta = \theta(k, A) > 0$ such that for the values of γ_0, η_0 satisfying (11) and every truth assignment σ, we have

$$\frac{\mathbf{E}\left[\gamma_0^{H(\sigma,F)}\eta_0^{U(\sigma,F)}\mathbf{1}_{\sigma\in\mathcal{S}^*(\mathcal{F})}\right]}{\mathbf{E}\left[\gamma_0^{H(\sigma,F)}\eta_0^{U(\sigma,F)}\right]} \to \theta \ . \tag{21}$$

Recalling that formulas in our model are sequences of i.i.d. random literals $\ell_1, \ldots, \ell_{km}$, let $\mathbf{P}(\cdot)$ denote the probability assigned by our distribution to any such sequence, i.e. $(2n)^{-km}$. Now, fix any truth assignment σ and consider an auxiliary distribution $\mathbf{P}_\sigma$ on k-CNF formulas where the m clauses $c_1, \ldots, c_m$ are again i.i.d. among all $(2n)^k$ clauses, but where now for any fixed clause ν

$$\mathbf{P}_\sigma(c_i = \nu) = \frac{\gamma_0^{H(\sigma,\nu)}\eta_0^{U(\sigma,\nu)}\mathbf{P}(\nu)}{Z(\gamma_0, \eta_0)} \ , \tag{22}$$

where

$$Z(\gamma_0, \eta_0) = \mathbf{E}\left[\gamma_0^{H(\sigma,c)}\eta_0^{U(\sigma,c)}\right] \ , \tag{23}$$

was defined in (17). (Since each fixed clause ν receives probability proportional to $\gamma_0^{H(\sigma,\nu)}\eta_0^{U(\sigma,\nu)}$, indeed $Z(\gamma_0, \eta_0)$ provides the correct normalization to a probability distribution.) So, whereas under $\mathbf{P}(\cdot)$ every k-CNF formula F with m clauses had the same probability $\mathbf{P}(F) = (2n)^{-km}$, under $\mathbf{P}_\sigma$ its probability is

$$\mathbf{P}_\sigma(F) = \frac{\gamma_0^{H(\sigma,F)}\eta_0^{U(\sigma,F)}\mathbf{P}(F)}{Z(\gamma_0, \eta_0)^m} \ . \tag{24}$$

Let $\mathbf{E}_\sigma$ be the expectation operator corresponding to $\mathbf{P}_\sigma$. A calculation similar to that leading to (17), adding the equal contributions from the k literals, gives that for a single random clause c

$$\begin{aligned} Z(\gamma_0, \eta_0)\,\mathbf{E}_\sigma[H(\sigma, c)] &= \\ &= k(\gamma_0 - \gamma_0^{-1})\left(\frac{\gamma_0 + \gamma_0^{-1}}{2}\right)^{k-1} \\ &\quad + k(2\gamma_0)^{-k}(1 - \eta_0) \ . \end{aligned} \tag{25}$$

Moreover,

$$Z(\gamma_0, \eta_0)\mathbf{E}_\sigma[U(\sigma, c)] = (2\gamma_0)^k \eta_0 \, . \tag{26}$$

Thus (11) ensures that $\mathbf{E}_\sigma[H(\sigma, c)] = 0$ and also that $\mathbf{E}_\sigma[U(\sigma, c) - u_0] = 0$.

Next, we apply the multivariate central limit theorem (see, e.g. [Pol02], page 182) to the i.i.d. mean-zero random vectors $\Big(H(\sigma, c_i), U(\sigma, c_i) - u_0\Big)$ for $i = 1, \ldots, m$. Observe that, since $k \geq 2$, the common law of these random vectors is not supported on a line. We deduce that as $n \to \infty$

$$\mathbf{P}_\sigma[\sigma \in \mathcal{S}^*(F)] = \mathbf{E}_\sigma[\mathbf{1}_{\sigma\in\mathcal{S}^*(\mathcal{F})}] \to \theta(k, A) > 0. \tag{27}$$

Here, $\theta(k, A)$ is the probability that a certain nondegenerate bivariate normal law assigns to a certain open set. Its exact value is unimportant for our purpose. Combining (24) and (27) yields (21).

6. Proof of Lemma 4

Linearity of expectation implies

$$\begin{aligned} &\eta_0^{2u_0 m}\,\mathbf{E}[X_*^2] = \\ &\quad = \mathbf{E}\left[\left(\sum_\sigma \gamma_0^{H(\sigma,F)}\eta_0^{U(\sigma,F)}\,\mathbf{1}_{\sigma\in\mathcal{S}^*(F)}\right)^2\right] \\ &\quad = \sum_{\sigma,\tau}\mathbf{E}\Big[\gamma_0^{H(\sigma,F)+H(\tau,F)}\eta_0^{U(\sigma,F)+U(\tau,F)} \\ &\qquad\qquad \times \mathbf{1}_{\sigma,\tau\in\mathcal{S}^*(F)}\Big] \ . \end{aligned} \tag{28}$$

Observe now that since $\sigma \in \mathcal{S}^*$ implies $H(\sigma, F) \geq 0$ and $U(\sigma, F) \geq u_0 m$, we get that for every pair σ, τ and any $\gamma \geq \gamma_0$ and $\eta \geq \eta_0$,

$$\begin{aligned} &\mathbf{E}\left[\gamma_0^{H(\sigma,F)+H(\tau,F)}\eta_0^{U(\sigma,F)+U(\tau,F)}\,\mathbf{1}_{\sigma,\tau\in\mathcal{S}^*(F)}\right] \\ &\quad \leq \mathbf{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)}\,\mathbf{1}_{\sigma,\tau\in\mathcal{S}^*(F)}\right] \\ &\quad \leq \mathbf{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)}\right] \ . \end{aligned} \tag{29}$$

In other words, when using the right hand side of (29) to bound each term of the sum in (28), we are allowed to *adapt* the value of γ and η to the pair σ, τ, the only restrictions being $\gamma \geq \gamma_0$ and $\eta \geq \eta_0$. This is a crucial point and we will exploit it heavily when bounding the contribution of pairs with large overlap.

To estimate the right hand side of (29) for any fixed pair of truth assignments σ, τ we first observe that the m clauses $c_1, c_2, \ldots, c_m$ are i.i.d.. Therefore, if c is a random clause then

$$\begin{aligned} &\mathbf{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)}\right] = \\ &\quad = \mathbf{E}\left[\prod_{c_i}\gamma^{H(\sigma,c_i)+H(\tau,c_i)}\eta^{U(\sigma,c_i)+U(\tau,c_i)}\right] \\ &\quad = \prod_{c_i}\mathbf{E}\left[\gamma^{H(\sigma,c_i)+H(\tau,c_i)}\eta^{U(\sigma,c_i)+U(\tau,c_i)}\right] \\ &\quad = \left(\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\eta^{U(\sigma,c)+U(\tau,c)}\right]\right)^m . \end{aligned} \tag{30}$$

Next, we observe that for every pair σ, τ, by symmetry, the expectation in (30) depends only on the number of variables to which σ, τ assign the same value.

Specifically, let σ, τ be any pair of truth assignments that agree on exactly $z = \alpha n$ variables, i.e. have overlap z. By first rewriting $\gamma^H \eta^U$ as $\gamma^H + \gamma^H(\eta^U - 1)$ and then observing that $\eta^{U(\tau,c)}$ is distributed identically to $\eta^{U(\sigma,c)}$ we get

$$\begin{aligned}
&\mathbf{E}\left[\gamma^{H(\sigma,c)}\eta^{H(\sigma,c)}\gamma^{H(\tau,c)}\eta^{H(\tau,c)}\right] = \\
&= \mathbf{E}\left[\left(\gamma^{H(\sigma,c)} - \gamma^{H(\sigma,c)}\left(1-\eta^{U(\sigma,c)}\right)\right)\right. \\
&\qquad \left.\cdot\left(\gamma^{H(\tau,c)} - \gamma^{H(\tau,c)}\left(1-\eta^{U(\tau,c)}\right)\right)\right] \\
&= \mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] \\
&\quad -2\,\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\left(1-\eta^{U(\sigma,c)}\right)\right] \\
&\quad +\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\left(1-\eta^{U(\sigma,c)}\right)\left(1-\eta^{U(\tau,c)}\right)\right] \\
&= \mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] \\
&\quad -2(1-\eta)\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\mathbf{1}_{\sigma\not\models c}\right] \\
&\quad +2^{-k}\alpha^k\gamma^{-2k}(1-\eta)^2 \;. \qquad (31)
\end{aligned}$$

To estimate (31) we note that the literals $\ell_1, \ell_2, \ldots \ell_k$ comprising c are i.i.d. Therefore,

$$\begin{aligned}
\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] &= \mathbf{E}\left[\prod_i \gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\right] \\
&= \prod_i \mathbf{E}\left[\gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\right] \\
&= \left(\alpha\left(\frac{\gamma^2+\gamma^{-2}}{2}\right)+1-\alpha\right)^k \;.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbf{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\mathbf{1}_{\sigma\not\models c}\right] = \\
&\quad = \mathbf{E}\left[\prod_i \gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\mathbf{1}_{\sigma\not\models \ell_i}\right] \\
&\quad = \prod_i \mathbf{E}\left[\gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\mathbf{1}_{\sigma\not\models \ell_i}\right] \\
&\quad = \left(\frac{\alpha\gamma^{-2}+(1-\alpha)}{2}\right)^k \;.
\end{aligned}$$

Substituting these last two equations in (31) we get

$$\begin{aligned}
&\eta^{-2u_0}\,\mathbf{E}\left[\gamma^{H(\sigma,c)}\eta^{H(\sigma,c)}\gamma^{H(\tau,c)}\eta^{H(\tau,c)}\right] = \\
&= \eta^{-2u_0}\left[\left(\alpha\left(\frac{\gamma^2+\gamma^{-2}}{2}\right)+1-\alpha\right)^k\right. \\
&\quad -2(1-\eta)\left(\frac{\alpha\gamma^{-2}+(1-\alpha)}{2}\right)^k \\
&\quad \left.+(1-\eta)^2\left(\frac{\alpha\gamma^{-2}}{2}\right)^k\right] = f(\alpha,\gamma,\eta) \;. \qquad (32)
\end{aligned}$$

Since the number of ordered pairs with overlap z is $2^n\binom{n}{z}$ we thus conclude that

$$\mathbf{E}[X_*^2] \le 2^n \sum_{z=0}^{n} \binom{n}{z} f(z/n, \gamma(z), \eta(z))^m \;, \qquad (33)$$

for any set of choices for $\gamma(z), \eta(z)$ such that $\gamma(z) \ge \gamma_0$ and $\eta(z) \ge \eta_0$ for all $0 \le z \le n$.

6.1. Proof of Lemma 5

Lemma 5 is a consequence of the following Laplace-type lemma from [AM02]. The idea is that such sums are dominated by the contribution of $\Theta(n^{1/2})$ terms around the maximum term.

Lemma 8. *Let ϕ be any positive, function on $[0,1]$ and let*

$$S_n = \sum_{z=0}^{n} \binom{n}{z} \phi\left(\frac{z}{n}\right)^n \;.$$

Letting $0^0 \equiv 1$, define g on $[0,1]$ as

$$g(\alpha) = \frac{\phi(\alpha)}{\alpha^\alpha(1-\alpha)^{1-\alpha}} \;.$$

Assume that there exists $c > 0$ such that for all α, $g(\alpha) \le g(1/2) - c(\alpha - 1/2)^2$. Then there exists a constant $D = D(c) > 0$ such that $S_n < D \times g(1/2)^n$ for all sufficiently large n.

7. Bounds for finite k

As mentioned in Section 2, for any fixed k the simple adaptation scheme of Proposition 6 does not yield the best possible lower bound for $r_k(p)$ afforded by our method. For that, one has to use an optimal adaptation of γ, η with respect to α. Our lower bounds reported in Figure 1 are the result of finding a good adaptation of γ, η to α numerically. Specifically, for each k we computed a lower bound for 100 equally spaced values of p and then had Maple [Red94] "connect the dots". (For the upper bound plots and the plots of the lower bound from [CGHS03] we used the explicit formulas.)

For each of these values of p, to prove the corresponding lower bound for $r_k(p)$ we established that there exist a choice of functions χ, ω as in Lemma 5 such that for all $\alpha \in (1/2, 1]$ we have $g_r(1/2, \gamma_0, \eta_0) > g_r(\alpha, \chi(\alpha), \omega(\alpha))$. To that end, we partitioned $(1/2, 1]$ to 10,000 points and for each such point we searched for values of $\gamma \geq \gamma_0$ and $\eta \geq \eta_0$ such that this condition holds with a bit of room. (For $k > 4$ we solved (11), defining γ_0 and η_0, numerically to 10 digits of accuracy. For the optimization we exploited convexity to speed up the search.) Having determined such values, we (implicitly) extended the functions χ, ω to all $(1/2, 1]$ by assigning to every not-chosen point the value at the nearest chosen point. Finally, we computed a (crude) upper bound on the derivative of g_r with respect to α in $(1/2, 1]$. This bound on the derivative, along with our room factor, then implied that for every point that we did not check, the value of g_r was sufficiently close to its value at the corresponding chosen point to also be dominated by $g_r(1/2, \gamma_0, \eta_0)$.

Acknowledgements

We thank Cris Moore for helpful conversations in the early stages of this work.

References

[AM02] Dimitris Achlioptas and Cristopher Moore, *The asymptotic order of the random k-SAT threshold*, 43th Annual Symposium on Foundations of Computer Science, 2002, pp. 779–788.

[ANP03] Dimitris Achlioptas, Assaf Naor and Yuval Peres, *On the maximum satisfiability of random formulas*, available at: http://www.arxiv.org/abs/math.PR/0305151

[AP03] Dimitris Achlioptas and Yuval Peres, *The random k-SAT threshold is* $2^k \ln 2 - O(k)$, 35th Annual Symposium on Theory of Computing, 2003, pp. 223–231.

[AS91] Noga Alon and Joel H. Spencer, *The Probabilistic Method*, Wiley 1991.

[BFU93] Andrei Z. Broder, Alan M. Frieze, and Eli Upfal, *On the satisfiability and maximum satisfiability of random* 3*-CNF formulas*, 4th Annual Symposium on Discrete Algorithms, 1993, pp. 322–330.

[BKPS02] Paul Beame, Richard Karp, Toniann Pitassi, and Michael Saks, *The efficiency of resolution and davis-putnam procedures*, SIAM J. Comput. **31** (2002), no. 4, 1048–1075.

[BP96] Paul W. Beame and Toniann Pitassi, *Simplified and improved resolution lower bounds*, 37th Annual Symposium on Foundations of Computer Science, 1996, pp. 274–282.

[CGHS03] Don Coppersmith, David Gamarnik, Mohammad T. Hajiaghayi, and Gregory B. Sorkin, *Random MAX 2-SAT and MAX CUT*, 14th Annual Symposium on Discrete Algorithms, 2003, pp. 364-373.

[Coo71] Stephen A. Cook, *The complexity of theorem-proving procedures*, 3rd Annual Symposium on Theory of Computing, 1971, pp. 151–158.

[CS88] Vašek Chvátal and Endre Szemerédi, *Many hard examples for resolution*, J. Assoc. Comput. Mach. **35** (1988), no. 4, 759–768.

[DM95] Paul Deheuvels and David M. Mason *On the Fractal Nature of Empirical Increments*, The Annals of Probability **23** (1995), 355–387.

[DB97] Olivier Dubois and Yacine Boufkhad, *A general upper bound for the satisfiability threshold of random r-SAT formulae*, J. Algorithms **24** (1997), no. 2, 395–420.

[FdlVK02] W. Fernandez de la Vega and Marek Karpinski, *9/8-approximation algorithm for random max-3sat*, Technical Report TR02-070, Electonic Colloquium on Computational Complexity (2002).

[Fei02] Uriel Feige, *Relations between average case complexity and approximation complexity*, 34th Annual Symposium on Theory of Computing, 2002, pp. 534 – 543.

[GJ79] Michael R. Garey and David S. Johnson, *Computers and intractability*, Freeman, San Francisco, CA, 1979.

[GSCK00] C. P. Gomes, B. Selman, N. Crato, and H. Kautz, *Heavy-tailed phenomena in satisfiability and constraint satisfaction problems*, J. Automat. Reason. **24** (2000), no. 1-2, 67–100. MR 2000k:68070

[Pol02] David Pollard, *A user's guide to measure theoretic probability*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2002. MR 2002k:60003

[Red94] Darren Redfern, *The Maple Handbook: Maple V Release 3*, third ed., Springer Verlag, New York, 1994.

[SK93] Bart Selman and Henry Kautz, *Domain-independent extensions to GSAT: Solving large structured satisfiability problems*, Proc. 13th International Joint Conference on Artificial Intelligence, 1993, pp. 290–295.

[SLM92] Bart Selman, Hector Levesque, and D. Mitchell, *A new method for solving hard satisfiability problems*, Proc. 10th National Conference on Artificial Intelligence, 1992, pp. 440–446.