Tail Bounds for Occupancy and the Satisfiability Threshold Conjecture

Anil Kamath*
Department of Computer Science
Stanford University
Stanford, CA 94305-2140

Rajeev Motwani†
Department of Computer Science
Stanford University
Stanford, CA 94305-2140

Krishna Palem
Department of Computer Science
Courant Institute of Mathematical Sciences
New York, NY 10012

Paul Spirakis‡
Computer Technology Institute
Patras University
Greece

December 6, 1994

Abstract

The classical occupancy problem is concerned with studying the number of empty bins resulting from a random allocation of m balls to n bins. We provide a series of tail bounds on the distribution of the number of empty bins. These tail bounds should find application in randomized algorithms and probabilistic analysis. Our motivating application is the following well-known conjecture on threshold phenomenon for the satisfiability problem. Consider random 3-SAT formulas with cn clauses over n variables, where each clause is chosen uniformly and independently from the space of all clauses of size 3. It has been conjectured that there is a sharp threshold for satisfiability at $c^* \approx 4.2$. We provide the first non-trivial upper bound on the value of c^*, showing that for $c > 4.758$ a random 3-SAT formula is unsatisfiable with high probability. This result is based on a structural property, possibly of independent interest, whose proof needs several applications of the occupancy tail bounds.

*Supported by US Army Office Research Grant DAAL-03091-G-0102.

†Supported by NSF Young Investigator Award CCR-9357849, with matching funds from IBM, Schlumberger Foundation, Shell Foundation and Xerox Corporation.

‡Partially supported by the CEC ESPRIT Basic Research Project ALCOM II.

1 Introduction

Consider a *random allocation* of m balls to n bins where each ball is placed in a bin chosen uniformly and independently. The properties of the resulting distribution of balls among bins have been the subject of intensive study in the probability and statistics literature [JK77, KSC78]. In computer science, this process arises naturally in randomized algorithms and probabilistic analysis. Of particular interest is the *occupancy problem* where the random variable under consideration is the number of empty bins.

We develop a series of bounds (reminiscent of the Chernoff bound for binomial distributions) on the tail of the distribution of the number of empty bins; our tail bounds are successively tighter, but each new bound has a more complex closed form. Such strong bounds do not seem to have appeared in the earlier literature. Our motivating application was the *threshold conjecture* for satisfiability, and we present some significant progress towards settling this conjecture in the affirmative. The tail bounds for occupancy are an essential ingredient in the analysis of the satisfiability problem, as will become clear later.

The problem of determining the satisfiability of boolean formulas in the conjunctive normal form (CNF) has played a central role in the understanding of the complexity of computations. In addition, it is of tremendous practical interest, arising naturally in a variety of applications: program and machine testing, VLSI design and testing, logic programming, inference, machine learning, and constraint satisfaction. Given its NP-hardness the practitioners seek heuristic solutions to the problems of verifying the satisfiability of large formulas, or enumerating all satisfying truth assignments [Gol79]. Of late there has been increased interest in the analysis of the behavior of random formulas. Typically, it is assumed that the formula is in the CNF form and the clauses are chosen uniformly and independently. Recent research has concentrated on formulas with exactly k literals per clause (the k-SAT problem). The study of random 3-SAT formulas is particularly instructive as $k = 3$ is the smallest value at which the problem is NP-hard, and in any case the average-case behavior of 2-SAT is well-understood.

We define the probability space of random 3-SAT formulas as follows. The sample space $\Omega(n, c)$ consists of all 3-SAT formulas with $m = cn$ clauses over the n variables $X_1, X_2, \ldots, X_n$. A random formula is generated by choosing each clause independently from the following distribution: choose three distinct variables uniformly at random, and independently negate each variable with probability one half. The probability of satisfiability of a random 3-SAT formula decreases as $c = m/n$ increases. For sufficiently small c a random formula is very likely to be satisfiable since variables are unlikely to occur in both the negated and the unnegated form; similarly, for sufficiently large c it is easy to see that a random formula is very likely to be unsatisfiable. Empirical evidence [LT92, MSL92] suggests that there is a sharp threshold c^* such that when $c < c^*$ a random formula is almost surely satisfiable, and when $c > c^*$ a random formula is almost surely unsatisfiable. It is conjectured that the value of c^* is around 4.2. For random 2-SAT, a sharp threshold has been demonstrated at $m/n = 1$ by Chvátal-Reed [CR92], Goerdt [Goe92], and de la Vega [Veg92].

Our main result for satisfiability is a significant improvement over the obvious upper bound (from first-moment methods) on the value of the threshold. In particular, we show that random 3-SAT formulas with $c \geq 4.758$ are unsatisfiable with high probability.

The motivation for trying to understand the behavior of the threshold is two-fold. First, the precise knowledge of the behavior of this problem on either side of the threshold helps in the design of good heuristics. Second, it helps in isolating the region around the threshold where we do not — and most likely will not — have a precise understanding of the problem's behavior; intuitively, the intractability of the problem is likely to be "concentrated" around this region. It is expected that the structural results obtained in the process will help to establish some form of average-case hardness in this region. The following is our main structural result: for a random formula which is satisfiable, the number of satisfying truth assignments is almost surely an exponential factor larger than the expectation. A succinct encoding of these truth assignments can be determined in polynomial time.

Recent work has led to considerable progress on bounding the threshold from below. Chao-Franco [CF86] and Chvátal-Reed [CR92] have shown that some natural heuristics find satisfying truth assignments with high probability when $c \leq 1$. Broder-Frieze-Upfal [BFU93] have analyzed a simple heuristic which is shown to solve the problem for $c \leq 1.63$. Chao-Franco [CF90] have shown that a commonly-used heuristic succeeds in finding a satisfying truth assignment with a non-zero limiting probability for $c \leq 2.99$. Currently, the best known result is due to Frieze-Suen [FS93] who show that the lower bound on the threshold can be improved to 3.003.

Significantly less progress has been made on the bounding of the conjectured sharp threshold from above. An easy counting argument (due to Chvátal-Szemerédi [CS88] and Franco-Paull [FP83]) shows that for $c > 5.19$ a random formula is unsatisfiable with probability approaching 1 as n tends to infinity. Broder-Frieze-Upfal [BFU93] claimed an improvement to a bound of $5.19 - \epsilon$, where $\epsilon = 10^{-7}$. Recently, and independently of our results, Maftouhi-Vega [MV93] have claimed an upper bound of $c > 5.01$.

The rest of this paper is organized as follows. In Section 2 we present the tail bounds for the occupancy problem. The proofs of these new bounds are presented in Sections 2.1, 2.2, and 2.3. In Section 3 we present our main result for the satisfiability problem. For the sake of clarity, instead of the claimed bound of 4.758 on the threshold, we present a marginally weaker bound of 4.762. The proof of our main theorem is sketched in Section 3.1. A series of lemmas is presented in Section 3.2 establishing bounds on the number of satisfying truth assignments for a *satisfiable* random 3-SAT formula. These are used to obtain the desired results in Section 3.3. Establishing the promised bound requires computing the maximum value of a complicated algebraic function with several parameters. The reported result is based on a computational maximization, and an analytic proof of a marginally weaker bound is presented in the Appendix A.

2 Tail Bounds for Occupancy Problems

We will make use of the following notation in presenting sharp bounds on the tails of distributions. The notation $F \sim G$ will denote that $F = (1+o(1))G$; further, $F \asymp G$ will denote that $\ln F \sim \ln G$. When we prove that $f \sim g$, it will only be for the purposes of later claiming that $2^f \asymp 2^g$. These asymptotic equalities will be treated like actual equalities and it will be clear that the results claimed are unaffected by this "approximation."

We start by stating the Chernoff bound on the tail of the binomial distribution [AS92, MR94].

Theorem 1 (Binomial Bound) *Let $Y_1, Y_2, \cdots Y_n$ be i.i.d. Bernoulli trials, each with probability of success p. Define the binomial $B(n,p)$ random variable $Y = \sum_{i=1}^{n} Y_i$ and $\mu = \mathbf{E}[Y] = np > 0$. Further, let the function $F(n,p,y)$ denote the probability that $Y = y$. For, $0 \leq y \leq n$ we have*

$$F(n,p,y) \asymp \left(\frac{np}{y}\right)^{y} \left(\frac{n-np}{n-y}\right)^{n-y}. \tag{1}$$

Furthermore, if $y = (1+\theta)\mu$, then for $\theta \geq -1$, we have

$$F(n,p,y) \leq \exp(-((1+\theta)\ln[1+\theta] - \theta)\mu). \tag{2}$$

Consider now the probabilistic experiment of throwing m balls, independently and uniformly, into n bins.

Definition 1 *Let Z be the number of empty bins when m balls are placed randomly into n bins, and define $r = m/n$. Define the function $H(m, n, z)$ as the probability that $Z = z$. The expectation of Z is given by*

$$\mu = \mathbf{E}[Z] = n\left(1 - \frac{1}{n}\right)^m \sim ne^{-r}.$$

The following three theorems provide the bounds on the tail of the distribution of the random variable Z. The proof of the first bound[1] can be found in Section 2.1 and is based on a martingale argument. Recently, our first two bounds have been improved upon by Janson [Jan94] who has obvtained strong large deviation inequalities for sums of dependent random variables. Unfortunately, these new bounds do not help improve our main result for the satisfiability threshold since it is based on our third bound which is tight (but has no clean closed form).

Theorem 2 (Occupancy Bound 1) *For any $\theta > 0$,*

$$\mathbf{P}[|Z - \mu| \geq \theta\mu] \leq 2\exp\left(-\frac{\theta^2\mu^2(n - 1/2)}{n^2 - \mu^2}\right).$$

We remark that for large r this bound is asymptotically equal to

$$2\exp\left(-\frac{\theta^2 e^{-2r} n}{1 - e^{-2r}}\right).$$

The reader may wish to compare this with the following *heuristic* estimate of the tail probability assuming that the distribution of Z is well approximated by the approximating normal distribution also far out in the tails [JK77, KSC78].

$$\mathbf{P}[|Z - \mu| \geq \theta\mu] \leq 2\exp\left(-\frac{\theta^2 e^{-r} n}{2(1 - (1 + r)e^{-r})}\right).$$

The next two bounds are in terms of point probabilities rather than tail probabilities (as was the case in the Binomial Bound), but the unimodality of the distribution implies that the two differ by at most a small (linear) factor. These more general bounds on the point probability are essential for the application to the satisfiability problem. The next result is obtained via a generalization of the Binomial Bound to the case of *dependent* Bernoulli trials as described in Section 2.2.

Theorem 3 (Occupancy Bound 2) *For $\theta > -1$,*

$$H(m, n, (1 + \theta)\mu) \leq \exp(-((1 + \theta)\ln[1 + \theta] - \theta)\mu) \quad (3)$$

In particular, for $-1 \leq \theta < 0$,

$$H(m, n, (1 + \theta)\mu) \leq \exp\left(-\frac{\theta^2\mu}{2}\right). \quad (4)$$

[1] Although the first bound has a nice closed form, it is not strong enough for our purposes. The second bound is stronger and leads to the weaker result in Appendix A. The third bound is complex but exact, and is essential for deriving our main result for satisfiability.

The last result is proved in Section 2.3 using ideas from large deviations theory [Wei93].

Theorem 4 (Occupancy Bound 3) *For $|z-\mu| = \Omega(n)$,*

$$H(m,n,z) \asymp \exp\left(\left[-n\left(\int_0^{1-z/n} \ln\left[\frac{k-x}{1-x}\right] \mathrm{d}x - r\ln k\right)\right]\right) \tag{5}$$

where k is defined implicitly by the equation $z = n(1-k(1-e^{-r/k}))$.

2.1 The Martingale Bound for the Occupancy Problem

The proof of Theorem 2 uses a technique based on martingale analysis called *the method of bounded differences* [McD89]. We assume familiarity with the basics of martingale theory [AS92, HH80], including the martingale tail inequality due to Azuma [Azu67].

Theorem 5 (Azuma's Inequality) *Let X_0, X_1, ... be a martingale sequence such that for each k,*

$$|X_k - X_{k-1}| \leq c_k$$

where c_k may depend on k. Then, for all $t \geq 0$ and any $\lambda > 0$,

$$\mathbf{P}\left[|X_t - X_0| \geq \lambda\right] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_{k=1}^t c_k^2}\right).$$

Consider the process of placing m balls in n bins, and let time t refer to the stage at which the first t balls have been placed. Let the σ-field $\mathcal{F}_t$ consist of the events corresponding to the state at time t. Let Z be the random variable denoting the number of empty bins at time m, and let $Z_t = \mathbf{E}\left[Z \mid \mathcal{F}_t\right]$ denote the conditional expectation of Z at time t. It can be shown that the random variables $\{Z_t\}_{t=0}^m$ form a martingale, with $Z_0 = \mathbf{E}\left[Z\right]$ and $Z_m = Z$.

It can be verified by direct calculation that

$$z(Y,t) = \mathbf{E}\left[Z \mid Y \text{ bins are empty at time } t\right] = Y\left(1-\frac{1}{n}\right)^{m-t}.$$

Let the random variable Y_t denote the number of empty bins at time t. It is then easy to see that

$$Z_{t-1} = z(Y_{t-1}, t-1) = Y_{t-1}\left(1-\frac{1}{n}\right)^{m-t+1}.$$

Suppose we are at time $t-1$ (i.e. in the σ-field $\mathcal{F}_{t-1}$.), and therefore the values of Y_{t-1} and Z_{t-1} are determined. Now, at time t there are two possibilities:

Case 1: with probability $1 - Y_{t-1}/n$, the tth ball goes into a currently non-empty bin. Then, $Y_t = Y_{t-1}$, and

$$Z_t = z(Y_t, t) = z(Y_{t-1}, t) = Y_{t-1}\left(1-\frac{1}{n}\right)^{m-t}.$$

Case 2: with probability Y_{t-1}/n, the tth ball goes into a currently empty bin. Then, $Y_t = Y_{t-1} - 1$, and

$$Z_t = z(Y_t, t) = z(Y_{t-1}-1, t) = (Y_{t-1}-1)\left(1-\frac{1}{n}\right)^{m-t}.$$

Let us now focus on the difference random variable $\Delta_t = Z_t - Z_{t-1}$. Corresponding to Z_t, the distribution of Δ_t (given the state at time $t-1$) can be characterized as follows.

Case 1: with probability $1 - Y_{t-1}/n$, the value of Δ_t is

$$\delta_0 = Y_{t-1}\left(1-\frac{1}{n}\right)^{m-t} - Y_{t-1}\left(1-\frac{1}{n}\right)^{m-t+1} = \frac{Y_{t-1}}{n}\left(1-\frac{1}{n}\right)^{m-t}.$$

Case 2: with probability Y_{t-1}/n, the value of Δ_t is

$$\delta_1 = (Y_{t-1}-1)\left(1-\frac{1}{n}\right)^{m-t} - Y_{t-1}\left(1-\frac{1}{n}\right)^{m-t+1} = -\left(1-\frac{Y_{t-1}}{n}\right)\left(1-\frac{1}{n}\right)^{m-t}.$$

Observing that $0 \leq Y_{t-1} \leq n$, it follows that the value of the difference is bounded as follows:

$$-\left(1-\frac{1}{n}\right)^{m-t} \leq \Delta_t \leq \left(1-\frac{1}{n}\right)^{m-t}$$

Setting $c_t = \left(1-\frac{1}{n}\right)^{m-t}$, we obtain that

$$\sum_{t=1}^{m} c_t^2 = \frac{1-(1-1/n)^{2m}}{1-(1-1/n)^2} = \frac{n^2-\mu^2}{2n-1}.$$

Plugging this into Azuma's inequality with $\lambda = \theta\mu$ gives the desired result.

2.2 The Chernoff Bound for the Occupancy Problem

To prove Theorem 3, we will derive a Chernoff-type bound on the number of empty bins Z when $m = rn$ balls are thrown independently at random into n bins. Notice that in this setting the emptiness of the bins is not independent, thereby blocking any attempts at a direct application of the Chernoff bound. However, we will show that the correlations are such that the Chernoff bound from the independent case is still indirectly applicable.

Definition 2 *Let I_i be the indicator variable whose value is 1 if and only if bin i is empty. Let $J_i = 1 - I_i$ be the indicator variable for the bin i being non-empty. Define $Z = \sum I_i$ as the number of empty bins and $X = \sum J_i = n - Z$ as the number of non-empty bins.*

The problem is that the Bernoulli random variables I_i, for $1 \leq i \leq n$, are not independent. For the purposes of our analysis, we create corresponding *independent* Bernoulli random variables.

Definition 3 *Define*

$$p = \mathbf{E}[I_i] = \left(1-\frac{1}{n}\right)^m.$$

Let I'_i be mutually independent Bernoulli random variables with probability p of evaluating to 1, and observe that the sum $Y = \sum I'_i \sim B(n,p)$. Similarly define $J'_i = 1 - I'_i$ and $W = \sum J'_i$.

The following lemma relates the moment generating function for the random variables Y and Z.

Lemma 1 *For all real t, $\mathbf{E}\left[e^{tZ}\right] \leq \mathbf{E}\left[e^{tY}\right]$.*

Proof: First we consider the easier case where $t \geq 0$. By a Taylor series expansion of the exponential, and using the symmetry of the indices, it suffices to show that

$$\mathbf{E}\left[I_1 \cdots I_k\right] \leq \mathbf{E}\left[I'_1 \cdots I'_k\right]$$

for $k = 1, \ldots, n$. But we can show that

$$\mathbf{E}\left[I_1 \cdots I_k\right] = \left(1 - \frac{k}{n}\right)^m \leq \left(1 - \frac{1}{n}\right)^{km} = \mathbf{E}\left[I'_1 \cdots I'_k\right]$$

and the result follows immediately.

Now consider the case where $t < 0$, and let $t = -s$. We now claim that $\mathbf{E}\left[e^{-ns+sX}\right] \leq \mathbf{E}\left[e^{-ns+sW}\right]$, and this reduces to $\mathbf{E}\left[e^{sX}\right] \leq \mathbf{E}\left[e^{sW}\right]$. Again by the Taylor series expansion it suffices to show that

$$\mathbf{E}\left[J_1 \cdots J_k\right] \leq \mathbf{E}\left[J'_1 \cdots J'_k\right]$$

for $k = 1, \ldots, n$. By an application of the FKG inequality [AS92] it follows that

$$\mathbf{P}\left[J_{i+1} = 1 | J_1 = 1, \ldots J_i = 1\right] \leq \mathbf{P}\left[J_{i+1} = 1\right],$$

i.e. the probability that bin $i+1$ is non-empty given that bins $1, \ldots, i$ are non-empty is bounded from above by the unconditional probability of bin $i+1$ being non-empty. The desired result can now be obtained by a straight-forward argument. ■

Recalling the structure of the argument used in Chernoff bounds in general, as well as the particular bounds for the binomial distribution [AS92, MR94], it can be verified that the lemma implies the following corollary, which in turn implies Theorem 3.

Corollary 1 *A Chernoff bound on either tail of the distribution of Y applies to Z as well. In particular, for $\theta > -1$ and $\mu = np$ we have*

$$H(m, n, (1+\theta)\mu) \leq \exp(-((1+\theta)\ln[1+\theta] - \theta)\mu).$$

The emptiness of bins are negatively correlated and so it ought to be the case that the distribution is tighter around the expected value than in the (hypothetical) case where the emptiness of bins is independent. However, we did not prove a dominance relationship for the actual tail probabilities; instead, the result of this section deals only with the Chernoff bounds on the tail. This is precisely why the current tail bound is weaker than Theorem 4 which is a tight bound but with a more complicated closed form.

2.3 The Large Deviations Bound for the Occupancy Problem

We shall derive here the large deviations result, as stated in Theorem 4, bounding the tail of the distribution of the number of empty bins when $m = rn$ balls (where $r > 0$ is a constant) are thrown independently at random into n bins.

Definition 4 *The stochastic process $\mathcal{M}$ is the sequence of states, $X_0, X_1, \ldots, X_m$, where nX_i denotes the number of non-empty bins after i balls have been thrown at random and independently into n bins.*

The process $\mathcal{M}$ is a time-dependent Markov process. Initially the bins are all empty and hence we have $X_0 = 0$.

Definition 5 *The increments of the process $\mathcal{M}$ represent the change in the number of non-empty bins and are defined as follows, for $i = 0, \ldots, m-1$, given that M is in state X_i,*

$$\Delta X_i = X_{i+1} - X_i = \begin{cases} 0 & \text{with probability } X_i \\ 1/n & \text{with probability } 1 - X_i \end{cases}$$

Definition 6 *The moment generating function for ΔX_i is defined by*

$$M_{X_i}(\theta) = \mathbf{E}\left[e^{n\theta\Delta X_i} | X_i\right] = \left((1 - X_i)\, e^{\theta} + X_i\right).$$

Definition 7 *We define the function*

$$l(x, y) = \sup_{\theta > 0} \left(\theta y - \ln M_x(\theta)\right)$$

The probability that the number of non-empty bins increases by a fixed amount upon the addition of a single ball can be bounded as follows.

Lemma 2 *The probability distribution of the increments in the number of non-empty bins satisfies the inequality*

$$\mathbf{P}\left[\Delta X_i = \frac{y}{n}\right] \leq e^{-l(X_i, y)}.$$

Proof: Using a Chernoff-type argument we obtain that for all $\theta > 0$,

$$\mathbf{P}\left[\Delta X_i = \frac{y}{n} | X_i\right] = \mathbf{P}\left[e^{n\theta\Delta X_i} = e^{\theta y} | X_i\right] \leq \frac{\mathbf{E}\left[e^{n\theta\Delta X_i} | X_i\right]}{e^{\theta y}} = \exp(-(\theta y - \ln M_{X_i}(\theta))).$$

This implies the desired result. ■

The following technical lemma now allows us to obtain a handle on the behavior of the function $l(x, y)$.

Lemma 3 *For $0 < y < 1$, eliminating the supremum from the definition of $l(x, y)$, we get,*

$$l(x, y) = y \ln\left(\frac{yx}{(1-x)(1-y)}\right) - \ln\left(\frac{x}{1-y}\right)$$

and the supremum is obtained at the point

$$\theta_{x,y} = \ln\left(\frac{yx}{(1-x)(1-y)}\right)$$

Proof: Differentiating the expression inside the supremum (in Definition 7) with respect to θ, and setting the derivative to 0, gives us

$$e^{\theta} = \frac{yx}{(1-x)(1-y)}.$$

Substituting this value of θ in the expression for $l(x, y)$ and using the convexity of $\ln M_x(\theta)$ (see Shwartz and Weiss [SW94]), gives us the desired result. ■

Definition 8 *A path $\gamma = (x_0, x_1, \ldots, x_m)$ of the Markov process $\mathcal{M}$ is an instantiation of its states where $X_0 = x_0$, $X_1 = x_1$, …, $X_m = x_m$.*

Given that $X_0 = 0$, the path γ can be uniquely determined by the increments $\Delta x_i = x_{i+1} - x_i$. Moreover the probability of a path can be bounded using the probability of the increments as shown in the following lemma.

Lemma 4 *The probability of a path $\gamma = (x_0, x_1, \ldots, x_m)$ is bounded as*

$$\mathbf{P}[\gamma] \leq \exp\left(-\sum_{i=0}^{m-1} l(x_i, n\Delta x_i)\right).$$

Proof: A path is uniquely determined by the increments, implying that

$$\mathbf{P}[\gamma] = \mathbf{P}[\Delta X_0 = \Delta x_0, \Delta X_1 = \Delta x_1, \ldots, \Delta X_{m-1} = \Delta x_{m-1}].$$

By the Markov property, we can show that

$$\mathbf{P}[\gamma] = \prod_{i=0}^{m-1} \mathbf{P}[\Delta X_i = \Delta x_i | X_i = x_i] \leq \exp\left(-\sum_{i=0}^{m-1} l(x_i, n\Delta x_i)\right),$$

which is the desired result. ■

Definition 9 *A path γ of the Markov chain $\mathcal{M}$ can be approximated by an absolutely continuous scaled process $x_n(t)$ which is defined such that for $nt = 0, 1, \ldots, m$, we have $x_n(t) = X_{nt}$,*

Definition 10 *Let Γ_x be the set of absolutely continuous functions $x(t)$ with derivative $\frac{\mathrm{d}x}{\mathrm{d}t} = x'(t)$, satisfying the boundary conditions $x(0) = 0$ and $x(r) = x$.*

Definition 11 *The function, $I(x)$, is defined as corresponding to the maximum probability path, as follows,*

$$I(x) = \inf_{x(t)\in\Gamma_x} \int_0^r l(x(t), x'(t))\mathrm{d}t.$$

Notice that from Lemma 4 one might expect that $e^{-nI(x)}$ would give an upper bound on the probability of the paths in Γ_x. We now state a theorem due to Azencott-Ruget [AR77](Theorem 6.3) which gives us the following surprising result: in case the final value $X_m = x$ has a large deviation from its expectation, then the upper bound on the probability of the paths in Γ_x provides a good upper bound on the probability of having $X_m = x$.

Theorem 6 ([AR77]) *Let μ be the expected value of X_m, and let $x = (1+\delta)\mu$ where $\delta > 0$ (i.e. we have a large deviation from the expectation). Then, for Γ_x and $I(x)$ as defined above,*

$$\lim_{n\to\infty} \frac{1}{n} \log P[X_m \geq x] = -\inf_{u\geq x} I(u).$$

The next two lemmas will be devoted to the computation of $I(x)$. The following lemma characterizes the function that gives the upper bound on the probability of paths in Γ_x.

Lemma 5 *Let $\xi(t)$ be the function in Γ_x that minimizes $I(x)$. Then, $\xi(t)$ obeys the Euler differential equation.*

$$\frac{\partial l}{\partial \xi} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial l}{\partial \xi'} = 0.$$

Proof: This is a standard well-known result in variational calculus. For a detailed proof please refer to Shwartz and Weiss [SW94]). ■

We can now determine $I(x)$ provided x is expressed suitably.

Lemma 6 *If $x = k(1 - e^{-r/k})$, where $k > 1$ is a constant that depends on x, then*

$$I(x) = \int_0^x [\ln(k - \xi) - \ln(1 - \xi)]\,\mathrm{d}\xi - r \ln k$$

Proof: Integrating the result of Lemma 5, gives us the Dubois-Raymond equation, $l - \xi' \frac{\partial l}{\partial \xi'} =$ constant, which on simplification gives for some constant k,

$$\frac{\xi}{1 - \xi'} = k. \tag{6}$$

Solving this differential equation gives the solution $\xi(t) = k(1 - e^{-t/k})$ and imposing the boundary conditions $\xi(0) = 0$ and $\xi(r) = x$ implicitly defines k, as follows

$$x = k(1 - e^{-r/k})$$

Finally, using the differential equation (6)and substituting the value of $l(\xi, \xi')$ as given in Lemma 3, we get

$$\begin{aligned} I(x) &= \int_0^r l(\xi, \xi')\mathrm{d}t \\ &= \int_0^r \left[\xi' \ln\left(\frac{\xi'\xi}{(1-\xi)(1-\xi')}\right) - \ln\left(\frac{\xi}{1-\xi'}\right)\right] \mathrm{d}t \\ &= \int_0^x \ln\left(\frac{k-\xi}{1-\xi}\right) \mathrm{d}\xi - r\ln k. \end{aligned}$$

■

The estimate applies whenever

$$x \in (1 - e^{-r}, \min(1, r))$$

If $x \in (0, 1 - e^{-r})$ then the estimate gives the approximate probability that $X_m \leq x$. Observing that the number of empty bins $z = n(1 - x)$, the occupancy bound in Theorem 4 follows by combining the results of Theorem 6 and Lemma 6.

3 Threshold for Satisfiability

We now turn our attention to the problem of determining bounds on the threshold for the satisfiability of 3-SAT formulas. We will need some additional notation concerning random 3-SAT formulas. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be the set of variables and $\mathcal{C} = \{C_1, \ldots, C_m\}$ be the set of clauses where $m = cn$. For convenience, we sometimes view a clause C as being a set of exactly three literals, no two of which are derived from the same variable. We use the notation $Z = (z_1, \ldots, z_n) \in \{0, 1\}^n$ to indicate a particular truth assignment to the variables. Setting $z_i = 1$ corresponds to assigning a value "TRUE" to variable X_i and similarly $z_i = 0$ indicates a "FALSE" value.

The clauses in a 3-SAT formula can be classified into four distinct types:

type 0: no literals are negated, e.g. the clause $X_1 \vee X_2 \vee X_3$,

type 1: exactly one literal is negated, e.g. the clause $\bar{X}_1 \vee X_2 \vee X_3$,

type 2: exactly two literals are negated, e.g. the clause $\bar{X}_1 \vee \bar{X}_2 \vee X_3$,

type 3: all three literals are negated, e.g. the clause $\bar{X}_1 \vee \bar{X}_2 \vee \bar{X}_3$.

Let τ_i indicate the event that a clause chosen at random is of type i, for $i \in \{0, 1, 2, 3\}$. Clearly in our model

$$\mathbf{P}[\tau_0] = \mathbf{P}[\tau_3] = 1/8$$

and

$$\mathbf{P}[\tau_1] = \mathbf{P}[\tau_2] = 3/8.$$

We would like to focus upon the set of formulas in $\Omega(n, c)$ that are satisfiable by an arbitrary truth assignment Z. It is easy to see that there is no loss of generality in considering only the set of formulas satisfiable by the particular truth assignment $Z_1 = (1, \ldots, 1)$. (We could just consistently negate all literals derived from variables which were assigned 0 by Z.) Consider a formula $F_1 \in \Omega(n, c)$ which is satisfiable by Z_1, i.e. $F_1(Z_1)$ is true. We will show that with high probability it has a large number of satisfying truth assignments.

A formula F_1 satisfiable by Z_1 does not have any clauses of type 3. This conditions the distribution of the clauses among the various types. Let $\mathbf{P}[\tau_i \mid Z_1]$ denote the probability that a clause satisfiable by Z_1 is of type i. Clearly,

$$\begin{aligned}
\mathbf{P}[\tau_0 \mid Z_1] &= \frac{1}{7}, \\
\mathbf{P}[\tau_1 \mid Z_1] &= \frac{3}{7}, \\
\mathbf{P}[\tau_2 \mid Z_1] &= \frac{3}{7}, \\
\mathbf{P}[\tau_3 \mid Z_1] &= 0.
\end{aligned}$$

Finally, we need the following definition in our analysis.

Definition 12 *A variable X_i is said to* **cover** *a clause C_j if X_i occurs unnegated in clause C_j, i.e. $X_i \in C_j$. A set of variables $\mathcal{V} \subseteq \mathcal{X}$ is said to* **cover** *a set of clauses $\mathcal{B} \subseteq \mathcal{C}$ if for each $C_j \in \mathcal{B}$ there exists a variable $X_i \in \mathcal{V}$ such that X_i covers C_j.*

Observe that any truth assignment which assigns "TRUE" to the variables in $\mathcal{V}$, must satisfy all the clauses in $\mathcal{B}$.

3.1 The Proof Outline

In this section we outline the proof which will be given in detail in the subsequent sections. To get the intuition behind our approach, consider first the rather straightforward proof of the bound of 5.19 on the threshold. The idea here is to use the first moment method. Consider a random 3-SAT formula F chosen from the above probability space. Let $\#F$ denote the number of satisfying truth assignments for F. We claim that

$$\mathbf{E}[\#F] = 2^n \left(\frac{7}{8}\right)^{cn}$$

because given a truth assignment, it satisfies $(7/8)^{cn}$ of all formulas (which is seen by first choosing the variables in the formula and then considering all possible choices of negations) and there are a total of 2^n distinct truth assignments. Using the Markov inequality, we obtain that

$$\mathbf{P}[\#F > 0] \leq \mathbf{E}[\#F] = \left(2\left(\frac{7}{8}\right)^c\right)^n.$$

It is easy to see that this probability is exponentially small when $c > \log_{8/7} 2 = 5.19089\ldots$.

The interesting point here is that in the range $4.2 < c < 5.19$ a random formula is expected to have a large number of satisfying truth assignments, and yet it is (empirically) known that a random formula is unlikely to have *any* satisfying truth assignments. This is exactly the reason why the first moment method cannot give a better upper bound on the threshold. Our proof is based on exploiting the only possible explanation behind these seemingly contradictory statements: there are a very few formulas with a tremendously large number of satisfying truth assignments, and the overwhelming majority of the formulas have either few or no satisfying truth assignments. Our goal then is to identify the formulas which have many more satisfying truth assignments than the expected value. In fact, we will consider a random formula which is satisfiable by Z_1 and show that with high probability there exists a small set of variables which cover all its clauses; then, it follows that assigning these variables "TRUE" and trying all possible assignments to the rest gives a large number of satisfying truth assignments for the formula under consideration. Finally, we devise a variant of the first moment argument which uses this structural property to prove the desired result.

The following is an outline of the argument which establishes the existence of formulas with a large number of satisfying truth assignments. In the following, we are focusing upon a random formula F_1 which is satisfiable by Z_1, i.e. it is chosen from our probability space but its distribution is then conditioned by the event that Z_1 satisfies it. Denote by $\mathcal{C}$ the set of clauses in F_1. Let $\mathcal{C}_i \subseteq \mathcal{C}$ be the subset of clauses in F_1 of type i and define the random variable $m_i = |\mathcal{C}_i|$, for $i \in \{0, 1, 2, 3\}$. We shall try to determine the smallest cover of $\mathcal{C}$ in the following 4 steps.

First, we determine the expected value of m_2 and show that it is sharply concentrated around this value. Since each clause in $\mathcal{C}_2$ has one unnegated variable, the set of all such variables in the clauses of $\mathcal{C}_2$ form a minimal cover of $\mathcal{C}_2$. Let $\mathcal{S}$ denote the set of variables which *do not* appear unnegated in $\mathcal{C}_2$. Then, assigning "TRUE" to the variables in $\mathcal{X} \setminus \mathcal{S}$ satisfies the clauses of $\mathcal{C}_2$, and the variables in $\mathcal{S}$ can be set freely to handle the remaining clauses. A combinatorial argument establishes sharp bounds on the size of $\mathcal{S}$. The remaining clauses in F_1 are those in $\mathcal{C}_0$ and $\mathcal{C}_1$, but some of these are already satisfied (or covered) by assigning "true" to the variables in $\mathcal{X} \setminus \mathcal{S}$. Let $\mathcal{U}$ denote the clauses in $\mathcal{C}_0 \cup \mathcal{C}_1$ all of whose unnegated variable are in the set $\mathcal{S}$, clearly these are the only clauses which need to be covered by the variables in $\mathcal{S}$. We show that the number of these clauses is small (with high probability), and therefore only a few of the variables in $\mathcal{S}$ suffice to form a covering set for them. The remaining variables in $\mathcal{S}$, called the independent set $\mathcal{I}$, have the property that $\mathcal{X} \setminus \mathcal{I}$ forms a cover for *all* clauses in F_1. Thus, assigning "TRUE" to the variables in $\mathcal{X} \setminus \mathcal{I}$ ensures that F_1 is satisfied, and the remaining variables in $\mathcal{I}$ can be assigned freely to generate a total of $2^{|\mathcal{I}|}$ satisfying truth assignments for F_1. Our analysis will establish that $|\mathcal{I}|$ is large with high probability.

3.2 From One to Many Assignments

We now flesh out the details of the proof outlined above, using the following sequence of lemmas. Consider a random formula F_1 satisfiable by Z_1, and denote by $\mu_2 = \mathbf{E}[m_2]$ the expected number of type 2 clauses in F_1. Since each of the clauses in F_1 is of type 2 with probability 3/7, we obtain that

$$\mu_2 = \frac{3cn}{7}.$$

Consider the function $t(\alpha) = (1 + \alpha)\mu_2$, for $\alpha \in [-1, 4/3]$. We define the probability function

$$P(\alpha) = \mathbf{P}[m_2 = t(\alpha)].$$

We would like to obtain an upper bound on this probability function. Observe that the number of clauses in $\mathcal{C}_2$ is a binomial random variable with expectation μ_2 and hence we obtain the following lemma. The result follows from the linearity of expectation.

Lemma 7 *For $-1 \leq \alpha \leq 4/3$ and $t(\alpha) = (1+\alpha)\mu_2$,*

$$P(\alpha) = F(cn, 3/7, t(\alpha)).$$

We now determine the size of a minimal covering set for the clauses in $\mathcal{C}_2$. Recall that each clause in $\mathcal{C}_2$ contains exactly one unnegated variable. Let $\mathcal{S}$ be the set of variables which do not appear unnegated in any of the clauses in this set. Then, $\mathcal{X} \setminus \mathcal{S}$ is a covering set for $\mathcal{C}_2$. Let the conditional expectation of $|S|$, given that $m_2 = t(\alpha)$, be

$$\mu_{\mathcal{S}}(\alpha) = \mathbf{E}\left[|\mathcal{S}| \mid m_2 = t(\alpha)\right].$$

Define the function

$$s(\alpha, \beta) = (1+\beta)\mu_{\mathcal{S}}(\alpha),$$

where $-1 \leq \beta \leq n/\mu_{\mathcal{S}}(\alpha) - 1$, and the probability function

$$P(\beta \mid \alpha) = \mathbf{P}\left[|\mathcal{S}| = s(\alpha, \beta) \mid m_2 = t(\alpha)\right].$$

Lemma 8 *Suppose that $m_2 = t(\alpha)$. Then,*

$$\mu_{\mathcal{S}}(\alpha) \sim ne^{-t(\alpha)/n}$$

and

$$P(\beta \mid \alpha) = H(t(\alpha), n, s(\alpha, \beta)).$$

Proof: Since the clauses in $\mathcal{C}_2$ have exactly one unnegated variable each, the minimal cover of $\mathcal{C}_2$ consists of all variables that occur unnegated in $\mathcal{C}_2$. We can look at this problem as an occupancy problem in which each clause of $\mathcal{C}_2$ corresponds to a ball and the variables in $\mathcal{X}$ correspond to bins. Hence, the number of variables in the minimal cover of $\mathcal{C}_2$ corresponds to the nonempty bins when $t(\alpha)$ balls are thrown at random into n bins; the set $\mathcal{S}$ corresponds to the empty bins. It follows that the expected size of $\mathcal{S}$ given the size of $\mathcal{C}_2$ is

$$\mu_{\mathcal{S}}(\alpha) \sim ne^{-t(\alpha)/n},$$

and

$$P(\beta \mid \alpha) = H(t(\alpha), n, s(\alpha, \beta)).$$

■

Next, we turn to the clauses $\mathcal{U} \subseteq \mathcal{C}_0 \cup \mathcal{C}_1$ which are not covered by the above covering set for $\mathcal{C}_2$, i.e. those clauses whose unnegated variables are drawn only from the set $\mathcal{S}$. Assume that the size of $\mathcal{C}_2$ is $t(\alpha)$ and that of $\mathcal{S}$ is $s(\alpha, \beta)$, for some α and β, and define the conditional expectation of $|\mathcal{U}|$ as follows.

$$\mu_{\mathcal{U}}(\alpha, \beta) = \mathbf{E}\left[|\mathcal{U}| \mid |\mathcal{S}| = s(\alpha, \beta) \wedge m_2 = t(\alpha)\right].$$

Define the function

$$u(\alpha, \beta, \gamma) = (1+\gamma)\mu_{\mathcal{U}}(\alpha, \beta),$$

where $\gamma \in \left[-1, \frac{cn - t(\alpha)}{\mu_{\mathcal{U}}(\alpha,\beta)} - 1\right]$, and the probability function

$$P(\gamma \mid \alpha, \beta) = \mathbf{P}\left[|\mathcal{U}| = u(\alpha, \beta, \gamma) \mid m_2 = t(\alpha) \wedge |\mathcal{S}| = s(\alpha, \beta)\right].$$

Lemma 9 *Suppose that $m_2 = t(\alpha)$ and $|\mathcal{S}| = s(\alpha,\beta)$. Then,*

$$\mu_{\mathcal{U}}(\alpha,\beta) \sim (cn - t(\alpha))\left(\frac{3}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^2 + \frac{1}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^3\right)$$

and

$$P(\gamma \mid \alpha,\beta) \asymp F\left(cn - t(\alpha), \frac{3}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^2 + \frac{1}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^3, u(\alpha,\beta,\gamma)\right).$$

Proof: The expected size of $\mathcal{U}$ is obtained by taking the $cn - m_2$ clauses of types 0 and 1, and determining for each one of them whether all their unnegated variables belong to $\mathcal{S}$. Notice that a clause which is not in $\mathcal{C}_2$ belongs to $\mathcal{C}_0$ with probability 1/4 and to $\mathcal{C}_1$ with probability 3/4.

$$\begin{aligned}
\mu_{\mathcal{U}}(\alpha,\beta) &= \mathbf{E}\left[|\mathcal{U}| \mid |\mathcal{S}| = s(\alpha,\beta) \wedge m_2 = t(\alpha)\right] \\
&= (cn - t(\alpha))\left(\frac{3}{4}\frac{\binom{s(\alpha,\beta)}{2}}{\binom{n}{2}} + \frac{1}{4}\frac{\binom{s(\alpha,\beta)}{3}}{\binom{n}{3}}\right) \\
&\sim (cn - t(\alpha))\left(\frac{3}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^2 + \frac{1}{4}\left(\frac{s(\alpha,\beta)}{n}\right)^3\right)
\end{aligned}$$

The probability bound is the Binomial Bound. ■

Finally, we are left with the situation where all the clauses of F_1 except those in $\mathcal{U}$ are covered by the variables in $\mathcal{X} \setminus \mathcal{S}$. We now determine an upper bound on the number of variables in $\mathcal{S}$ required to cover the clauses of $\mathcal{U}$. In fact, we will identify a large set of *independent* variables $\mathcal{I} \subseteq \mathcal{S}$ such that the clauses of $\mathcal{U}$ have a cover in $\mathcal{S} \setminus \mathcal{I}$. This is done in two substeps. Each clause in $\mathcal{U}$ has at least two unnegated variables . Label the unnegated variables in eac clause as the first, second and (if there are three) third independently. For the first substep, always take the first variable. For the second substep, take the first variable with probability $1/s$ and the second with probability $1 - 1/s$. These two substeps will give us two independent candidates for $\mathcal{I}$. We choose $\mathcal{I}$ to be the largest of these two independent sets. Given that $|C_2| = t(\alpha)$, $|\mathcal{S}| = s(\alpha,\beta)$ and $|\mathcal{U}| = u(\alpha,\beta,\gamma)$, we define the conditional expectation of $|\mathcal{I}|$ as follows.

$$\mu_{\mathcal{I}}(\alpha,\beta,\gamma) = \mathbf{E}\left[|\mathcal{I}| \mid m_2 = t(\alpha) \wedge |\mathcal{S}| = s(\alpha,\beta) \wedge |\mathcal{U}| = u(\alpha,\beta,\gamma)\right]$$

We define the function

$$i(\alpha,\beta,\gamma,\delta) = (1+\delta)\mu_{\mathcal{I}}(\alpha,\beta,\gamma)$$

where $\delta \in \left[-1, \frac{s(\alpha,\beta)}{\mu_{\mathcal{I}}(\alpha,\beta,\gamma)} - 1\right]$, and the probability function

$$P(\delta \mid \alpha,\beta,\gamma) = \mathbf{P}[|\mathcal{I}| = i(\alpha,\beta,\gamma,\delta) \mid m_2 = t(\alpha) \wedge |\mathcal{S}| = s(\alpha,\beta) \wedge |\mathcal{U}| = u(\alpha,\beta,\gamma)].$$

Lemma 10 *Suppose that $m_2 = t(\alpha)$, $|\mathcal{S}| = s(\alpha,\beta)$ and $|\mathcal{U}| = u(\alpha,\beta,\gamma)$. Then,*

$$\mu_{\mathcal{I}}(\alpha,\beta,\gamma) \sim s(\alpha,\beta)\exp\left(-\frac{u(\alpha,\beta,\gamma)}{s(\alpha,\beta)}\right)$$

and for $\delta < 0$ we have

$$P(\delta \mid \alpha,\beta,\gamma) \leq (H(u(\alpha,\beta,\gamma), s(\alpha,\beta), i(\alpha,\beta,\gamma,\delta)))^2.$$

Proof: The current random process reduces to an occupancy problem in which the clauses of $\mathcal{U}$ correspond to balls and the variables in $\mathcal{S}$ correspond to bins. Recall that each clause in $\mathcal{U}$ chooses one unnegated variable as its covering variable, and this is equivalent to choosing a variable from $\mathcal{S}$ uniformly at random. The empty bins then correspond to a subset $\mathcal{I}$ of $\mathcal{S}$ which is not needed to cover the clauses in $\mathcal{U}$. The value of the expectation is then obvious. Since the set $\mathcal{I}$ of independent variables is taken to be the largest of the independent sets obtained in the two substeps, the probability that the size of $\mathcal{I}$ is smaller than the expectation is bounded by the product of the probabilities of getting such a small independent set in each substep. ∎

We shall call $\mathcal{I}$ the set of independent variables because the variables in $\mathcal{I}$ can be assigned any truth value and we would still have a satisfying truth assignment for the formula F_1, provided the remaining variables are always assigned "TRUE". It follows that F_1 has at least $2^{|\mathcal{I}|}$ solutions.

3.3 The Probability of (Dis)Satisfaction

Consider the matrix A with rows corresponding to the truth assignments and columns to formulas in $\Omega(n, c)$. The matrix entry (k, l) is 1 if the truth assignment Z_k satisfies formula F_l, and 0 otherwise. Let $\mathcal{T} \subseteq \Omega(n, c)$ be the subset of satisfiable formulas. Let $\mathcal{T}_k \subseteq \mathcal{T}$ be the subset of formulas satisfiable by Z_k. Let N be the number of formulas in $\Omega(n, c)$. Let $\#F_l$ be the number of satisfying truth assignments to a formula F_l.

Lemma 11 [2] *The fraction of satisfiable formulas is given by*

$$\frac{|\mathcal{T}|}{N} = 2^n \left(\frac{7}{8}\right)^{cn} \sum_{F_l \in \mathcal{T}_1} \frac{1}{\#F_l} \frac{1}{|\mathcal{T}_1|}.$$

Proof: Simply noting the following sequence of equalities leads to the desired proof.

$$|\mathcal{T}| = \sum_{F_l \in \mathcal{T}} \frac{\sum_k a_{kl}}{\#F_l} = \sum_k \sum_{F_l \in \mathcal{T}} \frac{a_{kl}}{\#F_l} = \sum_k \sum_{F_l \in \mathcal{T}_k} \frac{1}{\#F_l} = 2^n |\mathcal{T}_1| \sum_{F_l \in \mathcal{T}_1} \frac{1}{\#F_l} \frac{1}{|\mathcal{T}_1|}$$

We know that any truth assignment satisfies $(7/8)^{cn}$ fraction of the formulas, i.e. $|\mathcal{T}_1|/N = (7/8)^{cn}$. This gives us

$$\frac{|\mathcal{T}|}{N} = 2^n \left(\frac{7}{8}\right)^{cn} \sum_{F_l \in \mathcal{T}_1} \frac{1}{\#F_l} \frac{1}{|\mathcal{T}_1|}.$$

∎

Note that the sum in this lemma represents the expected value of the reciprocal of the number of satisfying truth assignments over the formulas that are satisfiable by a given assignment. Formulas that have an independent set of size at least $i(\alpha, \beta, \gamma, \delta)$ have at least $2^{i(\alpha,\beta,\gamma,\delta)}$ satisfying truth assignments and occur with probability $P(\alpha, \beta, \gamma, \delta)$, where

$$P(\alpha, \beta, \gamma, \delta) = P(\alpha) P(\beta \mid \alpha) P(\gamma \mid \alpha, \beta) P(\delta \mid \alpha, \beta, \gamma). \tag{7}$$

Since the cardinality of the sets $\mathcal{C}_2$, $\mathcal{S}$, $\mathcal{U}$, and $\mathcal{I}$ are integers bounded between 0 and $max(n, cn)$, there is only a polynomial number of tuples $(\alpha, \beta, \gamma, \delta)$ for which $P(\alpha, \beta, \gamma, \delta)$ is non-zero. Let $\mathcal{A}$ be the set of such tuples, then $|\mathcal{A}| = O(n^4)$ and we can write

$$\frac{|\mathcal{T}|}{N} \leq 2^n \left(\frac{7}{8}\right)^{cn} \sum_{\mathcal{A}} \frac{P(\alpha, \beta, \gamma, \delta)}{2^{i(\alpha,\beta,\gamma,\delta)}} = O\left(2^n \left(\frac{7}{8}\right)^{cn} n^4 \max_{\alpha,\beta,\gamma,\delta} \frac{P(\alpha, \beta, \gamma, \delta)}{2^{i(\alpha,\beta,\gamma,\delta)}}\right),$$

[2] David Aldous informs us that this is related to the *harmonic mean formula* used by him earlier to obtain improved bounds on the probability of a union of events [Ald89].

where the maximum is taken over all feasible choices of α, β, γ and δ. We can compute an upper bound on the probability $P(\alpha, \beta, \gamma, \delta)$ by using equation (7) to express it as a product of probabilities which can then be bounded using Lemmas 7, 8, 9 and 10. We define for all feasible α, β, γ and δ, an upper bound function h to be a function such that

$$h(\alpha, \beta, \gamma, \delta) \geq \frac{P(\alpha, \beta, \gamma, \delta)}{2^{i(\alpha,\beta,\gamma,\delta)}}.$$

We will get different functions h depending upon our estimate of the size of the independent set and how we choose to bound the probabilities. For $c \geq 4.762$ we can find an exponentially small upper bound on the probability of satisfaction using an appropriate function h.

Theorem 7 *For $c \geq 4.762$, the probability that a random 3-SAT formula from the space $\Omega(n, c)$ is satisfiable with a probability approaching* 0 *as n approaches infinity.*

Proof: Observe that the probability of satisfaction at $c = c^*$ is an upper bound on the probability of satisfaction for $c > c^*$. Hence to prove the theorem it suffices to prove it for a given value of c. Using equation (7) and Lemmas (7)-(10), we can express an upper bound on $P(\alpha, \beta, \gamma, \delta)$ in terms of $F(n, p, y)$ and $H(m, n, z)$. We compute an upper bound function h by substituting the bound (1) of Theorem 1 for $F(n, p, y)$, and the bound (5) of Theorem 4 for $H(m, n, z)$. The upper bound function h depends also on n and is of the type $exp(nf(\alpha, \beta, \gamma, \delta) + o(n))$. Note in particular that Theorem 4 is not an exact bound without a term $o(n)$ in the exponent and is valid only for large deviations from the mean value μ. But since the expected sizes of all the sets τ_2, $\mathcal{S}, \mathcal{U}$ and $\mathcal{I}$ are $\theta(n)$, asymptotically small deviations have very little effect on the size of these sets and hence do not affect our analysis.

The function $f(\alpha, \beta, \gamma, \delta)$ so obtained was maximized using *Mathematica* and we later verified the solution independently using a Fortran program doing an exhaustive check. The verifying Fortran program stepped through the feasible space of values for α, β, γ and δ in small steps of size ϵ. For each such grid point in the four dimensional space, it computed a (tight) upper bound on f in an ϵ-neighborhood of that point. The upper bound on f was computed by bounding (in the ϵ-neighborhood of the grid point) the simpler individual terms that comprise f. The maximum of such upper bounds over all grid points then gives us an upper bound on the value of f, which is then used to bound the probability of satisfaction.

For $c = 4.762$, we computed the maximum value h^* of h to be $(0.944324)^{n+o(n)}$, attained at $\alpha = 0.0232163$, $\beta = -0.226729$, $\gamma = 0.666931$, and $\delta = -0.013987$. Observe that $2^n(7/8)^{4.762n}h^* = (0.944324/0.94438)^{n+o(n)}$ tends to 0 as n tends to infinity, giving us the desired result. ■

A marginally better threshold of $c = 4.758$ can be obtained by using a more involved technique for computing the independent set in the final step, by considering the graph with nodes in $\mathcal{S}$ and edges corresponding to the unnegated variables occuring in clauses of $\mathcal{U}$ and then using Turan's Theorem to construct an independent set. We do not describe this technique here, since the resulting improvement in the threshold value is insignificant. Finally, for the sake of completeness, the Appendix A contains a completely analytic proof of a bound of $c = 4.87$.

Acknowledgements

We are grateful to Svante Janson and Joel Spencer for technical discussions regarding occupancy bounds in general, and for valuable help in the proof of Occupancy Bound 2. We thank Alan Weiss for his help with the proof of Occupancy Bound 3. Thanks also to David Aldous for helpful comments.

References

[Ald89] D.J. Aldous. The Harmonic Mean Formula for Probabilities of Unions: Applications to Sparse Random Graphs. *Discrete Mathematics*, vol. 76 (1989), pages 167–176.

[AS92] N. Alon and J.H. Spencer. *The Probabilistic Method.* John Wiley & Sons (1992).

[AR77] R. Azencott, and G. Ruget. Mélanges d'équations différentielles et grands écarts à la loi des grands nombres. *Z.Wahrscheinlichkeitstheorie verw. Gebeite*, vol. 38 (1977), pages 1–54.

[Azu67] K. Azuma. Weighted sums of certain dependent random variables. *Tokuku Mathematical Journal*, vol. 19 (1967), pages 357–367.

[BFU93] A. Broder, A. Frieze, and E. Upfal. On the satisfiability and maximum satisfiability of random 3-CNF formulas. *Proceedings of the Fourth ACM-SIAM Symposium on Discrete Algorithms* (1993).

[CF86] M.T. Chao and J. Franco. Probabilistic analysis of two heuristics for the 3-satisfiability problem. *SIAM Journal on Computing*, vol. 15 (1986), pages 1106–1118.

[CF90] M.T. Chao and J. Franco. Probabilistic analysis of a generalization of the unit-clause literal selection heuristics for the k satisfiable problem. *Information Science*, vol. 51 (1990), pages 289–314.

[CR92] V. Chvátal and B. Reed. Mick Gets Some (the Odds Are on His Side). *Proceedings of the 33rd IEEE Symposium on Foundations of Computer Science* (1992), pages 620–627.

[CS88] V. Chvátal and E. Szemerédi. Many hard examples for resolution. *Journal of the ACM*, vol. 35 (1988), pages 759–768.

[FP83] J. Franco and M. Paull. Probabilistic analysis of the Davis Putman procedure for solving the satisfiability problem. *Discrete Applied Mathematics*, vol. 5 (1983), pages 77–87.

[FS93] A.M. Frieze and S. Suen. Analysis of simple heuristics for random instances of 3-SAT. Manuscript (1993).

[Goe92] A. Goerdt. A threshold for unsatisfiability. *Proceedings of the 7th International Symposium on Mathematical Foundations of Computer Science* (1992).

[Gol79] A. Goldberg. Average case complexity of the satisfiability problem. *Proceedings of the 4th Workshop on Automated Deduction* (1979), pages 1–6.

[HH80] P. Hall and C.C. Heyde. *Martingale limit theory and its application.* Academic Press (1980).

[Jan94] S. Janson. Large Deviation Inequalities for Sums of Indicator Variables. Technical Report No. 34, Department of Mathematics, Uppsala University (1994).

[JK77] N.L. Johnson and S. Kotz. *Urn Models and Their Applications.* John Wiley & Sons (1977).

[KSC78] V.F. Kolchin, B.A. Sevastyanov, and V.P. Chistyakov. *Random Allocations*, John Wiley & Sons (1978).

[LT92] T. Larrabee and Y. Tsuji. Evidence for Satisfiability Threshold for Random 3CNF Formulas. Technical Report UCSC-CRL-92-42, University of California Santa Cruz (1992).

[McD89] C.J.H. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics: Invited Papers at the 12th British Combinatorial Conference* (Ed: J. Siemons), Cambridge University Press (1989), pages 148–188.

[MV93] H. El Maftouhi and W. Fernandez de la Vega. *Personal Communication* (1993).

[MSL92] D. Mitchell, B. Selman, and H. Levesque. Hard and Easy Distributions of SAT problems. *Proceedings of the 10th National Conference on Artificial Intelligence* (1992), pages 459–465.

[MR94] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, New York, 1994. In Press.

[SW94] A. Shwartz, and A. Weiss. *Large Deviations for Performance Analysis.* Chapman-Hall (1994).

[Veg92] W. Fernandez de la Vega. On random 2-SAT. Manuscript (1992).

[Wei93] A. Weiss. *Personal Communication* (1993).

A An Analytic Proof of a Weaker Bound

We have described a computational proof to compute an upper bound on the probability of satisfaction. In this section, we give an analytic proof for a slightly weaker threshold.

As mentioned earlier, we need to bound the sum defined by

$$\sigma = \sum_{\mathcal{A}} \frac{P(\alpha, \beta, \gamma, \delta)}{2^{i(\alpha,\beta,\gamma,\delta)}}$$

To simplify the analysis, we make the following changes.

1. We deterministically estimate the independent set to be of size $i \geq s - u$. This eliminates the parameter δ from the definition of the upper bound function h.
2. To further simplify the formulas, in Lemma 9 we over-estimate the size u of the set of uncovered clauses by substituting the $(s/n)^3$ term in $F(\gamma|\alpha, \beta)$ by $(s/n)^2$. Note that these modifications will only underestimate the size of i. Therefore, the thresholds obtained using these estimates will be an upper bound on the real threshold.
3. For $\alpha < 0$ the number of clauses with one unnegated variable is less than the expected value and the number of clauses with two or more unnegated variables is more than the expected value; hence, we should get independent sets of size larger than the expected value. Thus the maximum value of σ must be at a point $\alpha \geq 0$, and we only need to consider the upper bound on the function for non-negative α. Similarly, $\beta > 0$ implies fewer than expected variables are required to cover the clauses in τ_2, which in turn should lead to a larger than expected independent set. It follows that the maximum should occur for $\beta \leq 0$. These simplifications give us the following inequality.
$$\sigma \leq cn^2 \sum_{\alpha \geq 0} \sum_{\beta \leq 0} \sum_{\gamma} \frac{P(\alpha, \beta, \gamma)}{2^{i(\alpha,\beta,\gamma)}}.$$
4. To compute an upper bound on the above sum we use the weaker but simpler bounds of equation (2) in Theorem 1 and equation (4) in Theorem 3 for the Binomial and Occupancy Bounds respectively.

The upper bound function (over the range of the sum) is then given by

$$h(\alpha, \beta, \gamma) = e^{-f} \tag{8}$$

where

$$f = ((1+\alpha)\ln[1+\alpha]-\alpha)\mu_t + \beta^2\mu_s/2 + ((1+\gamma)\ln[1+\gamma]-\gamma)\mu_u + (s-u)\ln 2.$$

We obtain that

$$\sigma \leq cn^2 \sum_{\alpha\geq 0}\sum_{\beta\leq 0}\sum_{\gamma} h(\alpha,\beta,\gamma).$$

The restriction of the range of β was required to make h a valid upper bound. We shall now analytically compute a bound on the above sum.

1. Let $\gamma^*(\alpha,\beta)$ be the the point at which h attains its maximum value given α and β. Setting to 0 the derivative of the function h with respect to γ gives us the value of the optimal γ to be

$$\gamma^* = 1.$$

2. Next for a given α and for $\gamma = \gamma^*$, we determine the value of β at which f attains its maximum value. Setting the derivative of f with respect to β to zero allows us to express β^* in terms of α as follows:

$$\beta^*(\alpha) = \frac{(1-\ln 2)}{(1+2A(c-t/n)\mu_s/n)} - 1$$

where

$$A = [(1+\gamma^*)\ln(1+\gamma^*) - \gamma^*] - (1+\gamma^*)\ln 2.$$

3. We observe that

$$\sum_{\alpha\geq 0}\sum_{\beta\leq 0}\sum_{\gamma} h(\alpha,\beta,\gamma) \leq cn\sum_{\alpha\geq 0}\sum_{\beta\leq 0} h(\alpha,\beta,\gamma^*) \leq cn^2\sum_{\alpha\geq 0} h(\alpha,\beta^*(\alpha),\gamma^*)$$

4. Let $g(\alpha) = h(\alpha,\beta^*(\alpha),\gamma^*)$. It can be shown that for $c = 4.87$ the function g is decreasing for $\alpha \geq 0$. We conclude that

$$\sigma \leq c^3 n^5 h(0,\beta^*(0),1) = O(n^5(0.957759)^n).$$

Observing that $2(7/8)^{4.87}0.957759 < 1$ proves that the probability of satisfaction tends to 0 as n tends to infinity, giving us an analytic bound of 4.87 on the threshold.