The Fraction of Satisfiable Clauses in a Typical Formula

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

Given n Boolean variables $x_1, \ldots, x_n$, a k-clause is a disjunction of k literals, where a literal is a variable or its negation. A k-CNF formula is a conjunction of a finite number of k-clauses. Call such a formula *p-satisfiable* if there exists a truth assignment satisfying a fraction $1-2^{-k}+p\,2^{-k}$ of all clauses and observe that every k-CNF formula is 0-satisfiable. Our goal is to determine for which ratios of clauses to variables, a typical k-CNF formula is p-satisfiable. Let $F_k(n,m)$ denote a formula on n variables formed by selecting uniformly and independently m out of all $(2n)^k$ possible k-clauses on $x_1, \ldots, x_n$. Given $p \in (0,1]$, let $r_k(p)$ be the supremum of all r such that the probability that $F_k(n, rn)$ is p-satisfiable tends to 1 as $n \to \infty$. Similarly, let $r_k^*(p)$ be the infimum of all r such that the probability that $F_k(n, rn)$ is p-satisfiable tends to 0 as $n \to \infty$. It is widely believed that $r_k(p) = r_k^*(p)$ for every $p \in (0,1]$ and $k \geq 2$. We prove that this is true asymptotically in k. More precisely, we prove that for every $p \in (0,1]$, both $r_k(p)$ and $r_k^*(p)$, when divided by 2^k, tend exponentially fast to $\log 2/[p + (1-p)\log(1-p)]$ as $k \to \infty$. As a corollary, for every $k \geq 2$ and $r > 2^k \log 2$ we give $p = p(k,r)$ such that with probability tending to 1 as $n \to \infty$, $F_k(n, rn)$ is p-satisfiable but not $(p+\delta)$-satisfiable, where $\delta = \delta(k) = O(k2^{-k/2})$. The proof is based on a delicate asymptotic analysis of a random bivariate generating function.

1 Introduction

This work is devoted to the asymptotic resolution of the maximum satisfiability threshold problem, a problem which originated in theoretical computer science in the early 1970s and has since been studied by mathematicians and statistical physicists due to its connections to spin glasses. Our solution is based on delicate analytic and probabilistic techniques, whereas previous rigorous results on the problem were algorithmic.

Definition 1. *Given n Boolean variables $x_1, \ldots, x_n$, a clause is a disjunction of literals, where a literal is a variable or its negation. A CNF formula F (CNF stands for conjunctive normal form) is the conjunction of finitely many clauses. When all clauses have k literals F is called a k-CNF formula.*

Given a CNF formula F, the Satisfiability problem asks whether there exists a truth assignment under which F evaluates to true. In his seminal 1971 paper [7], Cook proved that Satisfiability is complete for the complexity class NP, i.e., that it is at least as hard as any decision problem for which positive answers can be verified in polynomial time. Cook also proved that Satisfiability remains NP-complete when the input is restricted to be a k-CNF formula, where $k \geq 3$, and this version of the problem (known as k-SAT) has since been the canonical problem of computational complexity.

*Part of this work was done while visiting UC Berkeley.

†Research supported by NSF Grant DMS-0104073 and a Miller Professorship at UC Berkeley.

Given a k-CNF formula F, a natural generalization of k-SAT is to ask whether there exists a truth assignment satisfying at least a certain number of clauses of F (rather than all of them). This problem is known as Max k-SAT and it is NP-complete for $k \geq 2$ (see [13]). Observe that if a k-CNF formula has m clauses, the average over all 2^n truth assignments of the number of satisfied clauses is precisely $(1 - 2^{-k})m$. In particular, one can always produce a truth assignment satisfying all but $2^{-k}m$ clauses, e.g. by iteratively satisfying any literal that appears in the largest number of yet unsatisfied clauses. We will say that F is *p-satisfiable*, where $p \in [0, 1]$, if there exists a truth assignment satisfying $1 - 2^{-k} + p\,2^{-k}$ of all clauses.

A recent deep theorem of Håstad [14] states that, in general, the above algorithm is best possible: any polynomial-time algorithm that can distinguish between ε-satisfiable formulas and 1-satisfiable formulas, can be used to solve the original satisfiability problem in polynomial time. This suggests that in the *worst case* (over the choice of formulas), even approximating the maximum fraction of satisfiable clauses in a k-CNF formula is an extremely hard problem.

In this paper we study the maximum fraction of satisfiable clauses in *typical* k-CNF formulas, i.e., formulas chosen uniformly at random among all k-CNF formulas with a given number of variables and clauses. Our main result establishes that for such formulas the maximum fraction of satisfiable clauses can be determined with very high accuracy. The study of optimization problems with random inputs is pervasive in operations research (e.g., the travelling salesman problem and variants – see the recent special issue of TCS [11] and the references therein), in statistical physics (determining ground states of spin glasses – see for example Talagrand's recent book [19]) and in computer science.

Let $F_k(n, m)$ denote a random formula on n variables formed by selecting uniformly and independently m out of all $(2n)^k$ possible k-clauses on $x_1, \ldots, x_n$. When dealing with random k-CNF formulas[1] we say that a sequence of random events $\mathcal{E}_n$ occurs *with high probability* (w.h.p.) if $\lim_{n\to\infty} \mathbf{P}[\mathcal{E}_n] = 1$ and *with uniformly positive probability* if $\liminf_{n\to\infty} \mathbf{P}[\mathcal{E}_n] > 0$. We emphasize that throughout the paper k is arbitrarily large but fixed, while $n \to \infty$. We introduce the following parameters for every $k \geq 2$ and $p \in (0, 1]$:

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

One of the most intriguing aspects of random formulas is the *Satisfiability Threshold Conjecture* asserting that $r_k(1) = r_k^*(1)$ for every $k \geq 3$. Much work has been done to bound $r_k(1)$ and $r_k^*(1)$. Currently, the best rigorous bounds for general $k \geq 3$, from [2, 10] respectively, are: $2^k \log 2 - O(k) < r_k \leq r_k^* < 2^k \log 2 - O(1)$. For all $p < 1$, though, the ratio of the previously known upper bound for $r_k^*(p)$ and the lower bound for $r_k(p)$ tended to infinity with k.

The state of the art for general k was presented in an important recent paper by Coppersmith, Gamarnik, Hajiaghayi, and Sorkin [6], where it was proved that there exists an absolute constant $c > 0$ such that for all k and all $p \in (0, p_0(k)]$,

$$\frac{2^{k+1}\log 2}{ckp^2} < r_k(p) \leq r_k^*(p) < \frac{2^{k+1}\log 2}{p^2} \ . \tag{1}$$

For small k the two bounds in (1) are reasonably close, but the ratio between them grows linearly in k; this naturally raises the question which bound is closer to the truth. Our main result resolves this question by pinpointing 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

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

and let $T_k(1) = 2^k \log 2$ so that $T_k(\cdot)$ 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 $p \in (0, 1]$,*

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

Theorem 1 readily implies the following.

[1] Our results hold in all common models for random k-CNF, e.g. when literal replacement is not allowed; see Section 3.

Corollary 1. *For $k \geq 2$ and $r > 2^k \log 2$, let $p_c(k, r)$ be the supremum over all p such that w.h.p. $F_k(n, rn)$ is p-satisfiable. Let Ψ be the inverse of the function $f(p) = p + (1-p)\log(1-p)$. Then*

$$\Psi\left(\frac{2^k \log 2}{r}\right)(1 - \delta_k) \leq p_c(k, r) \leq \Psi\left(\frac{2^k \log 2}{r}\right),$$

where $\delta_k = O(k 2^{-k/2})$ is the same as in Theorem 1.

Proof. Denote $s = \Psi\left(\frac{2^k \log 2}{r}\right)$. We have $r = T_k(s) > r_k^*(s)$ and, therefore, w.h.p. $F_k(n, rn)$ is not s-satisfiable. If $p \leq s(1 - \delta_k)$, then $\frac{2^k \log 2}{r} = f(s) \geq f(p/(1-\delta_k)) > f(p)/(1-\delta_k)$, where the second inequality follows from the convexity of f. Thus, $r < T_k(p)(1-\delta_k) < r_k(p)$, implying that w.h.p. $F_k(n, rn)$ is p-satisfiable. $\square$

The upper bound in (3) follows from well-known tail estimates. Taylor expansion gives that as $p \to 0$,

$$T_k(p) = \frac{2^k \log 2}{p^2/2 + O(p^3)} \ ,$$

so for $p \to 0$ we can sharpen (1) to

$$(1 - \delta_k)\frac{2^{k+1} \log 2}{p^2 + O(p^3)} \ \leq \ r_k(p) \ \leq \ r_k^*(p) \ \leq \ \frac{2^{k+1} \log 2}{p^2 + O(p^3)} \ . \qquad (4)$$

Our proof of Theorem 1 actually 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 [6] dominates our lower bound uniformly, i.e., for every value of p, it yields a better lower bound for $r_2(p)$. Already for $k \geq 3$, though, our methods yield a better bound. The following plots indicate that even for relatively small k, our bounds are quite tight:

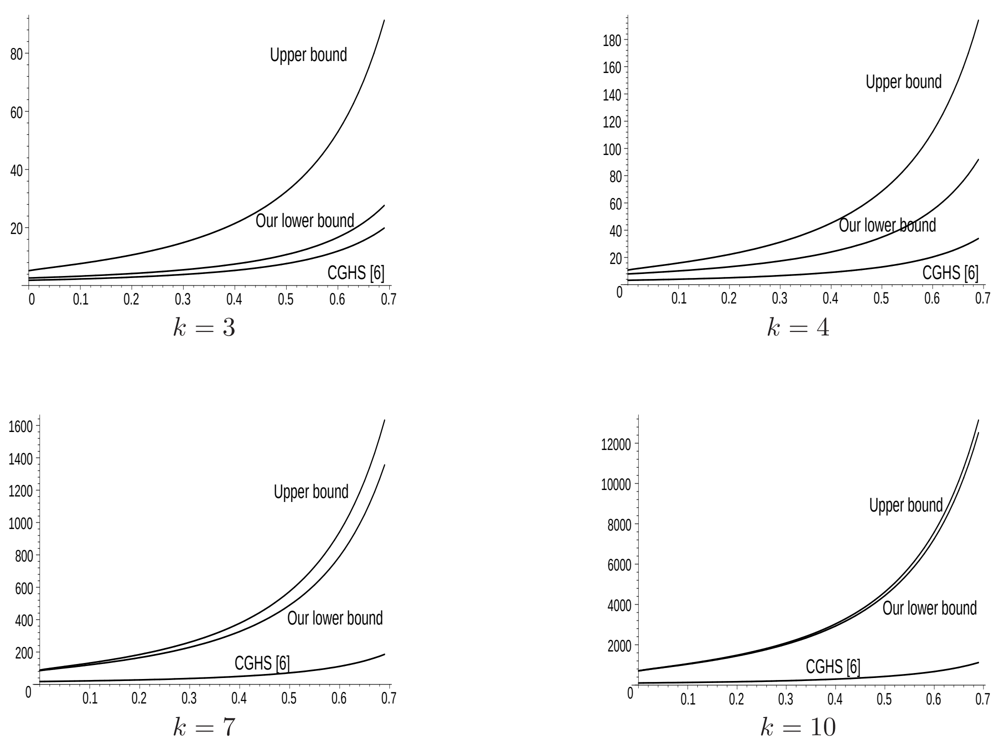

Figure 1. Bounds for r_k^* and r_k as functions of $q = 1 - p$.

The proof of Theorem 1 is based on a delicate application of the second moment method. The second moment method is a natural approach to many probabilistic and combinatorial problems, and we refer to Janson's survey article [16] for a presentation of this method and a useful conditional variant. For certain problems, a direct application of the second moment method fails due to high correlations, yet once the source of correlations is recognized, a suitable truncation or weighting can control it. Multiscale truncation was used in the solution of the Erdős-Taylor (1960) conjecture on simple random walk in the planar square lattice (see [12], [9]) and a weighting scheme motivated by entropy maximization was the key to the work of the first and third authors on random k-SAT [2].

In the present paper, in order to deal with correlations we study a random bivariate generating function that weights truth assignments according to the fraction of satisfied clauses and the fraction of satisfied literal occurrences. This approach builds upon insights from both [9] and [2], yet obtaining tight bounds in the presence of an additional parameter (the fraction of satisfied clauses) presents new analytic challenges. The crucial new ingredient is a truncation that allows us to *adapt* the weight assigned to each pair of assignments to their overlap. We emphasize that for Max k-SAT this adaptation is necessary even for determining the first order asymptotics (see Section 2.2). In general, it is notoriously difficult to obtain precise asymptotics from such random multivariate 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.

2 Locating the p-satisfiability threshold via moment estimates

The upper bound in Theorem 1 follows readily by applying Markov's inequality to the number of p-satisfying truth assignments (see Section 3.1). While for any fixed value of k one can get a slightly better upper bound for $r_k^*(p)$ by employing a more refined counting argument, such as that of Janson, Stamatiou and Vamvakari [15], it is remarkable that the naive first moment bound is asymptotically tight.

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. The first rigorous study of random Max k-SAT appeared in the work of Broder, Frieze and Upfal [4] where it was shown that $s_k(n,m)$ is sharply concentrated around its mean. Specifically,

Theorem 2 ([4]). $\mathbf{P}\left[\left|s_k(n,m) - \mathbb{E}[s_k(n,m)]\right| > t\right] < 2\exp\left(-\dfrac{2t^2}{m}\right)$.

In proving the lower bound of Theorem 1, we will use the following corollary of Theorem 2:

Corollary 2. *Assume that there exists $c = c(k,p,r)$ such that for n large enough $F_k(n,rn)$ is p-satisfiable with probability greater that n^{-c}. Then $F_k(n,rn)$ is p'-satisfiable w.h.p. for every constant $p' < p$.*

Proof. Let $S \equiv (1 - 2^{-k} + p2^{-k})rn$. Then $\mathbb{E}[s_k(n,rn)] > S - n^{2/3}$, since otherwise Theorem 2 would imply that the probability of p-satisfiability is less than $2e^{-2r^{-1}n^{1/3}}$, contradicting our assumption. By the same token, $\mathbf{P}[s_k(n,rn) < S - 2n^{2/3}] = o(1)$. □

Thus to establish the lower bound in (3) we will find for every $p \in (0,1]$ a value $r = r(p)$ such that $F_k(n,rn)$ is p-satisfiable with probability $\Omega(1/n)$ and rely on Corollary 2 to get a corresponding high probability result. A natural way of bounding probabilities from below is the second moment method, which is based on the following easy consequence of the Cauchy-Schwartz inequality:

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

$$\mathbf{P}[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \ . \tag{5}$$

2.1 Why the standard second moment method fails

A natural way to apply Corollary 2 and Lemma 1 is the following: for any fixed $p \in (0,1]$ one can let X denote the number of p-satisfying assignments and apply (5) to bound $\mathbf{P}[X > 0]$ from below. Unfortunately, it turns out that for all $k, p, r > 0$ there exists a constant $\beta = \beta(k,p,r) > 0$ such that $\mathbb{E}[X^2] > (1+\beta)^n \mathbb{E}[X]^2$. As a result, this straightforward approach only gives a trivial lower bound on the probability of p-satisfiability. As shown in [2] for the case $p = 1$, a major factor in the excessive correlations behind the above failure is that p-satisfying truth assignments tend to lean toward the majority vote truth assignment. To see this, first observe that truth assignments that satisfy more literal occurrences than the average number $km/2$, have higher probability of being p-satisfying. Yet, in order to satisfy many literal occurrences, such assignments tend to agree with each other (and the majority truth assignment) on more than half the variables. As a result, the successes of such assignments tend to be highly correlated and dominate $\mathbb{E}[X^2]$.

2.2 A bivariate weighting scheme

An attractive feature of the second moment method is that we are free to apply it to any random variable X such that $X > 0$ implies p-satisfiability. Hence, in order to avoid the above pitfall, we would like to apply the second moment method to truth assignments that satisfy, approximately, half of all literal occurrences; we call such truth assignments *"balanced"*.

In what follows we fix $p \in (0,1]$ and $k \geq 2$. We will denote by $F = F_k(n,m)$ a random k-CNF formula on n variables with $m = rn$ clauses. For 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 the central limit theorem, i.e., truth assignments σ such that for some constant $A > 0$,

$$|H(\sigma,F)| \leq A\sqrt{m} \tag{6}$$

$$|U(\sigma,F) - (1-p)2^{-k}m| \leq A\sqrt{m}\ . \tag{7}$$

To do this let us fix $0 < \gamma, \eta < 1$ and define $X(\gamma,\eta)$ as

$$X(\gamma,\eta) = \sum_{\sigma} \gamma^{H(\sigma,F)} \eta^{U(\sigma,F)-u_0 m}\ , \tag{8}$$

where, and this is a hard-coded definition with inconsistent but convenient notation.

$$u_0 = \frac{1-p}{2^k}\ . \tag{9}$$

Since $\gamma, \eta < 1$ we see that in $X(\gamma,\eta)$ the truth assignments σ for which $H(\sigma,F) > 0$ or $U(\sigma,F) > u_0 m$ are suppressed exponentially, while 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 $X(\gamma,\eta)$ with $\eta = 0$ corresponds to the approach of [2] for the random k-SAT threshold, where only satisfying assignments receive non-zero weight $\gamma^{H(\sigma,F)}$. Unfortunately, when attempting to apply the second moment method to $X(\gamma,\eta)$ with $\eta > 0$ we immediately encounter two problems.

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

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

Correspondingly, we define a truncated version of $X(\gamma,\eta)$ as

$$X_*(\gamma,\eta) = \sum_{\sigma\in\mathcal{S}^*} \gamma^{H(\sigma,F)}\eta^{U(\sigma,F)-u_0 m} \quad . \tag{10}$$

By definition, when $X_*(\gamma,\eta) > 0$ at least one truth assignment must falsify at most $u_0 m + A\sqrt{m}$ clauses. Thus, if we prove that there exists a constant $D > 0$ (which may depend on k, p but not on n) such that

$$\mathbb{E}X_*(\gamma,\eta)^2 < Dn \cdot [\mathbb{E}X_*(\gamma,\eta)]^2 \tag{11}$$

then Corollary 2 will imply that $F_k(n,m)$ is w.h.p. p'-satisfiable for all $p' < p$.

The second, more severe, problem is that $\mathbb{E}X_*(\gamma,\eta)^2$ becomes exponentially greater than $[\mathbb{E}X_*(\gamma,\eta)]^2$ when r is only, roughly, half the lower bound of Theorem 1. To overcome this difficulty we proceed as follows. Fix $\gamma,\eta > 0$ and note that

$$\begin{aligned}
\mathbb{E}[X_*(\gamma,\eta)]^2 &= \mathbb{E}\left[\left(\sum_{\sigma} \gamma^{H(\sigma,F)}\eta^{U(\sigma,F)-u_0 m}\,\mathbf{1}_{\{\sigma\in\mathcal{S}^*\}}\right)^2\right] \\
&= \sum_{\sigma,\tau} \mathbb{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)-2u_0 m}\,\mathbf{1}_{\{\sigma,\tau\in\mathcal{S}^*\}}\right] \quad .
\end{aligned} \tag{12}$$

Now, 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 $\tilde{\gamma} \geq \gamma$ and $\tilde{\eta} \geq \eta$,

$$\begin{aligned}
\mathbb{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)-2u_0 m}\,\mathbf{1}_{\{\sigma,\tau\in\mathcal{S}^*\}}\right] &\leq \mathbb{E}\left[\tilde{\gamma}^{H(\sigma,F)+H(\tau,F)}\tilde{\eta}^{U(\sigma,F)+U(\tau,F)-2u_0 m}\,\mathbf{1}_{\{\sigma,\tau\in\mathcal{S}^*\}}\right] \\
&\leq \mathbb{E}\left[\tilde{\gamma}^{H(\sigma,F)+H(\tau,F)}\tilde{\eta}^{U(\sigma,F)+U(\tau,F)-2u_0 m}\right] \quad .
\end{aligned} \tag{13}$$

In other words, when using the right hand side of (13) to bound each term of the sum in (12), we are allowed to *adapt* the value of $\tilde{\gamma}$ and $\tilde{\eta}$ to the pair σ,τ, the only restrictions being $\tilde{\gamma} \geq \gamma$ and $\tilde{\eta} \geq \eta$. This is a crucial point and we will exploit it heavily when bounding the contribution of pairs σ,τ which agree on many variables. The resulting adaptive weighting scheme leads to an extremely delicate asymptotic analysis in the proof of (11).

3 Probabilistic Preliminaries

Relationship to other k-CNF models: Recall that the m clauses of $F_k(n,m)$ are chosen independently with replacement among all $(2n)^k$ 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. Nevertheless, observe that the probability that any given clause is improper is smaller than k^2/n and, moreover, that 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, $F_k(n, m = rn)$ 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.

3.1 The upper bound

As remarked in the Introduction, the upper bound in (3) can be readily derived from the entropic-form Chernoff bound for the binomial distribution (see Lemma A.10 in [3] or Lemma 3.8 in [8]). Nevertheless, in deriving the lower bound of Theorem 1 it will be informative to have a self-contained proof of the upper bound.

Define for $\alpha, \gamma, \eta \in [0,1]$,

$$f(\alpha,\gamma,\eta) = \eta^{-2u_0}\left[\left(\alpha\left(\frac{\gamma^2+\gamma^{-2}}{2}\right)+1-\alpha\right)^k - 2(1-\eta)\left(\frac{\alpha\gamma^{-2}+(1-\alpha)}{2}\right)^k + (1-\eta)^2\left(\frac{\alpha\gamma^{-2}}{2}\right)^k\right], \quad (14)$$

where u_0 is as in (9). Additionally, let

$$g_r(\alpha,\gamma,\eta) = \frac{f(\alpha,\gamma,\eta)^r}{\alpha^\alpha(1-\alpha)^{1-\alpha}} \quad . \quad (15)$$

Lemma 2. *For $X(\gamma,\eta)$ as in (8) with $m = rn$, for every $0 < \gamma, \eta < 1$,*

$$\mathbb{E}X(\gamma,\eta) = \frac{2^n}{\eta^{u_0 m}}\left[\left(\frac{\gamma+\gamma^{-1}}{2}\right)^k - \frac{1-\eta}{(2\gamma)^k}\right]^m = \left[2g_r\left(\frac{1}{2},\gamma,\eta\right)\right]^{n/2} \quad . \quad (16)$$

Proof. The second equality in (16) follows directly from the definitions of f, g_r. Write $F = c_1 \wedge \cdots \wedge c_m$, where the c_i's are k-clauses. Observe that for every truth assignment σ, $H(\sigma,F) = \sum_{i=1}^{rn} H(\sigma,c_i)$, and similarly for $U(\sigma,F)$. Since $c_1,...,c_{rn}$ are i.i.d. it follows that if c denotes a random clause

$$\eta^{u_0 m}\mathbb{E}X(\gamma,\eta) = 2^n\left[\mathbb{E}\left(\gamma^{H(\sigma,c)}\eta^{U(\sigma,c)}\right)\right]^m .$$

Observe that $U(\sigma,c) = 1$ when c is violated by σ, i.e., with probability 2^{-k}; otherwise $U(\sigma,c) = 0$. Writing $c = \ell_1 \vee \cdots \vee \ell_k$, where $\ell_1,\ldots,\ell_k$ are i.i.d. uniformly distributed literals, it follows that:

$$\begin{aligned}
\mathbb{E}\left[\gamma^{H(\sigma,c)}\eta^{U(\sigma,c)}\right] &= \mathbb{E}\gamma^{H(\sigma,c)} - \mathbb{E}\left[\gamma^{H(\sigma,c)}\left(1-\eta^{U(\sigma,c)}\right)\right] \\
&= \left[\mathbb{E}\gamma^{H(\sigma,\ell_1)}\right]^k - \frac{1-\eta}{2^k\gamma^k} \\
&= \left(\frac{\gamma+\gamma^{-1}}{2}\right)^k - \frac{1-\eta}{2^k\gamma^k} \\
&= f(1/2,\gamma,\eta) \quad . \qquad (17)
\end{aligned}$$

□

Lemma 3. *For all $k \geq 2$ and $p \in (0,1]$, if $q = 1-p$ then*

$$r_k^*(p) \leq \frac{2^k \log 2}{q\log q - (2^k-q)\log\left(\frac{2^k-1}{2^k-q}\right)} < T_k(p) \quad , \quad (18)$$

where $T_k(\cdot)$ was defined in (2).

Proof. The right hand inequality of (18) follows from the inequality $\log t \leq t-1$ applied to $t = \frac{2^k-1}{2^k-q}$, so we just need to verify the left hand inequality. Recall that $u_0 = 2^{-k}q$. 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

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

By Lemma 2 we have

$$\mathbf{P}[X(1,\eta) \geq 1] \leq \mathbb{E}[X(1,\eta)] = 2^n\eta^{-qrn2^{-k}}\left(1-(1-\eta)2^{-k}\right)^{rn} . \quad (19)$$

Thus, the probability of p-satisfiability decays exponentially in n if the the n-th root of the RHS of (19) is strictly smaller than 1. Taking $\eta = q(2^k-1)/(2^k-q)$ yields the lemma. □

3.2 The Lower Bound: Groundwork

Our first order of business is to show that for an appropriate choice of γ and η, the truncation replacing $X(\gamma,\eta)$ by $X_*(\gamma,\eta)$ does not reduce the expectation by more than a constant factor. The idea behind the proof below is motivated by Cramer's classical "change of measure" technique in large deviation theory.

Lemma 4. *Given $u_0 \in (0, 2^{-k})$, let γ_0, η_0 satisfy*

$$1 - \eta_0 \ = \ (1-\gamma_0^2)(1+\gamma_0^2)^{k-1} \qquad \textit{and} \qquad u_0 \ = \ \frac{\eta_0}{(1+\gamma_0^2)^k - (1-\eta_0)} \ . \tag{20}$$

There exists $\theta = \theta(k, \gamma_0, A) > 0$ such that as $n \to \infty$,

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

Remark. Since the function $\gamma \mapsto \frac{1-(1-\gamma^2)(1+\gamma^2)^{k-1}}{(1+\gamma^2)^k-(1-\gamma^2)(1+\gamma^2)^{k-1}}$ maps $(0,1)$ onto $(0,2^{-k})$, the existence of $\gamma_0, \eta_0 \in (0,1)$ satisfying (20) is guaranteed. Moreover, η_0, γ_0 are uniquely defined, but this fact will not be used.

Proof. It suffices to prove that there exists some $\theta = \theta(k,\gamma_0,A) > 0$ such that for the values of γ_0, η_0 satisfying (20) and every truth assignment σ, we have

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

Fix a truth assignment σ and consider an auxiliary distribution $\mathbf{P}_\sigma$ on k-CNF formulas where the clauses $c_1,\ldots,c_m$ are again i.i.d. among all $(2n)^k$ possible k-clauses, but where now for any fixed clause ω

$$\mathbf{P}_\sigma(c_i=\omega) = \frac{1}{(2n)^k}\cdot\frac{\gamma_0^{H(\sigma,\omega)}\eta_0^{U(\sigma,\omega)}}{Z(\gamma_0,\eta_0)} \ , \tag{22}$$

where

$$Z(\gamma_0,\eta_0) = \frac{1}{(2n)^k}\sum_c \gamma_0^{H(\sigma,c)}\eta_0^{U(\sigma,c)} = \mathbb{E}\left[\gamma_0^{H(\sigma,c)}\eta_0^{U(\sigma,c)}\right] = f(1/2,\gamma_0,\eta_0) \ . \tag{23}$$

Let $\mathbb{E}_\sigma$ 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

$$Z(\gamma_0,\eta_0)\mathbb{E}_\sigma[H(\sigma,c)] = k\ \frac{\gamma_0-\gamma_0^{-1}}{2}\left(\frac{\gamma_0+\gamma_0^{-1}}{2}\right)^{k-1} + \frac{k(1-\eta_0)}{(2\gamma_0)^k} \ . \tag{24}$$

Moreover,

$$Z(\gamma_0,\eta_0)\mathbb{E}_\sigma[U(\sigma,c)] = (2\gamma_0)^{-k}\eta_0 \, . \tag{25}$$

Thus, the first equation in (20) along with (24) ensure that $\mathbb{E}_\sigma[H(\sigma,c)] = 0$, while the second equation in (20) along with (23),(25) ensure that $\mathbb{E}_\sigma[U(\sigma,c)] = u_0$.

Next, we apply the multivariate central limit theorem (see, e.g. [17], 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 $n\to\infty$,

$$\mathbf{P}_\sigma[\sigma\in\mathcal{S}^*(F)] = \mathbf{P}_\sigma\Big[H(\sigma,F)\geq 0 \text{ and } U(\sigma,F)\in[mu_0, mu_0+A\sqrt{m}]\Big] \to \theta(k,\gamma_0,A) > 0 \ .$$

Here, the right hand side is the probability that a certain non-degenerate bivariate normal law assigns to a certain open set. Its exact value is unimportant for our purpose. By (22), this is equivalent to (21).

□

The next lemma bounds the second moment of $X_*(\gamma_0, \eta_0)$ from above:

Lemma 5. *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,*

$$\mathbb{E}[X_*(\gamma_0, \eta_0)]^2 \leq 2^n \sum_{z=0}^{n} \binom{n}{z} f\left(\frac{z}{n}, \gamma(z), \eta(z)\right)^{rn} \leq (n+1) \cdot \left[2 \cdot \max_{0 \leq z \leq n} g_r\left(\frac{z}{n}, \gamma(z), \eta(z)\right)\right]^n \quad , \tag{26}$$

where g_r is as in (15).

Proof. Fix $\gamma, \eta > 0$. For any pair of truth assignments σ, τ we first observe that since the m clauses $c_1, c_2, \ldots, c_m$ are i.i.d., letting c be a single random clause we have

$$\begin{aligned}
\mathbb{E}\left[\gamma^{H(\sigma,F)+H(\tau,F)}\eta^{U(\sigma,F)+U(\tau,F)}\right] &= \mathbb{E}\left[\prod_{i=1}^{m} \gamma^{H(\sigma,c_i)+H(\tau,c_i)}\eta^{U(\sigma,c_i)+U(\tau,c_i)}\right] \\
&= \prod_{i=1}^{m} \mathbb{E}\left[\gamma^{H(\sigma,c_i)+H(\tau,c_i)}\eta^{U(\sigma,c_i)+U(\tau,c_i)}\right] \\
&= \left(\mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\eta^{U(\sigma,c)+U(\tau,c)}\right]\right)^m \quad .
\end{aligned} \tag{27}$$

Next, we observe that for every pair σ, τ, by symmetry, the expectation in (27) depends only on the number of variables to which σ, τ assign the same value. So, 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 with $\eta^{U(\sigma,c)}$ we get

$$\begin{aligned}
&\mathbb{E}\left[\gamma^{H(\sigma,c)}\eta^{H(\sigma,c)}\gamma^{H(\tau,c)}\eta^{H(\tau,c)}\right] \\
&= \mathbb{E}\left[\left(\gamma^{H(\sigma,c)} - \gamma^{H(\sigma,c)}\left(1-\eta^{U(\sigma,c)}\right)\right)\left(\gamma^{H(\tau,c)} - \gamma^{H(\tau,c)}\left(1-\eta^{U(\tau,c)}\right)\right)\right] \\
&= \mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] - 2\,\mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\left(1-\eta^{U(\sigma,c)}\right)\right] + \mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\left(1-\eta^{U(\sigma,c)}\right)\left(1-\eta^{U(\tau,c)}\right)\right] \\
&= \mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] - 2(1-\eta)\mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\mathbf{1}_{\{\sigma \text{ violates } c\}}\right] + \frac{\alpha^k(1-\eta)^2}{2^k\gamma^{2k}} \quad .
\end{aligned} \tag{28}$$

To evaluate (28) we note that since the literals $\ell_1, \ell_2, \ldots \ell_k$ comprising c are i.i.d. we have

$$\mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\right] = \mathbb{E}\left[\prod_i \gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\right] = \prod_i \mathbb{E}\left[\gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\right] = \left(\alpha\left(\frac{\gamma^2+\gamma^{-2}}{2}\right) + 1 - \alpha\right)^k \tag{29}$$

and, similarly,

$$\mathbb{E}\left[\gamma^{H(\sigma,c)+H(\tau,c)}\mathbf{1}_{\{\sigma \text{ violates } c\}}\right] = \mathbb{E}\left[\prod_i \gamma^{H(\sigma,\ell_i)+H(\tau,\ell_i)}\mathbf{1}_{\{\sigma \text{ violates } \ell_i\}}\right] = \left(\frac{\alpha\gamma^{-2}+(1-\alpha)}{2}\right)^k \quad . \tag{30}$$

Substituting (29), (30) into (28) we get

$$\begin{aligned}
&\eta^{-2u_0}\,\mathbb{E}\left[\gamma^{H(\sigma,c)}\eta^{H(\sigma,c)}\gamma^{H(\tau,c)}\eta^{H(\tau,c)}\right] \\
&= \eta^{-2u_0}\left[\left(\alpha\left(\frac{\gamma^2+\gamma^{-2}}{2}\right)+1-\alpha\right)^k - 2(1-\eta)\left(\frac{\alpha\gamma^{-2}+(1-\alpha)}{2}\right)^k + (1-\eta)^2\left(\frac{\alpha\gamma^{-2}}{2}\right)^k\right] \\
&= f(\alpha, \gamma, \eta) \quad .
\end{aligned} \tag{31}$$

Since the number of ordered pairs with overlap z is $2^n \binom{n}{z}$, using the reasoning in (12) and (13), we get

$$\mathbb{E}\, X_*(\gamma_0, \eta_0)^2 \le 2^n \sum_{z=0}^{n} \binom{n}{z} f\left(\frac{z}{n}, \gamma(z), \eta(z)\right)^m \quad , \tag{32}$$

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 $0 \le z \le n$.

The final inequality in (26) follows from the fact that for all $n > 1$ and $z = 0, 1, \ldots, n$ Stirling's formula implies that $\binom{n}{z} \le [\tau(z/n)]^n$, where $\tau(\alpha) = \alpha^{-\alpha}(1-\alpha)^{\alpha-1}$ (see e.g. [1], eq. (18)) . ∎

The following corollary is a direct consequence of Lemmata 2, 4 and 5.

Corollary 3. *Let $\chi : [0,1] \to [\gamma_0, 1)$ and $\omega : [0,1] \to [\eta_0, 1)$ be arbitrary functions satisfying $\chi(1/2) = \gamma_0$, $\omega(1/2) = \eta_0$, and let $g_r(\alpha) = g_r(\alpha, \chi(\alpha), \omega(\alpha))$. Assume that $g_r(\alpha) \le g_r(1/2)$ for all $\alpha \in [0,1]$. Then there exists a constant $D = D_{\chi,\omega}(k, r, p) > 0$ such that for all sufficiently large n*

$$\mathbb{E}[X_*(\gamma_0, \eta_0)^2] \le Dn \cdot \left[\mathbb{E} X_*(\gamma_0, \eta_0)\right]^2.$$

Therefore, in order to prove Theorem 1 it is enough to show that for every $p \in (0,1]$ and $r = (1-\delta_k)T_k(p)$, there exist functions χ, ω for which the conditions of Corollary 3 hold. To simplify the analysis, we use the crudest possible such functions, paying the price of this simplicity in the value of the constant C in Proposition 6 below. We note that by choosing a more refined (and more cumbersome) adaptation of γ, η to α this value can be improved greatly. Moreover, we emphasize that for any fixed value of k, one can get a sharper lower bound for $r_k(p)$ by partitioning $[0,1]$ to a large number of intervals and numerically finding a good value of γ, η for each one. Indeed, the bounds reported in the plots in the Introduction are the result of such optimization. (We discuss this point further in Section 5).

Recall that $q = 1 - p = u_0 2^k$. We define $\varphi : [0,1) \to [1/2, 1)$ as

$$\varphi(q) = \frac{(1-\sqrt{q})^2}{1 - q + q\log q} \quad . \tag{33}$$

Theorem 1 will follow from the following proposition.

Proposition 6. *Let $g_r(\alpha, \gamma, \eta)$ be as in (15) and define*

$$G_r(\alpha) = \begin{cases} g_r(\alpha, \gamma_0, \eta_0) & \text{if } \alpha \in \left[\frac{3\log k}{k}, 1 - \frac{3\log k}{k}\right] \\ \\ g_r(\alpha, \sqrt{\gamma_0}, \sqrt{\eta_0}) & \text{otherwise.} \end{cases} \tag{34}$$

There exists a universal constant $C > 0$ such that if $r \le \frac{2^k \log 2}{1-q+q\log q}\left(1 - Ck2^{-k\varphi(q)}\right)$, then for all $\alpha \in [0,1]$, $G_r(\alpha) \le G_r(1/2)$.

The proof of Proposition 6 is presented in the following section. We remark that using the Laplace method from asymptotic analysis (see [5, §4.2] or Lemma 2 in [1]), a slight modification of the arguments below shows that the factor of n in Corollary 3 can be removed. However, this is immaterial for our purpose.

4 Asymptotic Analysis

We begin by observing that to prove Proposition 6 it suffices to consider $\alpha \in (1/2, 1]$.

Lemma 7. *For every $0 < x \le \frac{1}{2}$, $G_r(1/2 + x) > G_r(1/2 - x)$.*

Proof. Since the function $\alpha \mapsto \alpha^\alpha(1-\alpha)^{1-\alpha}$ is symmetric around $1/2$, it suffices to prove that for $x \in (0, 1/2]$

$$f\left(\frac{1}{2} + x, \gamma, \eta\right) > f\left(\frac{1}{2} - x, \gamma, \eta\right) \quad . \tag{35}$$

Substituting $\alpha = 1/2 + x$ in (14) and denoting $\varepsilon = 1 - \gamma^2$ we get that for all $x \in [-1/2, 1/2]$

$$\begin{aligned}
\eta^{y/2^{k-1}} 2^{2k} \gamma^{2k} f\left(\frac{1}{2} + x, \gamma, \eta\right) &= [2x\varepsilon^2 + (2-\varepsilon)^2]^k - 2\,[2x\varepsilon + (2-\varepsilon)]^k (1-\eta) + (1-\eta)^2 (1+2x)^k \\
&= \sum_{j=0}^{k} \binom{k}{j} 2^j x^j [\varepsilon^{2j} (2-\varepsilon)^{2(k-j)} - 2(1-\eta)\varepsilon^j (2-\varepsilon)^{k-j} + (1-\eta)^2] \\
&= \sum_{j=0}^{k} \binom{k}{j} 2^j x^j [\varepsilon^j (2-\varepsilon)^{k-j} - (1-\eta)]^2. \qquad (36)
\end{aligned}$$

Thus, $f(1/2 + x, \gamma, \eta) - f(1/2 - x, \gamma, \eta) = \sum_{j=1}^{\lceil k/2 \rceil} a_j x^{2j-1}$, where at most one of the a_j's is zero. $\square$

Before proceeding with the rest of the proof, we introduce some notation, and prove some elementary estimates that will be used throughout the remainder of this section.

Recalling the definition of η_0 from (20) and substituting into the definition of f from (14) we get

$$\begin{aligned}
&f(\alpha, \gamma_0, \eta_0) = \eta_0^{-q/2^{k-1}} \\
&\quad \cdot \left[\left(1 - \alpha + \alpha \frac{\gamma_0^2 + \gamma_0^{-2}}{2}\right)^k - 2\,(1-\gamma_0^2)(1+\gamma_0^2)^{k-1} \left(\frac{\alpha\gamma_0^{-2} + 1 - \alpha}{2}\right)^k + \frac{\alpha^k (1-\gamma_0^2)^2 (1+\gamma_0^2)^{2k-2}}{2^k \gamma_0^{2k}} \right].
\end{aligned}$$

For some parts of the ensuing calculations, it will be convenient to use the following normalizations of $f(\alpha, \gamma_0, \eta_0)$ and $g_r(\alpha, \gamma_0, \eta_0)$ denoted as f_0 and g_0, respectively,

$$f_0(\alpha) = 2^{2k} \gamma_0^{2k} \eta_0^{q/2^{k-1}} f(\alpha, \gamma_0, \eta_0) \quad \text{and} \quad g_0(\alpha) = 2^{2kr} \gamma_0^{2kr} \eta_0^{qr/2^{k-1}} g_r(\alpha, \gamma_0, \eta_0). \qquad (37)$$

We will also write $\varepsilon_0 = 1 - \gamma_0^2$. With this notation, we have the following formula valid for all $x \in [-1/2, 1/2]$

$$f_0\left(\frac{1}{2} + x\right) = [2x{\varepsilon_0}^2 + (2-\varepsilon_0)^2]^k - 2\varepsilon_0 (2-\varepsilon_0)^{k-1} [2 - \varepsilon_0 + 2x\varepsilon_0]^k + {\varepsilon_0}^2 (2-\varepsilon_0)^{2k-2} (1+2x)^k \ . \qquad (38)$$

In particular,

$$f_0\left(\frac{1}{2}\right) = (2-\varepsilon_0)^{2k} - 2\varepsilon_0 (2-\varepsilon_0)^{2k-1} + {\varepsilon_0}^2 (2-\varepsilon_0)^{2k-2} = 4(1-\varepsilon_0)^2 (2-\varepsilon_0)^{2k-2} \ . \qquad (39)$$

The function $y \mapsto 1 - y + y \log y$, defined on $[0,1]$, appears throughout our analysis. The following elementary inequalities, valid for all $y \in [0,1]$, will be used

$$\frac{(1-y)^2}{2} \le 1 - y + y \log y \le (1-y)^2 \ . \qquad (40)$$

Recalling that $q = 1 - p$ and Equations (9), (20) we claim that

$$\varepsilon_0 = \frac{2(1-q)}{2^k - k - 1} + O\left(\frac{k(1-q)^2}{2^{2k}}\right) \ , \qquad (41)$$

and

$$\eta_0 = q - \frac{(k+1)(1-q)}{2^k - k - 1} + O\left(\frac{k(1-q)^2}{2^k}\right) \ . \qquad (42)$$

By the first equation in (20) we have that $\eta_0 = 1 - \varepsilon_0 (2-\varepsilon_0)^{k-1}$. Therefore, (42) is a consequence on (41). To prove (41) let

$$\psi(t) = \sum_{j=1}^{k-1} \frac{1}{(2-t)^j} = \frac{(2-t)^{k-1} - 1}{(1-t)(2-t)^{k-1}} \ . \qquad (43)$$

Evidently, ψ is increasing, and for $0 \le t \le 1/k$,

$$\psi(t) = 1 - \frac{1}{2^{k-1}} + t - \frac{(k+1)t}{2^k} + O(t^2) \ . \tag{44}$$

Additionally, using the second equation in (20) we find that $\psi(\varepsilon_0) = 1 - \frac{q}{2^{k-1}}$. Since ψ is increasing, (41) will be proved once we show that for some constants c_1, c_2,

$$\psi\left(\frac{2(1-q)}{2^k-k-1} + \frac{c_1(1-q)^2}{2^{2k}}\right) \le 1 - \frac{q}{2^{k-1}} \le \psi\left(\frac{2(1-q)}{2^k-k-1} + \frac{c_2(1-q)^2}{2^{2k}}\right) \ , \tag{45}$$

and (45) is a straightforward consequence of (44).

The following lemma proves Proposition 6 in the first range of the definition of G_r.

Lemma 8. *For all $k \ge k_0$, if $r \le \frac{2^k \log 2}{1-q+q\log q}$ then for $\alpha \in \left[\frac{1}{2}, 1 - \frac{3\log k}{k}\right]$ we have $G_r(\alpha) \le G_r(1/2)$.*

Proof. In this proof we use the normalization (37). It is enough to prove that $g_0'(\alpha) < 0$ for $\frac{1}{2} < \alpha \le 1 - \frac{3\log k}{k}$. Now,

$$g_0'(\alpha) = \frac{f_0(\alpha)^{r-1}\Big\{rf_0'(\alpha) + f_0(\alpha)\big[\log(1-\alpha) - \log\alpha\big]\Big\}}{\alpha^\alpha(1-\alpha)^{1-\alpha}} \ . \tag{46}$$

Differentiating (38) at $x = 0$ we find that

$$f_0'\left(\frac{1}{2}\right) = 2k{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2} - 4{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2} + 2{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2} = 0 \ .$$

Since, by (36), $f_0(\alpha) > 0$ it is enough to show that the following function is decreasing on $\left[\frac{1}{2}, 1 - \frac{3\log k}{k}\right]$

$$\zeta(\alpha) = rf_0'(\alpha) + f_0(\alpha)[\log(1-\alpha) - \log\alpha] \ .$$

Now,

$$\zeta'(\alpha) = rf_0''(\alpha) + f_0'(\alpha)[\log(1-\alpha) - \log\alpha] - f_0(\alpha)\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right) \ .$$

Since for $1/2 < \alpha \le 1$, $\log(1-\alpha) < \log\alpha$ and, by (36), $f_0' > 0$ on $(1/2, 1]$, it is thus enough to prove that

$$rf_0''(\alpha) \le f_0(\alpha)\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right) \ .$$

Now, $\frac{1}{\alpha} + \frac{1}{1-\alpha} \ge 4$ and, using (39), we get that for $\alpha \ge 1/2$,

$$f_0(\alpha) \ge f_0\left(\frac{1}{2}\right) = 4(1-\varepsilon_0)^2(2-\varepsilon_0)^{2k-2} \ge (2-\varepsilon_0)^{2k-2} \ ,$$

where for the last inequality we used that $\varepsilon_0 \le 1/2$ for k large enough (by (41)). Thus, it suffices to prove that for all $x \le 1/2 - \frac{3\log k}{k}$

$$rf_0''\left(\frac{1}{2} + x\right) \le 4(2-\varepsilon_0)^{2k-2} \ . \tag{47}$$

Now, using that $x \le \frac{1}{2} - \frac{3\log k}{k}$, we differentiate (38) twice to get

$$
\begin{aligned}
&f_0''\left(\frac{1}{2}+x\right)\\
&= 4k(k-1)\left\{{\varepsilon_0}^4[2x{\varepsilon_0}^2+(2-\varepsilon_0)^2]^{k-2} - 2{\varepsilon_0}^3(2-\varepsilon_0)^{k-1}[2-\varepsilon_0+2x\varepsilon_0]^{k-2} + {\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}(1+2x)^{k-2}\right\}\\
&\le 4k^2\left\{{\varepsilon_0}^4(2-\varepsilon_0)^{2k-4}\left(1+\frac{2x{\varepsilon_0}^2}{(2-\varepsilon_0)^2}\right)^{k-2} + {\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}(1+2x)^{k-2}\right\}\\
&\le 4k^2\left\{{\varepsilon_0}^4(2-\varepsilon_0)^{2k-4}\left(1+2x\right)^{k-2} + {\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}(1+2x)^{k-2}\right\}\\
&\le 8k^2{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}(1+2x)^k\\
&\le 8k^2{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}\left[2-\frac{6\log k}{k}\right]^k\\
&\le 8k^2{\varepsilon_0}^2(2-\varepsilon_0)^{2k-2}\,\frac{2^k}{k^3}\\
&\le \frac{2^{k+3}}{k}\left(\frac{4(1-q)}{2^k}\right)^2(2-\varepsilon_0)^{2k-2}\ ,
\end{aligned}
$$

where in the last line we used the fact that for k large enough (41) implies $\varepsilon_0 \le 4(1-q)/2^k$.

Combining this estimate with (47), we see that we must show that for sufficiently large k

$$\frac{128\log 2}{1-q+q\log q}\cdot\frac{(1-q)^2}{k} \le 4\ ,$$

and this is indeed the case by (40). □

It remains to prove that $G_r(\alpha) < G_r(1/2)$ for $1-\frac{3\log k}{k} \le \alpha \le 1$. This inequality simplifies to:

$$\left[\frac{f(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0})}{f(1/2,\gamma_0,\eta_0)}\right]^r < 2\alpha^\alpha(1-\alpha)^{1-\alpha}\ . \tag{48}$$

The following lemma gives an upper bound for the left-hand side of (48).

Lemma 9. *There exists a constant $c > 0$ such that for all sufficiently large k and $\alpha \in [1/2, 1]$,*

$$\frac{f(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0})}{f(1/2,\gamma_0,\eta_0)} < q^{q/2^k}\left[1+\frac{2\sqrt{q}-2q+(1-\sqrt{q})^2\alpha^k}{2^k}+\frac{ck(1-q)^2}{2^{2k}}\right]\ . \tag{49}$$

Proof. Denote $\varepsilon_1 = 1-(\sqrt{\gamma_0})^2 = 1-\sqrt{1-\varepsilon_0} = \frac{\varepsilon_0}{2}+O(\varepsilon_0^2)$. Now,

$$
\begin{aligned}
\eta_0^{q/2^k}\gamma_0^k f(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0}) &= \left(1-\varepsilon_1+\frac{\alpha\varepsilon_1^2}{2}\right)^k - \frac{2(1-\sqrt{\eta_0})(1-\varepsilon_1(1-\alpha))^k+(1-\sqrt{\eta_0})^2\alpha^k}{2^k}\\
&= 1-\frac{k\varepsilon_0}{2}+\frac{k^2\varepsilon_0^2}{8}+O(k\varepsilon_0^2)-\frac{2(1-\sqrt{\eta_0})+(1-\sqrt{\eta_0})^2\alpha^k}{2^k}+O\left(\frac{k(1-\sqrt{\eta_0})\varepsilon_0}{2^k}\right)\\
&= 1-\frac{k\varepsilon_0}{2}+\frac{k^2\varepsilon_0^2}{8}-\frac{2(1-\sqrt{\eta_0})+(1-\sqrt{\eta_0})^2\alpha^k}{2^k}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\ ,
\end{aligned} \tag{50}
$$

where in (50) we have used (41) and (42). Observe now that $q > \eta_0$ since

$$\frac{q}{2^k} = u_0 = \frac{\eta_0}{(1+\gamma_0^2)^k-(1-\eta_0)} = \frac{\eta_0}{2^k(1-\varepsilon_0)(1-\varepsilon_0/2)^{k-1}} > \frac{\eta_0}{2^k}\ ,$$

where for the second and third equalities we used, respectively, the definitions of u_0 and η_0 from (20). The fact $q > \eta_0$ implies that for every $\alpha \in [0, 1]$,

$$(1-\sqrt{\eta_0})^2\alpha^k - 2(1-\sqrt{\eta_0}) \le (1-\sqrt{q})^2\alpha^k - 2(1-\sqrt{q}) + \eta_0 - q\ , \tag{51}$$

since the right-hand side minus the left-hand side of (51) equals $(1-\alpha^k)\left((2\sqrt{q}-q)-(2\sqrt{\eta_0}-\eta_0)\right) > 0$. Using (42) to bound η_0 in the right hand side of (51) we conclude that

$$(1-\sqrt{\eta_0})^2\alpha^k - 2(1-\sqrt{\eta_0}) \leq (1-\sqrt{q})^2\alpha^k - 2(1-\sqrt{q}) - \frac{(k+1)(1-q)}{2^k-k-1} + O\left(\frac{k(1-q)^2}{2^k}\right) \ .$$

Substituting this estimate into (50), we get

$$\begin{aligned} &\eta_0^{q/2^k}\gamma_0^k f\left(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0}\right) \\ &= \ 1-\frac{k\varepsilon_0}{2}+\frac{k^2\varepsilon_0^2}{8}-\frac{2(1-\sqrt{q})-(1-\sqrt{q})^2\alpha^k}{2^k}-\frac{(k+1)(1-q)}{2^k(2^k-k-1)}+O\left(\frac{k(1-q)^2}{2^{2k}}\right) \ . \end{aligned} \tag{52}$$

Using the identity (39) we can rewrite the ratio in (48) as

$$\frac{f(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0})}{f(1/2,\gamma_0,\eta_0)} = \frac{\eta_0^{q/2^k}(1-\varepsilon_0)^{\frac{k}{2}-2}}{(1-\varepsilon_0/2)^{2k-2}}\cdot\eta_0^{q/2^k}\gamma_0^k f\left(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0}\right) \ . \tag{53}$$

By (42)

$$\eta_0^{q/2^k} = q^{q/2^k}\left[1-\frac{(k+1)(1-q)}{2^k(2^k-k-1)}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right] \ . \tag{54}$$

Expanding around ε_0 and using (41) we get

$$\frac{(1-\varepsilon_0)^{\frac{k}{2}-2}}{(1-\varepsilon_0/2)^{2k-2}}=1+\left(\frac{k}{2}+1\right)\varepsilon_0+\frac{k^2}{8}\varepsilon_0^2+O(k\varepsilon_0^2) = 1+\left(\frac{k}{2}+1\right)\varepsilon_0+\frac{k^2}{8}\varepsilon_0^2+O\left(\frac{k(1-q)^2}{2^{2k}}\right) \ . \tag{55}$$

Direct substitution of (52),(54),(55) into (53) gives (56). Collecting terms, noting two cancellations and using (41) to bound the remaining terms involving ε_0 gives (57)

$$\begin{aligned} \frac{f(\alpha,\sqrt{\gamma_0},\sqrt{\eta_0})}{f(1/2,\gamma_0,\eta_0)} \ &= \ q^{q/2^k}\left[1-\frac{(k+1)(1-q)}{2^k(2^k-k-1)}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right]\cdot \\ &\quad \left[1+\left(\frac{k}{2}+1\right)\varepsilon_0+\frac{k^2}{8}\varepsilon_0^2+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right]\cdot \qquad (56) \\ &\quad \left[1-\frac{k\varepsilon_0}{2}+\frac{k^2\varepsilon_0^2}{8}-\frac{2(1-\sqrt{q})-(1-\sqrt{q})^2\alpha^k}{2^k}-\frac{(k+1)(1-q)}{2^k(2^k-k-1)}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right] \\ &= \ q^{q/2^k}\left[1+\varepsilon_0-\frac{2(k+1)(1-q)}{2^k(2^k-k-1)}-\frac{2(1-\sqrt{q})-(1-\sqrt{q})^2\alpha^k}{2^k}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right] \qquad (57) \end{aligned}$$

Using (41) once more, (57) becomes

$$q^{q/2^k}\left[1+\frac{2(1-q)}{2^k-k-1}-\frac{2(k+1)(1-q)}{2^k(2^k-k-1)}-\frac{2(1-\sqrt{q})-(1-\sqrt{q})^2\alpha^k}{2^k}+O\left(\frac{k(1-q)^2}{2^{2k}}\right)\right] \ . \tag{58}$$

Observe now that

$$\frac{2(1-q)}{2^k-k-1}-\frac{2(k+1)(1-q)}{2^k(2^k-k-1)}-\frac{2(1-\sqrt{q})-(1-\sqrt{q})^2\alpha^k}{2^k} = \frac{2\sqrt{q}-2q+(1-\sqrt{q})^2\alpha^k}{2^k}$$

which concludes the proof. □

We are now in position to conclude the proof of Proposition 6.

Proof of Proposition 6. By (49), there is a constant $c > 0$ such that for $\alpha \in [1/2, 1]$

$$\frac{f(\alpha, \sqrt{\gamma_0}, \sqrt{\eta_0})}{f(1/2, \gamma_0, \eta_0)} < q^{q/2^k} \left[1 + \frac{2\sqrt{q} - 2q + (1-\sqrt{q})^2 \alpha^k}{2^k} + \frac{c\,k(1-q)^2}{2^{2k}} \right] .$$

Hence, to prove (48) it is enough to show that

$$\left\{ q^{q/2^k} \left[1 + \frac{2\sqrt{q} - 2q + (1-\sqrt{q})^2 \alpha^k}{2^k} + \frac{ck(1-q)^2}{2^{2k}} \right] \right\}^r < 2\alpha^\alpha (1-\alpha)^{1-\alpha} .$$

Taking logarithms and using the inequality $\log(1+x) \leq x$, this amounts to showing

$$\frac{r}{2^k} \left[q \log q + 2\sqrt{q} - 2q + (1-\sqrt{q})^2 \alpha^k + \frac{ck(1-q)^2}{2^k} \right] \leq \log 2 - h(\alpha) \ , \tag{59}$$

where $h(\alpha) = -\alpha \log \alpha - (1-\alpha) \log(1-\alpha)$.

To simplify notation let us define

$$A = (1-\sqrt{q})^2 \quad \text{and} \quad B = q \log q + 2\sqrt{q} - 2q + \frac{ck(1-q)^2}{2^k} \tag{60}$$

and observe that

$$\frac{A}{A+B} = \varphi(q)(1 + O(k2^{-k})) \ . \tag{61}$$

With this notation (59) becomes

$$\frac{r}{2^k} \leq \frac{\log 2 - h(\alpha)}{A\alpha^k + B} \equiv M(\alpha) \ , \tag{62}$$

and this should hold for all $\alpha \geq 1 - \frac{3 \log k}{k}$.

We need to determine the minimal value of M on the interval $\left[1 - \frac{3\log k}{k}, 1\right]$. The derivative of M is

$$M'(\alpha) = \frac{(A\alpha^k + B) \cdot [\log \alpha - \log(1-\alpha)] - kA\alpha^{k-1}[\log 2 - h(\alpha)]}{(A\alpha^k + B)^2} \ . \tag{63}$$

In particular, $M'(1) = \infty$, so M cannot be minimized at $\alpha = 1$. Moreover, we claim that $M(\alpha) > M(1)$ for $\alpha \in \left[1 - \frac{3\log k}{k}, 2^{-1/k}\right]$. To prove this observe that $h(\alpha) = o(1)$ since $\alpha \geq 1 - \frac{3\log k}{k}$. Since, also, $\alpha \leq 2^{-1/k}$

$$M(\alpha) \geq \frac{\log 2 - o(1)}{A\left(2^{-1/k}\right)^k + B} \geq \frac{\log 2\,(1 - o(1))}{\frac{A}{2} + B} \ .$$

On the other hand, $M(1) = \log 2/(A+B)$. Thus, by (61), we have $M(1)/M(\alpha) \to 1 - \varphi(q)/2$.

It remains to bound $M(\alpha)$ from below when $\alpha > 2^{-1/k}$ and $M'(\alpha) = 0$. For such α, by (63), we have

$$-\log(1-\alpha) = -\log \alpha + \frac{kA\alpha^{k-1}}{A\alpha^k + B}[\log 2 - h(\alpha)] \ . \tag{64}$$

Since $\alpha > 2^{-1/k}$, $h(\alpha) = o(1)$. Hence (64) implies that $\alpha \geq 1 - e^{-k/8}$ since for k large enough we must have

$$-\log(1-\alpha) \geq \frac{kA}{2(A+B)} (\log 2 - o(1)) > \frac{k\varphi(q)}{4} \geq \frac{k}{8} \ . \tag{65}$$

Invoking (64) again, now starting with the premise $\alpha > 1 - e^{-k/8}$, we get that for all sufficiently large k,

$$-\log(1-\alpha) > k\varphi(q) \log 2\,(1 - e^{-k/9}) \tag{66}$$

and, thus, $\alpha > 1 - \exp(-k\varphi(q)\log 2\,(1-e^{-k/9}))$. Since $h(1-e^{-z}) \le e^{-z}(1+z)$ for all $z \ge 0$ we get

$$h(\alpha) \le k2^{-k\varphi(q)(1-e^{-k/9})} = O(k2^{-k\varphi(q)}) \ , \tag{67}$$

where for the equality above we used that $\varphi(q) \in [1/2, 1)$. Therefore,

$$M(\alpha) = \frac{\log 2 - h(\alpha)}{A\alpha^k + B} \ge \frac{\log 2}{A+B}\left(1 - O\left(k2^{-k\varphi(q)}\right)\right) \ge \frac{\log 2}{1-q+q\log q}\left(1 - O\left(k2^{-k\varphi(q)}\right)\right) \ .$$

Hence the restriction on r in (62) becomes $r \le \frac{2^k \log 2}{1-q+q\log q}\left(1 - O\left(k2^{-k\varphi(q)}\right)\right)$, proving Proposition 6. $\square$

5 Bounds for Small values of k

As mentioned in Section 3, for small values of k the simple adaptation scheme of Proposition 6 does not yield the best possible lower bound for p-satisfiability afforded by our method. For that, one has to use a significantly more refined adaptation of γ, η with respect to α. Our lower bounds reported in Figure 1 are, indeed, the result of performing such optimization of γ, η numerically (For both the upper bound plots and the plots of the lower bound from [6] we used the explicit formulas.)

Specifically, to create the plots of the lower bounds we computed a lower bound for 100 equally spaced values of p on the horizontal axis and then had Maple's [18] plotting function "connect the dots". For each of these values of p, to prove the corresponding lower bound for r we had to establish that there exist a choice of functions χ, ω as in Lemma 3 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 $\gamma \ge \gamma_0$ and $\eta \ge \eta_0$ such that this condition holds with a bit of room. (For $k > 4$ we solved (20), 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 $g_r(1/2, \gamma_0, \eta_0)$.

Acknowledgement

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

References

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

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

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

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

[5] N. G. de Bruijn. *Asymptotic methods in analysis.* Dover Publications Inc., New York, 1981.

[6] Don Coppersmith, David Gamarnik, Mohammad T. Hajiaghayi, and Gregory B. Sorkin, *Random MAX 2-SAT and MAX CUT*, 14th Annual Symposium on Discrete Algorithms (Baltimore, MD, 2003), New York, 2003.

[7] Stephen A. Cook, *The complexity of theorem-proving procedures*, 3rd Annual Symposium on Theory of Computing (Shaker Heights, OH, 1971), New York, 1971, pp. 151–158.

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

[9] A. Dembo, Y. Peres, J. Rosen and O. Zeitouni. Thick points for planar Brownian motion and the Erdős-Taylor conjecture on random walk. *Acta Math.*, 186:239–270, 2001.

[10] 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.

[11] Olivier Dubois, Rémi Monasson, Bart Selman and Riccardo Z. Zecchina editors, *Phase Transitions in Combinatorial Problems*, Theoretical Computer Science, volume 265, numbers 1-2 (2001).

[12] P. Erdős and S. J. Taylor. Some problems concerning the structure of random walk paths. *Acta Sci. Hung.* 11:137–162, 1960.

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

[14] Johan Håstad, *Some optimal inapproximability results*, Journal of ACM **48** (2001), 798–859.

[15] Svante Janson, Yannis C. Stamatiou and Malvina Vamvakari, *Bounding the unsatisfiability threshold of random 3-SAT*, Random Structures and Algorithms **17** (2000), no. 2, 103-116.

[16] Svante Janson, *The second moment method, conditioning and approximation*, Random discrete structures (Minneapolis, MN, 1993), 175–183, IMA Vol. Math. Appl., 76, Springer, New York, 1996.

[17] 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

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

[19] Michel Talagrand, *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models*, Ergebnisse Der Mathematik Und Ihrer Grenzgebiete, 3. Folge, Bd. 46. (2003).