ANALYSIS OF TWO SIMPLE HEURISTICS ON A RANDOM INSTANCE OF k-SAT

Alan Frieze[*]and Stephen Suen
Department of Mathematics,
Carnegie Mellon University,
Pittsburgh PA15213, U.S.A.

May 31, 1995

Abstract

We consider the performance of two algorithms, GUC and SC studied by Chao and Franco [2], [3], and Chvátal and Reed [4], when applied to a random instance ω of a boolean formula in conjunctive normal form with n variables and $\lfloor cn \rfloor$ clauses of size k each. For the case where $k = 3$, we obtain the exact limiting probability that GUC succeeds. We also consider the situation when GUC is allowed to have limited backtracking, and we improve an existing threshold for c below which almost all ω is satisfiable. For $k \geq 4$, we obtain a similar result regarding SC with limited backtracking.

1 Introduction

Given a boolean formula ω in conjunctive normal form, the *satisfiability problem* (SAT) is to determine whether there is a truth assignment that satisfies ω. Since SAT is NP-complete, one is interested in efficient heuristics that perform well "on average," or with high probability. The choice of the probabilistic space is crucial for the significance of such a study. In particular, it is easy to decide SAT in probabilistic spaces that generate formulas with large clauses [8]. To circumvent this problem, recent studies have focused on formulas with exactly k literals per clause (the k-SAT problem). Of particular interest is the case $k = 3$, since this is the minimal k for which the problem is NP-complete.

Let V_n be a set of n variables. We define a probability space $\Omega^{(k)}_{m,n}$ on the set of all $m = \lfloor cn \rfloor$ clause formulae over the variables which have exactly k literals per clause. We let $\mathcal{C}_j(V_n)$ be

[*]Supported by NSF grant CCR-9024935

the set of all clauses of size j chosen from V_n. We will assume that all variables occurring in a single clause are distinct. We then take $\Omega^{(k)}_{m,n} = \mathcal{C}_k(V_n)^m$.

This means that we consider the clauses to be ordered and we will consider the literals within clauses to be ordered too. Thus we can think of ω as a $k \times m$ array where $\omega_{i,j}$ is the i'th literal in the clause C_j. There is not a lot of difference between this model and other *unordered* models. We show later in Section 8 that our results can easily be extended to these models.

Experimental evidence [11, 13] strongly suggests that there exists a threshold γ, such that formulas are almost surely satisfiable for $c < \gamma$ and almost surely unsatisfiable for $c > \gamma$, where γ is about 4.2. This has not been proven rigorously, but such a threshold (namely c=1) is known to exist for 2-CNF formulas [7, 4].

Most practical algorithms for the satisfiability problem (such as the well-known Davis-Putnam algorithm [6]) work iteratively. At each iteration, the algorithm selects a literal and assigns it the value 1. All clauses containing this literal are erased from the formula, and the complement of the chosen literal is erased from the remaining clauses. Algorithms differ in the way they select the literal for each iteration. The following three rules are the most common ones:

1. *The unit clause rule:* If a clause contains only one literal, that literal must have the value 1;
2. *The pure literal rule:* If a formula contains a literal but does not contain its complement, this literal is assigned the value 1;
3. *The smallest clause rule:* Give value 1 to a (random) literal in a (random) smallest clause.

Broder, Frieze and Upfal [1] analysed an algorithm based entirely on the pure literal rule. They showed that in the $\Omega^{(3)}_{m,n}$ probabilistic space, the pure literal rule alone is sufficient to find, with high probability, a satisfying assignment for a random formula $\omega \in \Omega^{(3)}_{m,n}$, for $c = m/n \leq 1.63$. On the other hand, if $c > 1.7$, then the pure literal rule by itself does not suffice.

Chao and Franco [2],[3] and Chvátal and Reed [4] analysed two heuristics GUC and SC based on the small clause rule:

```
begin
  repeat
    choose a literal $x$;
    remove all clauses from $\omega$ that contain $x$ and remove $\bar{x}$ from any remaining clause;
    if a clause becomes empty - HALT, FAILURE;
  until no clauses left;
  HALT, SUCCESS
end
```

The algorithms GUC and SC differ in how the literal x is chosen. In GUC, x is chosen at random from a randomly selected clause of smallest size. SC (see Chvátal and Reed [4] for a complete description of SC) differs from GUC in that if there are no clauses of size one or two, then x is chosen at random from the set of all free literals. Since at least one clause is satisfied each time when GUC assigns a value to a variable, it is intuitively clear that GUC is likely (probabilistically) to perform better than SC. Algorithm SC however has the advantage of being simpler to analyse. The reason for this is that since SC only takes care of clauses of size one and two, there are fewer cases to consider when analysing SC.

The combined results (among other things) in Chao and Franco [2], [3] and Chvátal and Reed [4] can be summarized as follows. For 3-SAT, if $c < 2/3$ then SC succeeds with probability tending to 1 [4] and if $c < 2.99$ then the probability that UC (a variant of GUC using only the unit clause rule) succeeds does not tend to zero [2]. For k-SAT where $k \geq 4$, if

$$c < \left(\frac{k-1}{k-3}\right)^{k-3} \frac{k-1}{k-2} \frac{2^{k-3}}{k},$$

then SC succeeds with probability tending to 1 [4], and if $k \leq 40$ and

$$c < 0.7725 \left(\frac{k-1}{k-2}\right)^{k-2} \frac{2^k}{k+1}, \tag{1.1}$$

then the probability that GUC succeeds does not tend to zero [3].

Our first theorem gives the *precise* limiting probability that GUC succeeds when applied to a random instance of 3-SAT. Let $c_3 \approx 3.003$ be the solution to the equation

$$3c - 2\log c = 6 - 2\log(2/3),$$

and

$$f(x) = f_c(x) = \frac{3c}{4}(1 - x^2) + \log x, \qquad x \in (0, 1).$$

When $c < c_3$ we have $f(x) < 1$ for all $x \in (0, 1)$.

Theorem 1.1 *Consider applying GUC to a random instance of 3-SAT with n variables and $\lfloor cn \rfloor$ clauses.*
(a) Suppose that $c < 2/3$. Then

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUC succeeds}) = 1.$$

(b) Suppose that $2/3 \leq c < c_3$. Let α be the unique root of $f(x) = 0$ that is strictly less than 1. Then

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUC succeeds}) = \exp\left(-\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))}dx\right).$$

(c) If $c \geq c_3$ then

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUC succeeds}) = 0.$$

Chao and Franco [2] report that using GUC in a backtracking algorithm can be quite successful (and possibly be polynomial expected time for certain values of c). We describe (in Section 6) a modification of GUC called GUCB that allows a limited amount of backtracking when an empty clause is produced. We obtain the following result by showing that for sufficiently small c, the backtracking does not change the state of GUC by a great deal.

Theorem 1.2 *Consider GUCB when applied to a random instance of 3-*SAT *with n variables and $\lfloor cn \rfloor$ clauses. If $c < c_3$ then*

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUCB succeeds}) = 1.$$

Thus Theorem 1.2 raises the lower threshold for almost sure satisfiability from about $1.65n$ [1] to just above $3n$. On the other hand, the upper threshold giving almost sure unsatisfiability has been reduced to below $5n$ by El Maftouhi and de la Vega [12] and to about $4.758n$ by Kamath, Motwani, Palem and Spirakis [9]. Thus the current gap in our knowledge of the satisfiabiity or unsatisfiability of random ionstances of 3-SAT is still rather large.

Furthermore, even though it is very easy to prove that an instance of 3-SAT with $100n$ random clauses is almost surely unsatisfiable, there are no known polynomial time algorithms which can prove this. Chvátal and Szemerédi [5] have proved negative results on this problem.

We next turn our attention to algorithm SC. It is possible to show that the assertions in Theorems 1.1 and 1.2 hold for SC. In fact, our proof of Theorem 1.1 can be extended to obtain the precise limiting probability that SC succeeds when applied to a random instance of k-SAT. However, the more interesting question is: for what values of c will SC, with limited backtracking as in GUCB, succeed with probability close to 1? We answer this question with our next result.

Assume $k \geq 4$. Let

$$p_3(x) = \binom{k}{3} \frac{c}{2^{k-3}} x^2 (1-x)^{k-3}.$$

It is easy to see that $p_3(x)$ is unimodal, achieving a maximum of

$$\frac{2}{3} \frac{kc}{2^{k-3}} \frac{k-2}{k-1} \left(\frac{k-3}{k-1} \right)^{k-3}$$

when $x = 2/(k-1)$. For

$$c > \frac{2^{k-3}}{k} \left(\frac{k-1}{k-3} \right)^{k-3} \frac{k-1}{k-2},$$

let $0 < \beta_0 = \beta_0(c) < \beta_1 = \beta_1(c) < 1$ be the two solutions of the equation $p_3(x) = 2/3$. We prove the following theorem.

Theorem 1.3 *Suppose that $k \geq 4$. Let c_k be the maximum value of c such that*

$$\frac{1}{(k-1)(k-2)} \left(\frac{1}{\beta_0^2} + \frac{k-3}{\beta_0} - \frac{1}{\beta_1^2} - \frac{k-3}{\beta_1} \right) + \ln(\beta_0/\beta_1) \leq 1.$$

*Then when SCB is applied to a random instance of k-*SAT *with n variables and $\lfloor cn \rfloor$ clauses where $c < c_k$, we have*

$$\lim_{n\to\infty} \mathbf{Pr}(\text{SCB succeeds}) = 1.$$

Write $c_k = \eta_k 2^k/k$. It is possible to show that as $k \to \infty$, $\eta_k \to \eta^*$ where η^* can be defined similarly as η_k. Numerical calculations show that $\eta^* \approx 1.817$, $\eta_4 \approx 1.3836$, $\eta_5 \approx 1.504$, $\eta_{10} \approx 1.686$, and that η_k is increasing in k. Theorem 1.3 gives a constant c_k such that almost every formula ω with n variables and $\lfloor cn \rfloor$, with $c < c_k$, clauses of size k is satisfiable. This improves, by only a constant factor, a similar result in [4]. Also, c_k (for $4 \leq k \leq 40$) is smaller than the right hand side of (1.1), and we believe that if the limiting probability that GUC succeeds is positive, then GUC with limited backtracking (as described later) succeeds with probability $1 - o(1)$. It is thus very likely that when applied to random instances of k-SAT for $k \geq 4$, GUCB has a higher threshold of success than SCB. At present, we can only characterize the critical behaviour of GUC and GUCB, when applied to random instances of k-SAT with $k \geq 4$, using a system of $k-2$ polynomial equations whose properties we have difficulty in penetrating analytically. It seems unlikely that the exact thresholds for GUCB can be rid of the factor $1/k$ (see definition of c_k).

2 Proof Strategy

The basis of our proof of Theorem 1.1 is that the intermediate states of GUC (or SC), when applied to a random instance of k-SAT, can be represented by a Markov chain which we describe as follows. Consider GUC when applied to a formula ω chosen at random (with equal probability) from the space $\Omega^{(k)}_{m,n}$ where $m = \lfloor cn \rfloor$. Use ν to denote the number of variables whose truth values are not yet determined by GUC at an intermediate stage. We call this stage ν and so GUC starts at stage n. For the purpose of analysis, all empty clauses are assumed to be removed by GUC as soon as they are created, and GUC is allowed to run until the set of clauses is exhausted. Hence, GUC succeeds if and only if the number of empty clauses created is zero.

We will assume that ω is not given to us in its entirety at the start of the algorithm. Instead we will learn about the formula as the algorithm proceeds. This scenario has been aptly named the *method of deferred decisions* by Knuth, Motwhani and Pittel [10].

At stage ν we will have partially filled in the $k \times m$ matrix ω and there remains ν free variables. Some columns, corresponding to satisfied clauses, will be completely filled in. We will refer to these as *removed.* The remaining columns will be partially filled in. If an entry in a partially filled in column is assigned a literal, then the value of this literal has been assigned false by previous steps of the algorithm. The remaining entries will be left blank. A partially filled in column with i blank entries will correspond to a residual clause of size i, $i = 0, 1, \ldots, k$. (A clause of size 0 is an empty clause, previous assignments have assured us that GUC will fail to satisfy this clause.) Let $N_i = N_i(\nu)$, $i = 0, 1, 2, \ldots, k$ be the number of residual clauses of size i remaining at the start of stage ν of GUC.

To carry out stage ν we choose a clause C of minimum size. We randomly choose a literal x from the remaining 2ν possibilities. We assign x to one unfilled entry of C and then randomly fill in the remaining positions, subject to the condition that all variables must be distinct. We then go through the partially filled columns of ω. Suppose we have a column j with ℓ unfilled entries:

- With probability $1 - \ell/\nu$ we do nothing.
- With probability ℓ/ν we choose one of the unfilled positions of column j, position i say.
    - With probability $1/2$ place x in position i, randomly fill in the rest of column j and remove it from further consideration, as it corresponds to a satisfied clause.
    - With probability $1/2$ we place $\bar{x}$ in position i, leaving the remaining positions of column j blank.

The reader can easily convince himself (herself) that at the end of the algorithm the columns have been filled in with random clauses.

The important and now obvious property of this process is that conditional on $N_i(\nu)$, $i = 0, 1, 2, \ldots, k$ the remaining clauses are random and independent of previous steps of the algorithm. For future reference we refer to this as *complete independence*. It follows that $N = (N_0, N_1, \ldots, N_k)$ is a Markov chain.

We next write down the transition probability of N. Use $B(\tau, p)$ to denote a binomial variable with parameters τ and p and note that ν decreases by 1 at each stage. Write $\Delta N_i(\nu) = N_i(\nu - 1) - N_i(\nu)$ as the change from stage ν to stage $\nu - 1$. Then ΔN_i are binomial variables (conditional upon $N(\nu)$). We shall write down the distributions of ΔN_i under the different cases where the minimum size of the clauses is i. For $i = 1, 2, \ldots, k$, we write $\chi_i((y_0, y_1, \ldots, y_k)) = 1$ if $\min\{j \mid y_j \neq 0, 1 \leq j \leq k\} = i$, and $\chi_i((y_0, y_1, y_2, \ldots, y_k)) = 0$ if otherwise. Also, $\chi_0(y) = 0$ always. Consider the stage ν when GUC has just assigned 1 to a literal x in clause C and is about to remove clauses that contain x and all occurrences of $\bar{x}$ from other clauses. Let $\Delta_{j,0}$ be the number of clauses of size j containing literals x or $\bar{x}$ (but not including C). Let $\Delta_{j,1}$ be the number of clauses of size j containing literal $\bar{x}$ (but not x as all variables in a clause are different). It follows that conditional on $N = N(\nu)$, we have for $j = 1, 2, \ldots, k$ that

$$\begin{aligned}
\Delta_{j,0}(\nu) &= B(N_j - \chi_j(N), j/\nu), \qquad \text{in distribution,} \\
\Delta_{j,1}(\nu) &= B(\Delta_{j,0}, 1/2), \qquad \text{in distribution.}
\end{aligned}$$

Note that *complete independece* implies that the variables, $\Delta_{j,0}, j = 1, 2, \ldots, k$, are independent. Then for $j = 0, 1, \ldots, k$,

$$\Delta N_j(\nu) = \Delta_{j+1,1}(\nu) - \Delta_{j,0}(\nu) - \chi_j(N(\nu)),$$

where $\Delta_{0,0} = \Delta_{k+1,1} = 0$. Note that if $N_1(\nu) = 0$, then $\Delta N_0(\nu) = 0$ with probability 1. Note also that if $N_1(\nu) \geq 1$ is given, then a clause of size one (with literal x say) is chosen at stage ν and that $\Delta N_0(\nu)$ is distributed as a binomial variable with parameters $N_1(\nu) - 1$ and $1/(2\nu)$. Theorem 1.1(b) is obtained by showing that in the case of 3-SAT, the total number of empty clauses created is asymptotically distributed as a Poisson variable with mean $\int_\alpha^1 f(x)^2/(4x(1-f(x)))dx$. Theorem 1.1(a) and (c) are shown using monotonicity arguments.

We shall also require similar statements for SC. Let $N'_j(\nu)$ be the number of size j clauses remaining at stage ν when SC is applied to a random instance of k-SAT with n variables and m clauses. Then similary to GUC, $N'(\nu)$ is a Markov chain with initial state $N'(n) = (0, \ldots, 0, m)$ and transition probabilities given by

$$\Delta N'_j(\nu) = \begin{cases} \Delta'_{j+1,1}(\nu) - \Delta'_{j,0}(\nu) - \chi_j(N'(\nu)), & \text{if } j = 0, 1, 2, \\ \Delta'_{j+1,1}(\nu) - \Delta'_{j,0}(\nu), & \text{otherwise,} \end{cases}$$

where $\Delta'_{0,0} = \Delta'_{k+1,1} = 0$ and for $j = 1, 2, \ldots, k$

$$\begin{aligned} \Delta'_{j,0}(\nu) &= \begin{cases} B(N'_j - \chi_j(N'), j/\nu), & \text{if } j = 0, 1, 2, \\ B(N'_j, j/\nu), & \text{otherwise,} \end{cases} \\ \Delta'_{j,1}(\nu) &= B(\Delta'_{j,0}, 1/2). \end{aligned}$$

Also, conditional on $N'_1(\nu)$, the distribution of the number of empty clauses created at stage ν is binomial with parameters $(N_1(\nu) - 1)^+$ and $1/(2\nu)$.

The layout of this paper is as follows. We concentrate on showing Theorems 1.1 and 1.2, while we shall only sketch our proof of Theorem 1.3. In the next section, we collect some useful properties of a Markov chain X_t which will be used to approximate N_1 in proving Theorem 1.1(b). We shall then prove parts (a) and (c) of Theorem 1.1 in Section 4 by developing monotonicity arguments for comparing different Markov chains. Theorem 1.1(b) is proved in Section 5 by applying the results stated in Section 3. In Section 6, we describe how GUC is allowed to backtrack, and prove Theorem 1.2. In Section 7, we sketch briefly how our proof of Theorem 1.2 can be extended to proving Theorem 1.3. Section 8 briefly discusses other models.

3 A Markov chain

Use $B(m, p)$ to denote a binomial variable with parameters m and p, and write $b_j = b_j(m, p)$ for the probability that $B(m, p)$ equals j. We assume throughout this section that $mp \leq \lambda^* < 1$. The big O terms in this section are uniform in m and p (but may depend on λ^*). We consider a Markov chain X_t with transition probabilities defined as follows. If $X_t = 0$, then $\Delta X_t = X_{t+1} - X_t$ equals $B(m, p)$ in distribution; otherwise ΔX_t equals $B(m, p) - 1$ in distribution. We assume $X_0 \geq 0$ and so $X = 0$ is a reflecting barrier. As we are interested in bounds that are uniform in m and p, we need to consider a Markov chain Y_t which is similar to X_t except that in the one-step transitions of Y_t, we have a Poisson variable $P(\lambda)$ in place of $B(m, p)$. It will be clear that the two chains X_t and Y_t are very similar when $mp = \lambda$, although it is not possible to couple them so that $X_0 = Y_0$ and $X_t \leq Y_t$ for all $t \geq 0$. We let $\lambda = mp$ in this section.

We first prove the existence of a steady state distribution denoted by π for our walk. The following existence proof was kindly provided by Boris Pittel.

Let T_i, $i > 0$ denote the expected number of steps to visit the state 0 if the walk starts at i. Then $T_i = \lim_{n\to\infty} T_i^{(n)}$, where $T_0^{(n)} = T_i^{(0)} = 0$ and for $n \geq 1$ and $i \geq 1$,

$$T_i^{(n)} = 1 + \mathbf{E}[T^{(n-1)}_{i-1+B(m,p)}]$$

is the expected value of $\min\{n, \text{time to reach 0 from } i\}$.

Now, if $mp < 1$ then $\overline{T}_i = \frac{i}{1-mp}$ satisfies

$$\overline{T}_i = 1 + \mathbf{E}[\overline{T}_{i-1+B(m,p)}],$$

so by induction $T_i^{(n)} \leq \overline{T}_i$, and consequently $T_i \leq \overline{T}_i$. Thus, T_0 the expected time of *return* to zero is at most

$$\begin{aligned} 1 + \sum_{j>0} \mathbf{Pr}(B(m,p) = j) \cdot \overline{T}_j &= 1 + \frac{1}{1-mp}\mathbf{E}[B(m,p)] \\ &= 1 + \frac{mp}{1-mp} = \frac{1}{1-mp} < \infty. \end{aligned}$$

Thus the stationary distribution $\{\pi_i\}$ exists and $\pi_0 = 1/T_0 \geq 1 - mp$. (Note that $\pi_0 = 1 - mp$ from (3.1) below, and so $T_i = \bar{T}_i, i \geq 1$.)

Note next that π satisfies

$$\pi_i = \pi_0 b_i + \sum_{j=1}^{i+1} \pi_j b_{i-j+1}, \quad \forall i \geq 0.$$

Writing $G_X(s) = \sum_{i=0}^{\infty} s^i \pi_i$ as the probability generating function of the steady state distribution, it follows from the above equations that

$$\begin{aligned} G_X(s) &= \pi_0 \sum_{i\geq 0} s^i b_i + \sum_{j\geq 1} \pi_j s^{j-1} \sum_{i\geq 0} b_i s^i \\ &= \pi_0(1-p+ps)^m + \frac{1}{s}(G_X(s) - \pi_0)(1-p+ps)^m, \end{aligned}$$

giving

$$G_X(s) = \frac{\pi_0(s-1)}{s(1-p+ps)^{-m} - 1}.$$

As $G_X(1) = 1$, we actually have

$$\pi_0 = 1 - mp \tag{3.1}$$

and

$$G_X(s) = \frac{(s-1)(1-mp)}{s(1-p+ps)^{-m} - 1}. \tag{3.2}$$

Since $(1-p+ps)^m \leq \exp(-\lambda + \lambda s)$ for all s, we see that

$$G_X(s) \leq G(s) = \frac{(s-1)(1-\lambda)}{s\exp(\lambda - \lambda s) - 1}, \tag{3.3}$$

for all s between 1 and the radius of convergence of G. (It can be checked that $G(s)$ is the probability generating function of the steady state distribution of Y_t.) Since $\lambda \leq \lambda^*$, $G(s)$ exists for all $s < r_1^*$, where $r_1^* > 1$ is a constant depending on λ^* only. (r_1^* is in fact the unique root bigger than 1 of $s\exp(\lambda^* - \lambda^* s) = 1$.) Thus, (3.3) holds for all s satisfying $1 < s < r_1^*$. Note also that from (3.2), the mean of the steady state distribution of X_t is

$$\mu = \mu(m,p) = \sum_{i \geq 0} i\pi_i = \frac{mp(2-p-mp)}{2(1-mp)}. \tag{3.4}$$

Also,

$$\mathbf{Pr}(X_t = 0) = G_X(0) = 1 - mp. \tag{3.5}$$

We would like to consider the number of times that X_t returns to 0 in a certain time period. To do this, we need to collect some preliminary results. Suppose $X_0 = 1$. Let H_X be the time elapsed when X_t first hits 0. (H is defined accordingly for Y_t with $Y_0 = 1$.) Note that

$$H_X = 1 + L_1 + \ldots + L_B \text{ in distribution,}$$

where $B = B(m,p)$ in distribution and $L_1, \ldots, L_B$ are independent copies of H_X. This last equation follows from the fact that if the first step of the walk jumps to state B, it takes B independent copies of H_X for the walk to get back to the origin because all moves of the walk toward the origin have magnitude 1. Hence, writing $M_X(\theta) = \mathbf{E}[\exp(\theta H_X)]$, we have

$$M_X(\theta) = \mathrm{e}^\theta (1 - p + pM_X(\theta))^m. \tag{3.6}$$

By considering the functions $f_1(y) = \mathrm{e}^\theta(1-p+py)^m$, $f_2(y) = \exp(\theta - \lambda + \lambda y)$, $f_3(y) = \exp(\theta - \lambda^* + \lambda^* y)$ and $f(y) = y$, and by noting that $f_1(y) \leq f_2(y)$ for all θ and y and that $f_2(y) \leq f_3(y)$ for all θ and $y \geq 1$, we have

$$M_X(\theta) \leq M(\theta) \leq M^*(\theta), \tag{3.7}$$

where the first inequality holds for all $\theta < r_2^*$ and the second inequality holds for $0 \leq \theta < r_2^*$, and r_2^* is the radius of convergence of $M^*(\theta)$, and $M(\theta)$ and $M^*(\theta)$ respectively are the smallest roots of

$$\begin{aligned} M(\theta) &= \exp(\theta - \lambda + \lambda M(\theta)), & (3.8) \\ M^*(\theta) &= \exp(\theta - \lambda^* + \lambda^* M^*(\theta)). & (3.9) \end{aligned}$$

(Again, it can be checked that M is the moment generating function for H.) By observing that r_2^* is the value of θ at which the line $f(y) = y$ is a tangent to the curve $f(y) = \exp(\theta - \lambda^* + \lambda^* y)$, we find that $r_2^* = \lambda^* - \log \lambda^* - 1$. Further, by considering θ close to r_2^*, we see that $\lambda^* M^*(\theta) < 1$. Also, we shall need to bound $M''(\theta) = \frac{d^2M}{d\theta^2}$. From (3.8), we have

$$\begin{aligned} M'(\theta) &= M(\theta)/(1 - \lambda M(\theta)) \\ M''(\theta) &= M(\theta)/(1 - \lambda M(\theta))^3. \end{aligned}$$

Using the fact that $\lambda M^*(\theta) \leq \lambda^* M^*(\theta) < 1$, it follows from the second inequality in (3.7) that for $0 \leq \theta < r_2^*$,

$$M''(\theta) \leq \frac{M^*(\theta)}{(1-\lambda^* M^*(\theta))^3}.$$

Also, for $\theta \leq 0$, we have

$$M''(\theta) \leq \frac{1}{(1-\lambda)^3} \leq \frac{1}{(1-\lambda^*)^3}.$$

Thus, for any $\theta \leq (1-\epsilon) r_2^*$ (where $\epsilon > 0$ is any fixed constant), we have

$$M''(\theta) \leq A, \tag{3.10}$$

where A is a fixed constant (depending only on λ^*). Note that from (3.6) and (3.8), we have

$$\mathbf{E}[H_X] = \mathbf{E}[H] = 1/(1-\lambda). \tag{3.11}$$

Consider next that $X_0 = 0$. For $r \geq 1$, let τ_r be the time elapsed when X_t first returns to 0 for the r-th time. We shall obtain a concentration result for τ_r (when r is large). Observe that τ_1 equals H_X in distribution (this is because X_1 has the same distribution when $X_0 = 0$ or $X_0 = 1$) and so τ_r is distributed as a sum of r independent copies of H_X. Hence, $\mathbf{E}[\tau_r] = r/(1-\lambda)$. We shall use the inequalities

$$\begin{aligned} \mathbf{Pr}(\tau_r \geq A) &\leq M_X(\theta)^r \exp(-A\theta), \\ \mathbf{Pr}(\tau_r \leq A) &\leq M_X(-\theta)^r \exp(A\theta), \end{aligned}$$

for any $\theta > 0$. As $M_X(\theta) \leq M(\theta)$ by (3.7), we shall bound $M(\theta)$. Using Taylor's theorem and (3.11),

$$M(\theta) = 1 + \theta/(1-\lambda) + M''(\xi)\theta^2/2,$$

for some ξ between 0 and θ. Using (3.10), we have that as $\theta \to 0$,

$$M''(\xi) = O(1),$$

which implies that

$$M(\theta) = 1 + \theta/(1-\lambda) + O(\theta^2).$$

Hence, for any $A > 0$ and small $\theta > 0$,

$$\begin{aligned} & \mathbf{Pr}(\tau_r \geq r/(1-\lambda) + Ar^{1/2}) \\ \leq\ & M(\theta)^r \exp(-r\theta/(1-\lambda) - A\theta r^{1/2}) \\ \leq\ & \exp(O(r\theta^2) - A\theta r^{1/2}). \end{aligned}$$

Also, we have for any $A > 0$ and small $\theta > 0$,

$$\begin{aligned} & \mathbf{Pr}(\tau_r \leq r/(1-\lambda) - Ar^{1/2}) \\ \leq\ & M(-\theta)^r \exp(r\theta/(1-\lambda) - A\theta r^{1/2}) \\ \leq\ & \exp(O(r\theta^2) - A\theta r^{1/2}). \end{aligned}$$

By putting $\theta = r^{-1/2}$, we have for any $A > 0$ and for large r

$$\mathbf{Pr}(|\tau_r - r/(1-mp)| \geq Ar^{1/2}) = O(\mathrm{e}^{-A}). \tag{3.12}$$

We therefore have the following lemma.

Lemma 3.1 *Let τ_r be the time elapsed when X_t first returns to 0 for the r-th time given that $X_0 = 0$. Then for any $A > 0$, we have as $r \to \infty$,*

$$\mathbf{Pr}(|\,\tau_r - r/(1-\lambda)\,| \geq Ar^{1/2}) = O(\mathrm{e}^{-A}).$$

Lemma 3.2 *Suppose that $X_0 = r$ for any integer $r \geq 1$. Let $H_r = \min\{t \mid X_t = 0\}$. Then for any $A > 0$,*

$$\mathbf{Pr}(|\,H_r - r/(1-\lambda)\,| \geq Ar^{1/2}) = O(\mathrm{e}^{-A}). \tag{3.13}$$

Also, we have for any $A > 0$ that

$$\mathbf{Pr}(\exists t \leq H_r \textit{ s.t. } X_t \geq r/(1-\lambda) + Ar^{1/2}) = O(\mathrm{e}^{-A}). \tag{3.14}$$

Proof Simply observe that H_r is distributed as a sum of r independent copies of H_X, and so H_r equals τ_r in distribution, which gives (3.13). Equation (3.14) follows from (3.13) and the fact that X_t decreases by at most 1 in each transition. □

Lemma 3.3 *Let N_T be the number of times that X_t equals 0 in the time interval $[0,T]$, given that $X_0 = O(\log^{10} T)$. Then for any $A > 0$, we have for any constant $A' > 0$ that*

$$\mathbf{Pr}(|\,N_T - T(1-\lambda)\,| \geq AT^{1/2}) = O(\mathrm{e}^{-A} + T^{-A'}). \tag{3.15}$$

Proof Use H to denote the minimum value of t such that $X_t = 0$. Using (3.13) with $r = O(\log^{10} T)$, we have for any constant $A' > 0$ that

$$\mathbf{Pr}(H \geq \log^{11} T) = O(\mathrm{e}^{-\log^6 T}) = O(T^{-A'}).$$

Hence if N'_T is the number of times that $X_t = 0$ in the interval $[0,T]$ given that $X_0 = 0$, then $N'_T \geq N_T \geq N'_{T-\log^{11} T}$ with probability at least $1 - O(T^{-A'})$ for any constant $A' > 0$. Now Lemma 3.1 implies that as $t \to \infty$,

$$\mathbf{Pr}(|\,N'_t - t(1-\lambda)\,| \geq At^{1/2}) = O(\mathrm{e}^{-A/(1-\lambda)}) = O(\mathrm{e}^{-A}).$$

The lemma now follows by taking $t = T$ and $t = T - \log^{11} T$. □

Lemma 3.4 *Suppose that $X_0 = 0$. With τ_1 (τ_r with $r = 1$) as defined in Lemma 3.1, we have for any $A > 0$, there exist a constant $\rho \in (0,1)$ and a constant $C > 0$ such that*

$$\mathbf{Pr}(\tau_1 \geq A) \leq C\rho^{-A}. \tag{3.16}$$

For each t, let $R_t = \min\{k \geq 1 \mid X_{t+k} = 0\}$. That is, R_t is the waiting time after time t until the next return to 0. Then for any $A > 0$, there is a constant $\rho \in (0,1)$ such that as $T \to \infty$,

$$\mathbf{Pr}(\max_{t \leq T} R_t \geq A) = O(T\rho^{-A}), \tag{3.17}$$

and

$$\mathbf{Pr}(\max_{t \leq T} X_t \geq A) = O(T\rho^{-A}), \tag{3.18}$$

Proof Since τ_1 equals H_X in distribution, we have

$$\mathbf{Pr}(\tau_1 \geq A) \leq M_X(\theta) \exp(-A\theta).$$

Inequality (3.16) follows by putting $\theta = r_2^*/2$. To show (3.17), let S_i be the time elapsed between the $(i-1)$-th and the i-th return to 0. That is, each S_i equals τ_1 in distribution. Let N be the number of times that $X_t = 0$ for $t \in [0, T]$. Then $N \leq T$ and (3.17) follows from (3.16) because

$$\mathbf{Pr}(\max_{t \leq T} R_t \geq A) \leq \mathbf{Pr}(\max_{t \leq T} S_t \geq A) = O(T\rho^{-A}).$$

Inequality (3.18) follows from (3.17) and the fact that X_t decreases by at most 1 in each transition. □

For the rest of the section, we will require coupling chain X_t with another chain X_t' having the same transition probability. The coupling is such that if $X_0 \leq X_0'$ then $X_t \leq X_t'$ for all $t \geq 0$. This coupling is specified by defining the transition probabilities of the coupled chain (X_t, X_t') as follows:

$$\begin{array}{llll} \Delta X_t & = \Delta X_t' & = B(m,p) - 1, & \text{if } X_t > 0 \text{ and } X_{t+1} > 0 \\ \Delta X_t - 1 & = \Delta X_t' & = B(m,p) - 1, & \text{if } X_t = 0 \text{ and } X_{t+1} > 0 \\ \Delta X_t & = \Delta X_t' & = B(m,p), & \text{if } X_t = 0 \text{ and } X_{t+1} = 0 \\ \Delta X_t & = \Delta X_t' - 1 & = B(m,p) - 1, & \text{if } X_t > 0 \text{ and } X_{t+1} = 0. \end{array}$$

Lemma 3.5 *Suppose that $X_0 = O(\log^{10} T)$. Then for any $A > 0$ and for large T, there is a constant $\rho \in (0,1)$ and a constant $C > 0$ such that*

$$\mathbf{Pr}(X_T \geq A) \leq C\rho^{-A} + O(T^{-A'}),$$

for any constant $A' > 0$.

Proof Use H to denote the minimum value of t such that $X_t = 0$. Note that for $t \geq H$, it follows from coupling X_t with the steady state chain $\hat{X}_t$ that the distribution of X_t is stochastically at most the steady state distribution. Hence,

$$\begin{aligned} \mathbf{Pr}(X_T \geq A) &\leq \mathbf{Pr}(X_T \geq A \mid H \leq T)\mathbf{Pr}(H \leq T) + \mathbf{Pr}(H > T) \\ &\leq \mathbf{Pr}(\hat{X}_t \geq A)\mathbf{Pr}(H \leq T) + \mathbf{Pr}(H > T). \end{aligned}$$

Now from (3.13), we have

$$\mathbf{Pr}(H > T) = O(T^{-A'}),$$

for any constant $A' > 0$. (Note that although $\mathbf{Pr}(H > T)$ should be exponentially small, our bound here will suffice for future applications.) To bound $\mathbf{Pr}(\hat{X}_t \geq A)$, we note that according to (3.2) and the comments that followed, the moment generating function $M_\pi(\theta)$ of $\hat{X}_t$ is properly defined for $\theta < \log r_1^*$. Hence, similar to proof of (3.16), there are constants $\rho \in (0,1)$ and $C > 0$ such that

$$\mathbf{Pr}(\hat{X}_t \geq A) \leq C\rho^{-A}.$$

The lemma now follows. □

For the next lemma, we let X_t denote the chain with initial state $X_0 = O(\log^2 n)$ and compare it with the steady state chain $\hat{X}$ after $h = \lfloor \log^9 n \rfloor$ steps.

Lemma 3.6 *As $n \to \infty$,*

$$\begin{aligned} \mathbf{Pr}(X_h = 0) &= \mathbf{Pr}(\hat{X} = 0) + o(1), \\ \mathbf{E}[X_h] &= \mathbf{E}[\hat{X}] + o(1). \end{aligned}$$

Proof We shall show the lemma for the case where $X_0 = \lceil \log^2 n \rceil$. Let $\mathcal{E}$ be the event that $\hat{X}_0 \geq \log^2 n$. Then from the last equation in the proof of the previous lemma, we have

$$\mathbf{Pr}(\mathcal{E}) = \mathbf{Pr}(\hat{X}_t \geq \log^2 n) = O(n^{-A}), \tag{3.19}$$

for any contstant A. Next, let H be the waiting time until X_t first hits 0. Then from (3.13), we have

$$\mathbf{Pr}(H > h) = O(n^{-A}), \tag{3.20}$$

for any constant A. Now in the coupling of X_t and $\hat{X}_t$, if $\mathcal{E}$ does not occur, X_h must equal $\hat{X}_h$ if $H \leq h$. (This is because $X_H = \hat{X}_H = 0$ on the event $\bar{\mathcal{E}}$.) Thus

$$\mathbf{Pr}(X_h = 0) \leq \mathbf{Pr}(\mathcal{E}) + \mathbf{Pr}(H > h) + \mathbf{Pr}(\hat{X}_h = 0),$$

and so the first assertion of the lemma follows. Also,

$$|\mathbf{E}[X_h] - \mathbf{E}[\hat{X}_h]| \leq \mathbf{E}[|X_h - \hat{X}_h|]$$

in the coupling. Let $\chi(\mathcal{D})$ be the indicator for the event $\mathcal{D}$. Then

$$\begin{aligned} \mathbf{E}[|X_h - \hat{X}_h|] &= \mathbf{E}[\chi(\mathcal{E})|X_h - \hat{X}_h|] + \mathbf{E}[(1 - \chi(\mathcal{E}))\chi(H > h)|X_h - \hat{X}_h|] \\ &\leq \mathbf{E}[\chi(\mathcal{E})|X_h - \hat{X}_h|] + \mathbf{E}[\chi(H > h)|X_h - \hat{X}_h|] \\ &\leq \mathbf{E}[\hat{X}_0\chi(\mathcal{E})] + (\log^2 n)\mathbf{E}[\chi(H > h)] \\ &\leq \mathbf{E}[\hat{X}_0^2]\mathbf{Pr}(\mathcal{E}) + (\log^2 n)\mathbf{Pr}(H > h), \end{aligned}$$

which equals $o(1)$ from (3.19), (3.20) and the fact that $\mathbf{E}[\hat{X}_0^2] = O(1)$. □

4 Proof of Theorems 1.1(a) and 1.1(c)

We shall first assume Theorem 1.1(b) and prove Theorem 1.1(c) by a monotonicity argument to show that when $c > c_3$, the probability that GUC succeeds is $o(1)$. We first consider the monotonicity argument. Suppose that we have two random instances of k-SAT on n variables with m and $\hat{m}$ clauses of size k respectively. Assume $m \leq \hat{m}$. Let $N(\nu) = (N_0(\nu), N_1(\nu), N_2(\nu), N_3(\nu))$ and $\hat{N}(\nu) = (\hat{N}_0(\nu), \hat{N}_1(\nu), \hat{N}_2(\nu), \hat{N}_3(\nu))$ denote their respective states in GUC when there are ν variables whose truth values remain undetermined. We aim to give a coupling of $N(\nu)$ and $\hat{N}(\nu)$ so that $N(\nu) \leq \hat{N}(\nu)$. Note that the transition probabilities of N are given at the end of Section 2 and that the transition probabilities of $\hat{N}$ are defined similarly with Δ replaced with $\hat{\Delta}$ and N with $\hat{N}$. Note also that $N(n) = (0, \ldots, 0, m)$ and $\hat{N}(n) = (0, \ldots, 0, \hat{m})$ and so $N(n) \leq \hat{N}(n)$. We shall show that if $N(\nu) \leq \hat{N}(\nu)$, then $N(t) \leq \hat{N}(t)$ for $t < \nu$ by coupling arguments.

Lemma 4.1 *If $N(\nu) \leq \hat{N}(\nu)$, then the chains N and $\hat{N}$ can be coupled so that $N(t) \leq \hat{N}(t)$ for $t < \nu$.*

Proof Let $i \geq 1$ be the minimum integer such that $\hat{N}_i(\nu) \neq 0$. Now for $j \neq i$, $\chi_j(\hat{N}(\nu)) = 0$ and $\chi_i(\hat{N}(\nu)) = 1$. Thus, for $j \neq i$,

$$N_j(\nu) - \chi_j(N(\nu)) \leq \hat{N}_j(\nu) - \chi_j(\hat{N}(\nu)).$$

For $j = i$, we have $\hat{N}_i(\nu) \geq 1$ and note that if $N_i(\nu) = 0$ then

$$N_i(\nu) - \chi_i(N(\nu)) = 0 \leq \hat{N}_i(\nu) - 1 = \hat{N}_i(\nu) - \chi_i(\hat{N}(\nu)),$$

and that if $N_i(\nu) \geq 1$ then $\chi_i(N(\nu)) = 1$, from which we have

$$N_i(\nu) - \chi_i(N(\nu)) = N_i(\nu) - 1 \leq \hat{N}_i(\nu) - 1 = \hat{N}_i(\nu) - \chi_i(\hat{N}(\nu)).$$

Therefore, we have for all $i = 1, \ldots, k$,

$$N_i(\nu) - \chi_i(N(\nu)) \leq \hat{N}_i(\nu) - \chi_i(\hat{N}(\nu)). \tag{4.1}$$

Observe next that for any two binomial variables $B = B(\tau, p)$ and $\hat{B} = B(\hat{\tau}, p)$ with $\tau \leq \hat{\tau}$, we can couple B and $\hat{B}$ so that

$$\begin{aligned} B &\leq \hat{B}, \\ \tau - B &\leq \hat{\tau} - \hat{B}, \end{aligned}$$

where the coupling is obtained by identifying B as the sum of the first τ Bernoulli variables from the $\hat{\tau}$ independent Bernoulli variables in $\hat{B}$. It follows from (4.1) that we may couple N and $\hat{N}$ so that for $i = 1, \ldots, k$,

$$\Delta_{i,0}(\nu) \leq \hat{\Delta}_{i,0}(\nu), \tag{4.2}$$

$$N_i(\nu) - \chi_i(N(\nu)) - \Delta_{i,0}(\nu) \leq \hat{N}_i(\nu) - \chi_i(\hat{N}(\nu)) - \hat{\Delta}_{i,0}(\nu). \tag{4.3}$$

It follows similarly from (4.2) that we may couple N and $\hat{N}$ so that for $i = 1, \ldots, k$,

$$\Delta_{i,1}(\nu) \leq \hat{\Delta}_{i,1}(\nu). \tag{4.4}$$

Combining (4.3) and (4.4) gives that $N(\nu - 1) \leq \hat{N}(\nu - 1)$. We can then repeat this coupling for $\nu - 1, \nu - 2, \ldots, 1$ to give the lemma. □

Proof of Theorem 1.1(c). For $c \geq c_3$, we have $c > c_3 - \epsilon$ for any $\epsilon > 0$. Now for a random instance $\mathcal{I}_\epsilon$ of 3-SAT with $\lfloor (c_3 - \epsilon)n \rfloor$ clauses and n variables, Theorem 1.1(b) gives that the limit (as $n \to \infty$) of the probability that GUC succeeds when applied to $\mathcal{I}_\epsilon$ is arbitrarily close to 0 for sufficiently small $\epsilon > 0$. Theorem 1.1(c) thus follows from monotonicity. □

To show Theorem 1.1(a), we apply a result of Chvátal and Reed [4] which can be stated as follows. Suppose that $c < 2/3$ and consider applying algorithm SC to a random instance of 3-SAT with n variables and $\lfloor cn \rfloor$ clauses. Then the probability that SC succeeds equals $1 - o(1)$ as $n \to \infty$. Theorem 1.1(a) now follows from the following lemma.

Lemma 4.2 *Consider applying both GUC and SC to a random instance of k-SAT with n variables and m clauses. Then*

$$\mathbf{Pr}(\text{SC succeeds}) \leq \mathbf{Pr}(\text{GUC succeeds}).$$

Proof Consider applying both SC and GUC to a random instance $\mathcal{I}$ of k-SAT with n variables and m clauses. Let $N(\nu) = (N_0(\nu), \ldots, N_k(\nu))$ and $N'(\nu) = (N_0'(\nu), \ldots, N_k'(\nu))$ denote the respective states of $\mathcal{I}$ in GUC and SC when there are ν variables whose truth values remain undetermined. Note that $N(n) = N'(n)$ initially and that the transition probabilities of $N(\nu)$ and $N'(\nu)$ are given at the end of Section 2. Note also that if $N(\nu) \leq N'(\nu)$ then $\Delta_{j,0}'$ and $\Delta_{j,0}$ can be coupled so that $\Delta_{j,0}' \leq \Delta_{j,0}$. Thus, by following the coupling arguments in proof of Lemma 4.1, we have that if $N(\nu) \leq N'(\nu)$ then the chains N and N' can be coupled so that $N(t) \leq N'(t)$ for $1 \leq t < \nu$. This shows in particular that $N_0(\nu) \leq N_0'(\nu)$, and so the lemma follows. □

5 Proof of Theorem 1.1(b)

Assume $c \in (2/3, c_3)$. Recall that

$$f(x) = f_c(x) = \frac{3c}{4}(1 - x^2) + \log x, \quad x \in (0, 1),$$

and c_3 is the maximum value of c such that $f(x) \leq 1$ for all $x \in (0, 1)$. Let $\alpha = \alpha(c)$ (for $c > 2/3$) be the root of the equation $f(x) = 0$ that is strictly less than 1. Note that α is uniquely defined and that α is positive. By investigating the behaviour of $f(\alpha(1 + \epsilon))$ for small $\epsilon > 0$, we see that $c\alpha^2 < 2/3$ and also if

$$\alpha_0 = \alpha + n^{-0.24}$$

then

$$nf(\alpha_0) = \Theta(n^{0.76}).$$

Note that both αn and $\alpha_0 n$ equal $\Omega(n)$. We shall show that if $c \in (2/3, c_3)$, then $N_2(\nu)$ can be approximated by $\nu f(\nu/n)$ as ν decreases from n to $\alpha_0 n$. We shall also show that if c

and ν are within these ranges, then $N_3(\nu)$ can be approximated by $c\nu(\nu/n)^2$. (Thus, when $\nu = \lfloor \alpha_0 n \rfloor$, we see that $N_2(\nu) = \Theta(n^{0.76})$ and $N_3(\nu) \approx c\alpha_0^3 n$). These estimates enable us to find the limit of the probability that GUC succeeds.

In order to minimize subscripts, we write $W(\nu) = N_1(\nu)$, $Y(\nu) = N_2(\nu)$ and $Z(\nu) = N_3(\nu)$. We shall also consider a process $X(\nu)$ which runs alongside $N(\nu)$, and so we have a Markov chain $(N_0, W(\nu), X(\nu), Y(\nu), Z(\nu))$. The transition probabilities of (N_0, W, Y, Z) are same as N, but those of X need defining. For completeness, we write down the one-step transitions of $(W(\nu), X(\nu), Y(\nu), Z(\nu))$ below.

$$\begin{aligned}
\Delta Z(\nu) &= -\Delta_{3,0} - \chi_3((N_0, W, Y, Z)) \\
\Delta Y(\nu) &= \Delta_{3,1} - \Delta_{2,0} - \chi_2((N_0, W, Y, Z)) \\
\Delta X(\nu) &= \Delta_{2,1} - \chi_1((N_0, X, Y, Z)), \\
\Delta W(\nu) &= \Delta_{2,1} - \Delta_{1,0} - \chi_1((N_0, W, Y, Z)), \\
\Delta N_0(\nu) &= \Delta_{1,1}(\nu),
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{3,0} = \Delta_{3,0}(\nu) &= B(Z - \chi_3((N_0, W, Y, Z)), 3/\nu), \\
\Delta_{3,1} = \Delta_{3,1}(\nu) &= B(\Delta_{3,0}, 1/2), \\
\Delta_{2,0} = \Delta_{2,0}(\nu) &= B(Y - \chi_2((N_0, W, Y, Z)), 2/\nu), \\
\Delta_{2,1} = \Delta_{2,1}(\nu) &= B(\Delta_{2,0}, 1/2), \\
\Delta_{1,0} = \Delta_{1,0}(\nu) &= B(W - \chi_1(((N_0, W, Y, Z)), 1/\nu), \\
\Delta_{1,1} = \Delta_{1,0}(\nu) &= B(\Delta_{1,0}, 1/2).
\end{aligned}$$

The initial state of the process is $(N_0(n), W(n), X(n), Y(n), Z(n)) = (0, 0, 0, 0, \lfloor cn \rfloor)$. As the transitions of $X(\nu)$ ignores the effects of $-\Delta_{10}(\nu)$, we have $W(\nu) \leq X(\nu)$ always (which can be checked by considering the cases where $X(\nu) = W(\nu)$ and $X(\nu) > W(\nu)$). We shall see that $X(\nu)$ is a good approximation of $W(\nu)$.

We shall need the following bounds for sums of independent binomial variables. Let $B_1(\tau_1, p_1)$, $\ldots, B_k(\tau_k, p_k)$ be independent binomial variables. Write $\tau = \tau_1 + \ldots + \tau_k$ and $\bar{p} = \sum_i \tau_i p_i / \tau$. Then for A satisfying $0 < A < \tau\bar{p}/3$

$$\mathbf{Pr}\left(| B_1 + \ldots + B_k - \tau\bar{p} | \geq \sqrt{3A\tau\bar{p}} \right) \leq 2\exp(-A). \tag{5.1}$$

Also, for a binomial variable $B(\tau, p)$, we have for $u \geq \mathrm{e}$,

$$\mathbf{Pr}(B \geq u\tau p) \leq (\mathrm{e}/u)^{u\tau p}. \tag{5.2}$$

All our subsequent error probabilities regarding sums like $\sum \Delta_{3,0}$ are derived from one of the above inequalities. We shall be bounding such sums by sums of independent binomial variables. Although the variables in sums like $\sum \Delta_{3,0}$ are usually not independent, it is not difficult to show the stochastic dominance by induction and by conditioning on the outcomes of the partial sums. Also, we say that an event $\mathcal{E}$ occurs with high probability (w.h.p. for short) if

$$\mathbf{Pr}(\mathcal{E}) = 1 - O(n^{-A}), \tag{5.3}$$

for any constant $A > 0$. Now the events $\mathcal{E}$ usually contain bounds, involving some big O terms, for random variables. In this situation, it will be clear that equations like (5.3) hold for any $A > 0$ by choosing sufficiently large constants (which may depend on A) in the big O terms. We first prove the following lemma which will be useful for future inductive proofs. Note that we make no attempt to minimize the powers of $\log n$.

Lemma 5.1 *Suppose that $\nu \geq \alpha_0 n$. Let $h = \lfloor n^{1/2} \rfloor$, $\nu' = \nu - h$ and $I = \{\nu' + 1, \ldots, \nu\}$. Suppose that at stage ν,*

$$\begin{aligned} Z(\nu) &= c\nu^3/n^2 + z(n), \\ Y(\nu) &= \nu f(\nu/n) + y(n), \\ W(\nu) &= w(n) \ \leq \ \log^{10} n, \end{aligned}$$

where $z(n) = o(n)$ and $y(n) = o(n^{0.76})$. Then with high probability,

$$\begin{aligned} Z(\nu') &= c\nu'^3/n^2 + O(z(n) + n^{1/4}\log n), && (5.4) \\ Y(\nu') &= \nu' f(\nu'/n) + O(y(n) + z(n)n^{-1/2} + n^{1/4}\log n), && (5.5) \\ W(\nu') &\leq \log^2 n, && (5.6) \end{aligned}$$

(The constants in the big O terms are independent of ν.)

When proving the above lemma, we shall obtain the following estimates which will be useful later.

Lemma 5.2 *With hypotheses of Lemma 5.1, we have with high probability that for all $j \in I$,*

$$\begin{aligned} Z(j) &= Z(\nu) + O(n^{1/2}), && (5.7) \\ Y(j) &= Y(\nu) + O(n^{1/2}). && (5.8) \end{aligned}$$

Let τ be the minimum value of $k \geq 0$ such that $W(\nu - k) = 0$, and for $j \in I$, let τ_j be the minimum value of $k \geq 1$ such that $W(j - k) = 0$. Then we have with high probability that

$$\begin{aligned} \tau &= O(w(n) + \sqrt{w(n)}\log n), && (5.9) \\ \tau_j &\leq \log^2 n, \quad \text{for } j \leq \nu - \tau. && (5.10) \end{aligned}$$

Also, we have with high probability that for $j \geq \nu - \tau$,

$$W(j) = O(w(n) + \sqrt{w(n)}\log n), \tag{5.11}$$

and that for $j \leq \nu - \tau$,

$$W(j) = O(\log^2 n). \tag{5.12}$$

Note that (5.9-5.12) imply that if $w(n) = O(\log^2 n)$, then we have with high probability that for all $j \in I$,

$$\begin{aligned} W(j) &= O(\log^2 n), & (5.13) \\ \tau_j &= O(\log^2 n). & (5.14) \end{aligned}$$

Proof We shall prove Lemma 5.1 and point out from where the statements in Lemma 5.2 follow. Note first that since $\alpha_0 n = \Omega(n)$, both ν and $Z(\nu)$ equal $\Omega(n)$. Define $\Delta' Z$ as the number of times that $Y(j) = W(j) = 0$ for $j \in I$, and $\Delta' Y$ be the number of times that $Y(j) \neq 0$ but $W(j) = 0$. Similarly, let $\Delta' W$ be the number of times that $W(j) = 0$. Therefore, we have

$$\begin{aligned} Z(\nu') - Z(\nu) &= -\sum_{j \in I} \Delta_{3,0}(j) - \Delta' Z, & (5.15) \\ Y(\nu') - Y(\nu) &= \sum_{j \in I} (\Delta_{3,1}(j) - \Delta_{2,0}(j)) - \Delta' Y. & (5.16) \end{aligned}$$

To estimate $U_Z = \sum_{j \in I} \Delta_{3,0}(j)$, we note that $\Delta_{3,0}(j)$ is bounded above in distribution by a binomial variable with parameters $Z(\nu)$ and $3/\nu'$. Thus it is not difficult to obtain that $\sum_{j \in I} \Delta_{3,0}(j)$ is bounded above by a sum of independent binomial variables, each with parameters $Z(\nu)$ and $3/\nu'$. This gives an upper bound (w.h.p.) $U_Z = O(h)$ for the sum of the variables. Since $Z(\nu') \leq Z(j) \leq Z(\nu)$, we have with high probability that $Z(j) = Z(\nu) - O(h)$, which is (5.7). Hence, with high probability, $\Delta_{3,0}(j)$ is bounded below by a binomial variable with parameters $Z(\nu) - U_Z$ and $3/\nu$. Since as $n \to \infty$,

$$\frac{3Z(\nu) - O(h)}{\nu - O(h)} = \frac{3Z(\nu)}{\nu} + O(n^{-1/2}),$$

we have with high probability that,

$$\sum_{j \in I} \Delta_{3,0}(j) = \frac{3hZ(\nu)}{\nu} + O(n^{1/4} \log n). \qquad (5.17)$$

Similarly, we have with high probability that

$$\sum_{j \in I} \Delta_{3,1}(j) = \frac{3hZ(\nu)}{2\nu} + O(n^{1/4} \log n), \qquad (5.18)$$

which gives us an upper bound $Y(\nu) + O(h)$ for $Y(j)$ where $j \in I$. As each $\Delta_{2,0}(j)$ is distributed as a binomial variable with parameters $Y(j) + O(1) = O(n)$ and $2/j = O(1/n)$, we have with high probability that

$$\sum_{j \in I} \Delta_{2,0}(j) = O(h).$$

Since $\Delta' Y = O(h)$, we therefore have a lower bound $Y(\nu) - O(h)$ for $Y(j)$ where $j \in I$. Thus, we have $Y(j) = Y(\nu) + O(h)$ with high probability (which is (5.8)). Hence, with high probability, each $\Delta_{2,0}(j)$ is bounded above and below in distribution by binomial variables with parameters $Y(\nu)+O(h)$ and $2/(\nu+O(h))$. It thus follows that

$$\sum_{j\in I}\Delta_{2,0}(j)=\frac{2hY(\nu)}{\nu}+O(n^{1/4}\log n) \tag{5.19}$$

with high probability. To estimate $\Delta' Z$, note first that if $\nu\le n-n^{0.76}$ (but $\nu\ge\alpha_0 n$), then from the hypotheses in the lemma, we have

$$Y(\nu)=\Omega(n^{0.76}).$$

Note that during the entire time interval I, the number of size two clauses removed is at most $\sum_{j\in I}\Delta_{2,0}(j)+h$, which equals $O(n^{1/2})$ with high probability (using (5.19)). Thus the quantity $Y(j)$, for $j\in I$, is never zero and so when $\nu\le n-n^{0.76}$,

$$\Delta' Z=0 \tag{5.20}$$

with high probability. For the case where $\nu\ge n-n^{0.76}$, we consider stage $k\in I$ with $k\le\nu-n^{0.1}$ and write $h'=\nu-k$. Then similar to (5.18), we have with high probability that

$$\sum_{j=k+1}^{\nu}\Delta_{3,1}(j)=\frac{3h'Z(\nu)}{2\nu}(1+o(1))=\frac{3c\nu^2h'}{n^2}(1+o(1))=\frac{3ch'}{2}(1+o(1)). \tag{5.21}$$

Also, note that $Y(\nu)=O(n^{0.76})$ (for $\nu\ge n-n^{0.76}$) and so similar to (5.19), we have that for any fixed $\epsilon>0$,

$$\sum_{j=k+1}^{\nu}\Delta_{2,0}(j)\le\epsilon h', \tag{5.22}$$

with high probability. Now in order for $Y(k)=0$, we must have

$$\sum_{i=k+1}^{\nu}(\Delta_{3,1}(i)-\Delta_{2,0}(i))\le h'.$$

which, since $c>2/3$ and according to (5.21) and (5.22), occurs with probability $O(n^{-A})$, for any constant $A>0$. This shows that with high probability, $Y(k)\ne 0$ for all $k\le\nu-n^{0.1}$. Thus with high probability, there are at most $n^{0.1}$ times when $Y(j)=0$ (where $j\in I$). Combining this with (5.20), we have with high probability that

$$\Delta' Z=O(n^{0.1}). \tag{5.23}$$

Using (5.15) and (5.17) and (5.23), we have with high probability that

$$\begin{aligned}Z(\nu') &= Z(\nu)-\frac{3hZ(\nu)}{\nu}+O(n^{1/4}\log n)\\ &= Z(\nu)(1-3h/\nu)+O(n^{1/4}\log n)\\ &= Z(\nu)(\nu'/\nu)^3(1+O(1/n))+O(n^{1/4}\log n)\\ &= c\nu'^3/n^2+O(z(n)+n^{1/4}\log n).\end{aligned}$$

This proves (5.4). Next, we like to estimate $\Delta'Y$ and $\Delta'W$. In view of (5.23), we have $\Delta'Y = \Delta'W - O(n^{0.1})$ with high probability. To estimate $\Delta'W$, we consider a process $\{X(j) \mid j \leq n\}$ with transition probabilities as defined in the beginning of this section. We also let $X(\nu) = W(\nu)$. Then as observed before, we have $W(j) \leq X(j)$ for all $j \leq \nu$. Let $\Delta'X$ be the number of times that $X(j) = 0$ for $j \in I$, and so $\Delta'W \geq \Delta'X$ (as $W(j) \leq X(j)$). Next, observe that similar to our proof of (5.19), we have with high probability that $\Delta_{2,1}(j)$ (for all $j \in I$) is bounded above and below in distribution by binomial variables with parameters $Y(\nu) + O(h)$ and $1/(\nu + O(h))$. Now according to the hypotheses of the lemma,

$$0 < \frac{Y(\nu) + O(h)}{\nu + O(h)} = f(\nu/n)(1 + o(1)),$$

which is bounded above by a constant less than 1 (since $c < c_3$). Hence, with high probability, we have that $X(j)$ (for all $j \in I$) is bounded above and below in distribution by the states of two Markov chains similar to the Markov chain described in the previous section. It therefore follows from Lemma 3.3 (by taking λ there as $(Y(\nu) + O(h))/(\nu + O(h))$, T there as h, A there as $O(\log h)$) that with high probability,

$$\begin{aligned} \Delta'X &= h\left(1 - \frac{Y(\nu) + O(h)}{\nu + O(h)}\right) + O(h^{1/2}\log h) \\ &= n^{1/2} - \frac{Y(\nu)n^{1/2}}{\nu} + O(n^{1/4}\log n). \end{aligned} \tag{5.24}$$

We shall show next that $\Delta'W$ and $\Delta'X$ do not differ by much. We do this by finding an estimate for

$$\sum_{j \in I} (X(j) - W(j)),$$

which will also be useful later. Let $\tau' = \max\{k \leq \nu \mid X(k) = 0\}$ and use τ'_j to denote the minimum value of $k \geq 1$ such that $X(j-k) = 0$. Note that when $X(j) = 0$, $W(j)$ is necessarily equal to 0 (as $W \leq X$). Hence whenever $\Delta_{1,0}(j) \geq 1$, its cumulative effect on $\sum W$ stops when X next gets to 0. Thus,

$$\sum_{j \in I} (X(j) - W(j)) \leq \sum_{j=\tau'+1}^{\nu} X(j) + \sum_{j=\nu'+1}^{\tau'} \Delta_{1,0}(j)\tau'_j.$$

Recall that as argued above, $X(j)$ behaves like the Markov chain X_j discussed in the previous section. To estimate τ', note that if $w(n) = 0$, then $\tau' = \nu$; otherwise we apply (3.13) (with n there as $w(n)$, and A there as $O(\log n)$) to obtain that

$$\nu - \tau' = O(w(n) + \sqrt{w(n)}\log n),$$

holds with high probability. (Since $W(j) \leq X(j)$, this gives (5.9).) Similarly, using (3.14), we have with high probability that for all j between ν and τ',

$$X(j) = O(w(n) + \sqrt{w(n)}\log n), \tag{5.25}$$

from which (5.11) follows. Thus, with $w(n) \leq \log^{10} n$, we have $\nu - \tau' = O(\log^{10} n)$ and $X(j) = O(\log^{10} n)$, from which we obtain that

$$\sum_{j=\tau'+1}^{\nu} X(j) = O(\log^{20} n)$$

holds with high probability. Next, for j between $\nu'+1$ and τ', we have from (3.18) that with high probability,

$$W(j) \leq X(j) \leq \log^2 n, \tag{5.26}$$

and so (5.12) follows. Next we use (3.17) to obtain that with high probability,

$$\tau'_j \leq \log^2 n, \tag{5.27}$$

from which (5.10) follows. (Note that strictly speaking, we have only showed that $X(j)$ can be approximated by a Markov chain X_j defined in the previous section for $j \in I$. This creates a problem when estimating τ'_j for j "close" to ν'. However, as it can be seen easily that our previous approximations for $Z(j)$ and $Y(j)$ work for j between $\nu' - \log^3 n$ and ν' also. This means that $X(j)$ can be approximated by X_j for all j between $\nu' - \log^3 n$ and ν. As (3.17) gives that $\tau'_{\nu'} = O(\log^2 n)$, inequality (5.27) now follows from (3.17) too.)

Note that $\Delta_{1,0}(j)$ is a binomial variable with parameters $W(j)+O(1)$ and $1/j$. Thus it follows from (5.26) that $\sum_{j=\nu'+1}^{\tau'} \Delta_{1,0}(j)$ is bounded above by a binomial variable with parameters $O(h\log^2 n)$ and $O(1/n)$. Hence (5.2) gives that

$$\sum_{j=\nu'+1}^{\tau'} \Delta_{1,0}(j) = O(\log n)$$

with high probability. It thus follows from (5.27) that

$$\sum_{j=\nu'+1}^{\tau'} \Delta_{1,0}(j)\tau'_j = O(\log^3 n)$$

with high probability. We thus conclude that with high probability,

$$\sum_{j\in I} (X(j) - W(j)) = O(\log^{20} n). \tag{5.28}$$

It follows that with high probability, we have

$$\Delta' W - \Delta' X = O(\log^{20} n).$$

This together with (5.24) give that with high probability,

$$\Delta' W = n^{1/2} - \frac{Y(\nu)n^{1/2}}{\nu} + O(n^{1/4}\log n). \tag{5.29}$$

Hence, combining (5.16), (5.18), (5.19), (5.29) and the fact that $\Delta' Y = \Delta' W - O(n^{0.1})$, we have with high probability that

$$Y(\nu') - Y(\nu) = \frac{3hZ(\nu)}{2\nu} - \frac{hY(\nu)}{\nu} - n^{1/2} + O(n^{1/4}\log n). \tag{5.30}$$

It follows from the hypotheses of the lemma that

$$\begin{aligned}
&Y(\nu') \\
= \;& Y(\nu)\left(1-\frac{n^{1/2}}{\nu}\right)+\frac{3cn^{1/2}}{2}\left(\frac{\nu}{n}\right)^2-n^{1/2}+O(n^{1/4}\log n+z(n)n^{-1/2}) \\
= \;& f(\nu/n)(\nu-n^{1/2})+\frac{3cn^{1/2}}{2}\left(\frac{\nu}{n}\right)^2-n^{1/2}+O(y(n)+n^{1/4}\log n+z(n)n^{-1/2}).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\nu' f(\nu'/n) \\
= \;& (\nu-n^{1/2})f\left(\frac{\nu}{n}\left(1-\frac{n^{1/2}}{\nu}\right)\right) \\
= \;& (\nu-n^{1/2})\left(f(\nu/n)+\left(\frac{\nu}{n}\right)^2\frac{3cn^{1/2}}{2\nu}-\frac{n^{1/2}}{\nu}\right)+O(1) \\
= \;& (\nu-n^{1/2})f(\nu/n)+\frac{3cn^{1/2}}{2}\left(\frac{\nu}{n}\right)^2-n^{1/2}+O(1).
\end{aligned}$$

This proves (5.5). For (5.6), we use the fact that $W(\nu') \le X(\nu')$. Since as observed previously that $X(j)$ can be approximated by a Markov chain X_t defined in the previous section, inequality (5.6) follows from Lemma 3.5. □

We now make use of Lemma 5.1 to show Theorem 1.1(b). Let $h = \lfloor n^{1/2} \rfloor$ as before, and write $n_i = n - ih$ and $I_i = \{n_i + 1, \ldots, n_{i-1}\}$. Define J as the greatest integer such that $n_J = n - Jh \ge \alpha_0 n$. Note first that by using induction and by applying Lemma 5.1 repeatedly, we have with high probability that for all $i \le J$,

$$\begin{aligned}
Z(n_i) &= cn_i^3/n^2 + O(in^{1/4}\log n), && (5.31)\\
Y(n_i) &= n_i f(n_i/n) + O(in^{1/4}\log n), && (5.32)\\
W(n_i) &\le \log^2 n, && (5.33)
\end{aligned}$$

where the constants in the big O terms are independent of i. Note that since $i \le J = O(n^{1/2})$, the error terms in the (5.31) and (5.32) are both equal to $O(n^{3/4}\log n) = o(n^{0.76})$. This implies that the values of $Z(n_i), Y(n_i)$ and $W(n_i)$ $(i \le J)$ satisfy the hypotheses of Lemma 5.1, and so induction works by applying Lemma 5.1 repeatedly. In particular, it follows from (5.13) that with high probability,

$$W(\nu) = O(\log^2 n), \qquad \text{for all } \nu \ge n_J. \tag{5.34}$$

We shall now prove the following two lemmas from which Theorem 1.1(b) follows immediately.

Lemma 5.3

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUC does not fail before stage } n_J) = \exp\left(-\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))}dx\right). \tag{5.35}$$

Lemma 5.4 *Suppose that at stage n_J,*

$$\begin{aligned} Z(n_J) &= cn_J^3/n^2 + o(n), \\ Y(n_J) &= n_J f(n_J/n) + o(n^{0.76}), \\ W(n_J) &\leq \log^{10} n. \end{aligned}$$

Then

$$\lim_{n\to\infty} \mathbf{Pr}(\text{GUC creates an empty clause at and after stage } n_J) = 0. \tag{5.36}$$

Proof of Lemma 5.3 Let ϕ_ν be the number of empty clauses created at stage ν and $\xi_\nu = \min\{1, \phi_\nu\}$. Note that conditional on $W(\nu) = w$, ϕ_ν is a distributed as a binomial variable with parameters $(w-1)^+$ and $1/(2\nu)$. Thus

$$\begin{aligned} \mathbf{Pr}(\phi_\nu \neq \xi_\nu \mid W(\nu) = w) &= \mathbf{Pr}(\phi_\nu \geq 2 \mid W(\nu) = w) \\ &= O(w^2/\nu^2). \end{aligned}$$

So if

$$\Phi = \sum_{\nu=n_J}^{n} \phi_\nu, \qquad \Xi = \sum_{\nu=n_J}^{n} \xi_\nu$$

then (5.34) implies

$$\begin{aligned} \mathbf{Pr}(\Phi \neq \Xi) &= O\left(n\frac{\log^2 n}{n^2}\right) \\ &= o(1). \end{aligned}$$

Since our aim is to show that Φ is asymptotically distributed (as $n \to \infty$) as a Poisson random variable with parameter

$$\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))} dx,$$

we need only show that Ξ is asympotically Poisson distributed with the right mean.

We shall do this by the method of moments. The r-th fractorial moment of Ξ is

$$\mathbf{E}[\Xi(\Xi-1)\dots(\Xi-r+1)] = r! \sum_{(i_1,i_2,\dots,i_r)\in S_r} \mathbf{Pr}(\mathcal{E}_r), \tag{5.37}$$

where $S_r = \{(i_1, i_2, \dots, i_r) : n_J \leq i_1 < i_2 < \dots < i_r\}$ and $\mathcal{E}_r = \{\xi_{i_1} = \xi_{i_2} = \dots = \xi_{i_r} = 1\}$. We next partition S_r into $S_r' \cup S_r''$ so that $S_r' = \{(i_1, i_2, \dots, i_r) \in S_r : i_1 \leq n - \log^{10} n$ and $i_{k+1} - i_k \geq \log^{10} n, k = 1, 2, \dots, r-1\}$ and $S_r'' = S_r \setminus S_r'$.

Let us first deal with S_r''. We have

$$|S_r''| = O(n^{r-1} \log^{10} n) \tag{5.38}$$

and we claim that for any $(i_1, i_2, \ldots, i_r) \in S_r$

$$\mathbf{Pr}(\mathcal{E}_r) = O\left(\left(\frac{\log^2 n}{n}\right)^r\right). \tag{5.39}$$

Combining (5.38) and (5.39) we will have

$$\sum_{(i_1,i_2,\ldots,i_r)\in S_r''} \mathbf{Pr}(\mathcal{E}_r) = O\left(n^{r-1}\log^{10} n \left(\frac{\log^2 n}{n}\right)^r\right) = o(1). \tag{5.40}$$

To prove (5.39) we write

$$\mathbf{Pr}(\mathcal{E}_r) = \prod_{t=1}^{r} \mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{E}_{t-1}). \tag{5.41}$$

We now consider a typical term in the product (5.41).

$$\mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{E}_{t-1}) = \sum_{\sigma_t \in N^3} \mathbf{Pr}(\xi_t = 1 \mid \mathcal{A}(\sigma_t, t), \mathcal{E}_{t-1})\mathbf{Pr}(\mathcal{A}(\sigma_t, t) \mid \mathcal{E}_{t-1})$$

where if $\sigma_t = (w, y, z)$ then $\mathcal{A}(\sigma_t, t) = \{W(i_t) = w, Y(i_t) = y, Z(i_t) = z\}$.

Now $\mathcal{E}_{t-1}$ refers to events in the history of the algorithm up to the start of stage i_t and so by complete independence

$$\begin{aligned}
\mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{A}(\sigma_t, t), \mathcal{E}_{t-1}) &= \mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{A}(\sigma_t, t)) \\
&\leq \frac{w}{i_t} \\
&\leq \frac{w}{n_J}.
\end{aligned}$$

So

$$\begin{aligned}
\mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{E}_{t-1}) &\leq \frac{1}{n_J}\sum_{\sigma_t} w(\sigma_t)\mathbf{Pr}(\mathcal{A}(\sigma_t, t) \mid \mathcal{E}_{t-1}) \\
&\leq \frac{1}{n_J}\sum_{w} w\mathbf{Pr}(W(i_t) = w \mid \mathcal{E}_{t-1}) \\
&\leq \frac{1}{n_J}\sum_{w} w\mathbf{Pr}(W(i_t) = w)/\mathbf{Pr}(\mathcal{E}_{t-1}) \\
&\leq \frac{1}{n_J}\left(B\log^2 n + \frac{n\mathbf{Pr}(W(i_t) \geq B\log^2 n)}{\mathbf{Pr}(\mathcal{E}_{t-1})}\right) \\
&\leq \frac{B\log^2 n}{\alpha_0 n} + \frac{n^{-2r}}{\mathbf{Pr}(\mathcal{E}_{t-1})}
\end{aligned}$$

from (5.34) (B denotes the hidden constant in (5.34)). Now either $\mathbf{Pr}(\mathcal{E}_{t-1}) \leq n^{-r}$ and we are done since $\mathbf{Pr}(\mathcal{E}_r) \leq \mathbf{Pr}(\mathcal{E}_{t-1})$ or

$$\begin{aligned}
\mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{E}_{t-1}) &\leq \frac{B\log^2 n}{\alpha_0 n} + n^{-r} \\
&\leq \frac{2B\log^2 n}{\alpha_0 n}.
\end{aligned}$$

Substituting in (5.41) gives (5.39).

We next find an estimate for $\mathbf{Pr}(\mathcal{E}_r)$ when $(i_1, i_2, \ldots, i_r) \in S'_r$. Let $h = \lfloor \log^9 n \rfloor$. Let $\Gamma_t = \{(w, y, z) : 0 \leq w \leq \log^2 n,\ y = i_t f(i_t/n) + O(n^{.76}),\ z = ci_t^3/n^2 + O(n^{.76})\}$ and $\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_r$. For $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_r) \in \Gamma$ let $\mathcal{A}(\sigma) = \bigwedge_{t=1}^r \mathcal{A}(\sigma_t, t)$. Then for $(i_1, i_2, \ldots, i_r) \in S'_r$ we have, where $\bar{\mathcal{D}} = \{\exists 1 \leq t \leq r : (W(i_t), Y(i_t), Z(i_t)) \notin \Gamma_t\}$,

$$\begin{aligned}
\mathbf{Pr}(\mathcal{E}_r) &= \sum_{\sigma \in S} \mathbf{Pr}(\mathcal{E}_r \mid \mathcal{A}(\sigma))\mathbf{Pr}(\mathcal{A}(\sigma)) + \mathbf{Pr}(\mathcal{E}_r \wedge \bar{\mathcal{D}}) \\
&= \sum_{\sigma \in S} \prod_{t=1}^r \mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{E}_r, \mathcal{A}(\sigma))\mathbf{Pr}(\mathcal{A}(\sigma)) + \mathbf{Pr}(\mathcal{E}_r \wedge \bar{\mathcal{D}}) \\
&= \sum_{\sigma \in S} \prod_{t=1}^r \mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{A}(\sigma_t, t))\mathbf{Pr}(\mathcal{A}(\sigma)) + O(n^{-A}), \qquad (5.42)
\end{aligned}$$

where the last equation follows from complete independence.

We now estimate $\mathbf{Pr}(\xi_{i_t} = 1 \mid \mathcal{A}(\sigma_t, t))$. As argued in our proof of Lemma 5.1, $W(\nu)$ can be approximated by a Markov chain defined in Section 2. Thus using (3.4), (3.5) and Lemma 3.6, we have

$$\begin{aligned}
\mathbf{E}\left([W(i_k) - 1 + \chi(W(i_k)) \mid \mathcal{A}(\sigma_t, t)]\right) &= \frac{f(i_k/n)(2 - f(i_k/n))}{2(1 - f(i_k/n))} - 1 + (1 - f(i_k/n)) + o(1) \\
&= \frac{f(i_k/n)^2}{2(1 - f(i_k/n))} + o(1).
\end{aligned}$$

Hence,

$$\prod_{k=1}^r \mathbf{Pr}(\xi_{i_k} = 1 | \mathcal{A}(\sigma_t, t)) = (1 + o(1)) \prod_{k=1}^r \frac{f(i_k/n)^2}{4i_k(1 - f(i_k/n))} + O(n^{-A}),$$

and $o(1)$ can be made independent of $(i_1, i_2, \ldots, i_r)$. Then applying (5.37), (5.40) and (5.42)

$$\begin{aligned}
\mathbf{E}[\Xi(\Xi - 1)\ldots(\Xi - r + 1)] &= (1 + o(1))r! \sum_{(i_1, i_2, \ldots, i_r) \in S'_r} \left(\prod_{k=1}^r \frac{f(i_k/n)^2}{4i_k(1 - f(i_k/n))} \right) + o(1) \\
&= (1 + o(1))r! \sum_{(i_1, i_2, \ldots, i_r) \in S_r} \prod_{k=1}^r \frac{f(i_k/n)^2}{4i_k(1 - f(i_k/n))} + o(1) \\
&= (1 + o(1)) \sum_{n_J \leq i_1, \ldots, i_k \leq n} \prod_{k=1}^r \frac{f(i_k/n)^2}{4i_k(1 - f(i_k/n))} + o(1) \\
&= (1 + o(1)) \left(\sum_{i=n_J}^n \frac{f(i/n)^2}{4i(1 - f(i/n))} \right)^r + o(1)
\end{aligned}$$

(To obtain the second equation from the first we use the fact that $f(x)/(x(1 - f(x))$ is bounded in the range $[1, \alpha]$.)

Note that $n_J/n \to \alpha$, and so

$$\sum_{i=n_J}^n \frac{f(i/n)^2}{4i(1 - f(i/n))} = \int_\alpha^1 \frac{f(x)^2}{4x(1 - f(x))} dx + o(1).$$

This gives that

$$\mathbf{E}[\Xi(\Xi-1)\dots(\Xi-r+1)] = (1+o(1))\left(\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))}dx\right)^r + o(1).$$

Thus, for any fixed integer $r \geq 1$,

$$\lim_{n\to\infty} \mathbf{E}[\Xi(\Xi-1)\dots(\Xi-r+1)] = \left(\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))}dx\right)^r.$$

This means that Ξ (and hence Φ) is asymptotically distributed as a Poisson variable with mean

$$\int_\alpha^1 \frac{f(x)^2}{4x(1-f(x))}dx.$$

The lemma now follows. □

Proof of Lemma 5.4 It is useful to note that as remarked when we defined α, the quantity $c\alpha^2$ is bounded above by a constant less than $2/3$. Note also that from the hypotheses of the lemma, we have $Z(n_J) = c\alpha^3 n(1+o(1))$ and $Y(n_J) = o(n^{0.8})$. We consider a further $h' = \lfloor n^{0.8} \rfloor$ stages after stage n_J. We claim that by that stage, GUC will have arrived at a stage n^* where $Y(n^*) = W(n^*) = 0$. To see this, it is not difficult to check that in these further h' stages, with high probability,

(I) at most $3c\alpha^2 n^{0.8}/2(1+o(1))$ new clauses of size 2 are created by GUC,

(II) at least h' clauses of minimal size are removed by GUC.

(Note that (I) is similar to (5.18) and can therefore be proved similarly.) Since $c\alpha^2 < 2/3$ and $Y(n_J) + W(n_J) = o(n^{0.8})$, it is not possible to have (I) and (II) unless some of the clauses of minimal size removed are of size 3. This shows that with high probability, there is $n^* \geq n_J - h'$ such that $Y(n^*) = W(n^*) = 0$. Note also that similar to (5.17), we have with high probability that between stages n_J and $n_J - h'$, only $O(n^{0.8})$ clauses of size 3 are removed. Thus at stage n^*, we have with high probability that there are $Z(n^*) = c\alpha^3 n(1+o(1))$ clauses of size three remaining, and that there are $n^* = \alpha n(1+o(1))$ variables whose truth values remain unassigned. Since the ratio of number of size three clauses to number of variables at stage n^* is strictly less than $2/3$, we know from part (a) of Theorem 1.1 that the probability that GUC creates an empty clause at and after stage n^* is $o(1)$. It therefore remains to argue that for n between $n' = n_J - h'$ and n_J, GUC creates no empty clauses with probability tending to 1 as $n \to \infty$. To do this, note that as in (I) above, we have with high probability that

$$Y(\nu) = O(n^{0.8}),$$

for all ν between n' and n_J. Since both n' and n_J equal $\Omega(n)$, we have with high probability that $Y(\nu)/\nu = o(1)$ for all $\nu \in [n', n_J]$. As indicated when showing (5.24), we have with high probability that for $\nu \in [n', n_J]$, $W(\nu)$ can be bounded above in distribution by a Markov chain X_n defined in the previous section with one-step transitions governed by a binomial variable with parameters $O(n^{0.8})$ and $1/n'$. Using (3.14) and (3.18) and by following arguments used in showing (5.11) and (5.12), we have with high probability that for all $\nu \in [n', n_J]$, $W(\nu) \leq \log^{11} n$. This in turn gives that

$$\sum_{\nu=n'}^{n_J} \frac{W(\nu)}{\nu} = O(n^{-0.2} \log^{11} n).$$

with high probability. Since the expected number of empty clauses created at stage ν equals $O(\mathbf{E}[W(\nu)/\nu])$ (see definition of $\Delta_{1,1}$), the above equation gives that the expected number of empty clauses created at stages $\nu \in [n', n_J]$ equals $o(1)$. Hence, as $n \to \infty$,

$$\mathbf{Pr}(\text{GUC creates an empty clause at stage } \nu \in [n', n_J]) = o(1). \tag{5.43}$$

This completes our proof of Lemma 5.4. □

6 GUC with backtracking and proof of Theorem 1.2

Since GUC succeeds with probability $1 - o(1)$ when $c < 2/3$, we consider only the case where $2/3 \leq c < c_3$. Note first that empty clauses can only be created by GUC when $N_1(\nu) \neq 0$. As our previous analysis shows, $N_1(\nu)$ behaves like a Markov chain in steady state with a reflecting barrier at 0. Also, given $N_1(\nu)$, the probability that GUC creates an empty clause is at stage ν is $O(N_1(\nu)/\nu)$. By allowing GUC to backtrack when it makes a "mistake", we shall see that a random instance of 3-SAT almost certainly has a satisfiable truth assignment when $c < c_3$.

Consider applying GUC to a 3-SAT problem. With $n_b > n_e$, we use $[n_b, n_e]$ to denote a "run" in which $N_1(\nu)$ is non-zero. That is, a run $[n_b, n_e]$ is such that $N_1(n_b + 1) = 0$, $N_1(k) > 0$ $(n_b \geq k > n_e)$, and $N_1(n_e) = 0$. We next describe how we allow GUC to backtrack. Recall that $N(\nu)$ is obtained from $N(\nu + 1)$ by setting a literal $x_{\nu+1}$ to 1 at stage $\nu + 1$ (using x_ν to denote the literal that is set to 1 at stage ν, and recall that x_ν is a literal picked randomly from a randomly chosen clause of minimal size). Also, use $\mathcal{S}(\nu)$ to denote the set of clauses at stage ν. Suppose that GUC is in a run with $N_1(n' + 1) = 0$, and $N_1(k) \geq 1$ for $k = n', n' - 1, \ldots, n''$ where $n'' \leq n'$ is the present stage. GUC then sets a literal $x_{n''}$ to 1. The backtracking is performed if the setting of $x_{n''}$ to 1 gives rise to the occurrence of two size one clauses $\{y\}$ and $\{\bar{y}\}$ for some variable y. If this occurs , then GUC is allowed a limited backtracking (see also the failure condition **(B)** later) by resetting the literals $x_{n'+1}, x_{n'}, x_{n'-1}, \ldots, x_{n''}$ to 0. We have to update the set of clauses by
(a) removing all clauses that contain $\bar{x}_k$ $(k = n' + 1, n', \ldots, n'')$ from the set $\mathcal{S}(n' + 1)$ of clauses,
(b) removing all occurrences of x_k $(k = n' + 1, n', \ldots, n'')$ from clauses in the set $\mathcal{S}(n' + 1)$. Hence this new set of clauses becomes $\mathcal{S}(n'' - 1)$ and the algorithm then proceeds as before by choosing a literal $x_{n''-1}$ and setting it to 1 to obtain $\mathcal{S}(n'' - 2)$. Stages $n'' - 2, n'' - 3, \ldots$ are carried out similarly as before. We call this algorithm GUCB. We say that GUCB fails if:

(A) An empty clause is created in the backtracking when resetting the truth values of some literals to 0, or

(B) It creates an empty clause in a stage after a backtracking and before the next time when the number of size one clauses becomes zero i.e. two separate occurrences of empty clauses in one run.

We use $\hat{N}(\nu) = (\hat{N}_0(\nu), \hat{N}_1(\nu), \hat{N}_2(\nu), \hat{N}_3(\nu))$ to denote the state of GUCB at stage ν when applied to a random instance of 3-SAT. With n' and n'' defined as above, we claim that at stage $n''-1$, the set $\mathcal{S}(n''-1)$ of clauses remains uniformly random.

<u>**Claim**</u>. If $V_{n''-1}$ is the set of variables whose truth values remain unassigned at stage $n''-1$, then for $i = 1, 2, 3$, a size i clause in $\mathcal{S}(n''-1)$ is equally likely to be any clause in $\mathcal{C}_i(V_{n''-1})$.
Proof Let C be a clause of size s in $\mathcal{S}(n'+1)$. Note that $s \geq 2$. It is clear that if $C \cap \{x_i, \bar{x}_i\} = \emptyset$ for all $i = n'+1, n', \ldots, n''$, then C is equally likely to be any clause in $\mathcal{C}_s(V_{n''-1})$. On the other hand, if $C \cap \{x_i, \bar{x}_i\} \neq \emptyset$ for some $i = n'+1, n', \ldots, n''$, then let j be the greatest value of such i's. If $\bar{x}_j \in C$, then no sub-clause of C is in $\mathcal{S}(n''-1)$ by definition of $\mathcal{S}(n''-1)$. If $x_j \in C$, then $C_1 = C - \{x_j\}$ is equally likely to be any clause in the set of all clauses with size $|C_1|$ made up of variables whose truth values remain unassigned immediately after stage j. Now since C contains x_j, C is not considered by GUCB until backtracking. During the backtracking, C is removed from $\mathcal{S}(n'+1)$ if C contains $\bar{x}_i$ for some $i = j-1, j-2, \ldots, n''$. Otherwise $C_2 = C - \{x_{n'+1}, x_{n'}, \ldots, x_{n''}\}$ is in $\mathcal{S}(n''-1)$, but then C_2 is equally likely to be any clause of size $|C_2|$ made up of variables in $V_{n''-1}$. □

Hence the behaviour of GUCB can be analysed by considering $\hat{N}(\nu)$. As before, we shall allow GUCB to continue after empty clauses are created, that is, we allow GUCB to continue even when it fails in cases (A) and (B) above. We shall show that the probability that GUCB fails is $o(1)$. This is done by showing that the effect of backtracking on $\hat{N}$ is negligible, and that with high probability, there are at most $\log^5 n$ times when GUCB backtracks. Note that we make no attempt to minimize the powers of $\log n$ in this section.

To minimize subscripts, we write $\hat{W}(\nu)$ for $\hat{N}_1(\nu)$, $\hat{Y}(\nu)$ for $\hat{N}_2(\nu)$ and $\hat{Z}(\nu)$ for $\hat{N}_3(\nu)$. Recall that

$$f(x) = \frac{3c}{4}(1-x^2) + \log x, \quad x \in (0,1).$$

The constant α is defined to be the unique root of $f(x) = 0$ within the range $(0,1)$, and $\alpha_0 = \alpha + n^{-0.24}$. Also, the integer J is defined as the greatest integer such that $n - Jh \geq \alpha_0 n$, where $h = \lfloor n^{1/2} \rfloor$. We next define some new quantities. Let $b_0 = n+1$, $l_0 = n+1$ and $f_0 = n+1$. For integers $1 \leq i \leq \log^5 n_0$, if GUCB backtracks for at least i times before stage n_J, then define b_i, l_i, f_i so that b_i equals the stage number at which GUCB backtracks for the i-th time, l_i equals the greatest integer $k \leq b_i$ such that $\hat{W}(k) = 0$, and f_i equals the smallest integer $k \geq b_i$ such that $\hat{W}(k+1) = 0$; if GUCB backtracks for less than i times before stage n_J, then define $b_i = b_{i-1}$, $l_i = l_{i-1}$ and $f_i = f_{i-1}$. (That is, $[f_i \searrow l_i]$ is essentially a "run" corresponding to GUCB in which the backtracking takes place at stage b_i). We shall use induction to show that with high probability, we have for all $i \leq \log^5 n_0$ that

$$\hat{Z}(b_i - 1) \quad = \quad cb_i^3/n^2 + O(in^{3/4}\log n), \tag{6.1}$$

$$\begin{aligned} \hat{Y}(b_i-1) &= b_i f(b_i/n) + O(i^2 n^{3/4}\log n), && (6.2)\\ \hat{W}(b_i-1) &= O(\log^4 n), && (6.3) \end{aligned}$$

where the constants in the big O terms are independent of i. Note that the quantities $\hat{Z}(b_i-1), \hat{Y}(b_i-1), \hat{W}(b_i-1)$ respectively are the numbers of size three, size two, size one clauses immediately after the backtracking at stage b_i. When proving the above estimates using induction, it is convenient to show at the same time the following estimates that for $i \leq \log^5 n$,

$$\begin{aligned} \mathbf{Pr}(\text{GUCB creates an empty clause at stage } j \in [b_i - 1 \searrow l_i + 1]) &= O(\log^8 n/n), && (6.4)\\ \mathbf{Pr}(\text{GUCB creates an empty clause at stage } b_i) &= O(\log^6 n/n). && (6.5) \end{aligned}$$

That (6.1 - 6.3) hold for $i = 0$ is trivial. We assume therefore that they hold for i, and show that (6.1 - 6.3) remain valid for $i+1$. Note that after stage b_i, GUCB behaves like GUC until the next backtracking. Therefore, consider applying GUC to a random instance $\mathcal{I}$ of a satisfiability problem on $b_i - 1$ variables with $\hat{Z}(b_i-1)$ size three clauses, $\hat{Y}(b_i-1)$ size two clauses and $\hat{W}(b_i-1)$ size one clauses. Use $Z(j), Y(j)$ and $W(j)$ to denote the numbers of size three, size two and size one clauses at stage $j \leq b_i - 1$. Also, for $j \leq b_i - 1$, use τ_j to denote the minimum value of $k \geq 1$ such that $W(j-k) = 0$. Note that until the next backtracking at stage b_{i+1}, we have $\hat{Z} = Z$, $\hat{Y} = Y$ and $\hat{W} = W$.

Note that the values of Z, Y, W satisfy the hypotheses of Lemma 5.1. Thus, we pply (5.9) and (5.11) to obtain that with high probability,

$$\begin{aligned} b_i - l_i &= O(\log^4 n), && (6.6)\\ W(j) &= O(\log^4 n), \quad \text{for all } j \in [b_i - 1, l_i]. && (6.7) \end{aligned}$$

We therefore have with high probability that

$$\sum_{j=b_i-1}^{l_i+1} \frac{W(j)}{j} = O(\log^8 n/n).$$

Hence, the expected number of empty clauses created in stages $j \in [b_i - 1 \searrow l_i + 1]$ equals $O(\log^8 n/n)$ (please refer to comments before (5.43)). Equation (6.4) now follows.

Next, we apply Lemma 5.1 to obtain that with high probability

$$\begin{aligned} Z(n') &= cn'^3/n^2 + O(in^{3/4}\log n + n^{1/4}\log n),\\ Y(n') &= n' f(n'/n) + O(i^2 n^{3/4}\log n + (i+1)n^{1/4}\log n),\\ W(n') &\leq \log^2 n, \end{aligned}$$

where $n' = b_i - 1 - h$. These estimates satisfy the hypotheses of Lemma 5.1. Therefore, if $n' \geq \alpha_0 n$, we may apply Lemma 5.1 repeatedly. Since we need only apply Lemma 5.1 at most $O(n^{1/2})$ times before we go past the stage $\lfloor \alpha_0 n \rfloor$, we have by using (5.7), (5.8), (5.11), (5.13), (5.10) and (5.14) that with high probability,

$$\begin{aligned} Z(j) &= cj^3/n^2 + O((i+1)n^{3/4}\log n), && (6.8)\\ Y(j) &= jf(j/n) + O((i+1)^2 n^{3/4}\log n), && (6.9)\\ W(j) &= O(\log^2 n) && (6.10)\\ \tau_j &= O(\log^2 n), && (6.11) \end{aligned}$$

for all $j \in [l_i \searrow n_J]$. Note that if there are at most i backtrackings before stage n_J, then (6.1 - 6.3) remain valid for $i+1$. Otherwise, we have $l_i > f_{i+1} \geq b_{i+1} \geq n_J$ by definitions of f_{i+1} and b_{i+1}. Therefore, using the above estimates, we have with high probability that

$$\begin{aligned}
Z(b_{i+1}) &= c{b_{i+1}}^3/n^2 + O((i+1)n^{3/4}\log n), && (6.12)\\
Y(b_{i+1}) &= b_{i+1}f(b_{i+1}/n) + O((i+1)^2 n^{3/4}\log n), && (6.13)\\
Z(f_{i+1}+1) &= c{f_{i+1}}^3/n^2 + O((i+1)n^{3/4}\log n), && (6.14)\\
Y(f_{i+1}+1) &= f_{i+1}f(f_{i+1}/n) + O((i+1)^2 n^{3/4}\log n_0). && (6.15)
\end{aligned}$$

Note that from (6.11), we have with high probability that the length of every "run" equals $O(\log^2 n)$ in the entire history when GUC is applied to a random instance $\mathcal{I}$ defined above. Thus, when GUCB backtracks at stage b_{i+1}, we have with high probability that GUCB need only reset the truth values of $v = O(\log^2 n)$ variables. Also, we have with high probability that

$$f_{i+1} - b_{i+1} = O(\log^2 n). \qquad (6.16)$$

We next show that the backtracking does not change the numbers of size three and size two clauses by much. Note first that by (6.12 - 6.16), we have

$$\begin{aligned}
Z(f_{i+1}+1) &= c{b_{i+1}}^3/n^2 + O((i+1)n^{3/4}\log n), && (6.17)\\
Y(f_{i+1}+1) &= b_{i+1}f(b_{i+1}/n) + O((i+1)^2 n^{3/4}\log n). && (6.18)
\end{aligned}$$

Recall that in the backtracking at stage b_{i+1}, GUCB resets the truth values of $v = O(\log^2 n)$ variables and obtain the set of clauses at stage $b_{i+1}-1$ by updating the set $\mathcal{S}(f_{i+1}+1)$ of clauses at stage $f_{i+1}+1$. We next observe that in the initial set of $\lfloor cn \rfloor$ (random) clauses of size three, the number of clauses containing a given literal is distributed as $B(m, 3/n)$. Thus, we have with high probability that for any literal x, the number of clauses containing x equals $O(\log^2 n)$. Hence, with high probability, the number of size three clauses in $\mathcal{S}(f_{i+1}+1)$ containing (at least) one of the v variables is $O(\log^4 n)$. This gives that with high probability,

$$\hat{Z}(b_{i+1}-1) - \hat{Z}(f_{i+1}+1) = O(\log^4 n). \qquad (6.19)$$

For size two clauses, we note first that at most $\hat{Z}(b_{i+1}-1) - \hat{Z}(f_{i+1}+1)$ clauses of size two are added to $\mathcal{S}(f_{i+1}+1)$. Also, similar to (6.19), we have with high probability that at most $O(\log^4 n)$ size two clauses are removed from $\mathcal{S}(f_{i+1}+1)$ in the backtracking. Therefore, we have with high probability that

$$\hat{Y}(b_{i+1}-1) - \hat{Y}(f_{i+1}+1) = O(\log^4 n). \qquad (6.20)$$

Similarly, it is easy to see that with high probability, at most $O(\log^4 n)$ clauses of size 1 are created from clauses of size two and size three in $\mathcal{S}(f_{i+1}+1)$. We thus have with high probability that

$$\hat{W}(b_{i+1}-1) = O(\log^4 n). \qquad (6.21)$$

The induction proof of (6.1 - 6.3) is now complete by noting that (6.1 - 6.3) follow from (6.17 - 6.21) and the fact that $\hat{Z}(f_{i+1}+1) = Z(f_{i+1}+1), \hat{Y}(f_{i+1}+1) = Y(f_{i+1}+1)$.

We next would like to show (6.5). Let $I = \{b_i, b_i{+}1, \ldots, f_i, f_i{+}1\}$ and use V_b to denote the set of variables whose truth values remain unassigned immediately before stage $b_i - 1$. For $j \in I$, use x_j to denote the literal that was set to 1 at stage j. (Note that $v = f_i - b_i + 2 = O(\log^2 n)$.) Now in the backtracking at stage b_i, GUCB resets these v literals to 0 and update the set $\mathcal{S}(f_i + 1)$ of clauses. For $j \in I$, let $\mathcal{S}_j^{(i)}$ be the set of clauses of size i in the set $\mathcal{S}(j)$ of clauses at stage j containing the literal x_j. That is,

$$\mathcal{S}_j^{(i)} = \{C \in \mathcal{S}(j) \mid x_j \in C \text{ and } |C| = i\}.$$

Note that if $C \in \mathcal{S}_j^{(1)}$, then C must come from a clause $C' \in \mathcal{S}(f_i + 1)$ where C' contains a literal $\bar{x}_{j'}$, for some $j' \in I$ and $j' > j$. Thus, no clause in $\cup_{j \in I} \mathcal{S}_j^{(1)}$ can become an empty clause during backtracking. Note also that if $C \in \mathcal{S}_j^{(i)}$ $(i = 2, 3)$, then the entire clause C is removed from $\mathcal{S}_j$ at stage j, and so no sub-clause of C can appear in $\mathcal{S}_{j'}^{(2)} \cup \mathcal{S}_{j'}^{(3)}$ for all $j' \in I$ and $j' < j$. Thus, if $C \in \mathcal{S}_j^{(i)}$ $(i = 2, 3)$, then *during backtracking*, $C - \{x_j\}$ is equally likely to be a size $i - 1$ clause chosen from the set

$$C_{i-1}(V_b \cup \{\hat{x}_{j'} \mid j' \in I \text{ and } j' < j\}),$$

where $\hat{x}$ here denotes the variable of the literal x. Thus, if $C \in \mathcal{S}_j^{(i)}$ $(i = 2, 3)$, then the probability that C becomes an empty clause after the backtracking is $O(v/b_i) = O(\log^2 n/n)$. Note that for a clause $C \in \mathcal{S}(f_i + 1)$ to become an empty clause after backtracking, the clause C must be contained in $\cup_{j \in I} \cup_{i=2,3} \mathcal{S}_j^{(i)}$. As argued in (6.19) and (6.20), the size of $\cup_{j \in I} \cup_{i=2,3} \mathcal{S}_j^{(i)}$ is $O(\log^4 n)$. Hence the probability that an empty clause is created in the backtracking at stage b_i equals $O(\log^6 n/n)$. This proves (6.5).

It now follows from (6.4) and (6.5) that

$$\begin{aligned} & \mathbf{Pr}(\text{GUCB creates an empty clause at stages } j \in [b_i, l_i + 1], \text{ for some } i \leq \log^5 n) \\ = \; & O(\log^{13} n/n). \end{aligned}$$

Therefore, it remains to show that

$$\mathbf{Pr}(\text{GUCB backtracks at least } \log^5 n \text{ times before stage } n_J) = o(1), \tag{6.22}$$

and that

$$\mathbf{Pr}(\text{GUCB backtracks at and after stage } n_J) = o(1). \tag{6.23}$$

To show (6.22), suppose that l_i is given and note that GUCB behaves like GUC after each l_i until the next backtracking at stage b_{i+1}. Note that using (6.8) and (6.9), we have with high probability that for $i \leq \log^5 n$,

$$\begin{aligned} \hat{Z}(l_i) &= c{l_i}^3/n^2 + O(n^{3/4} \log^6 n), \\ \hat{Y}(l_i) &= l_i f(l_i/n) + O(n^{3/4} \log^{11} n). \end{aligned}$$

Also, $\hat{W}(l_i) = 0$. Next, consider applying GUC to a random satisfiability problem $\mathcal{I}'$ with l_i variables, $Z'(l_i)$, $Y'(l_i)$ and $W'(l_i)$ clauses of size three, two and one respectively, where

$Z'(l_i) \geq \hat{Z}(l_i)$, $Y'(l_i) \geq \hat{Y}(l_i)$ and $W'(l_i) \geq \hat{W}(l_i)$. Then by the monotonicity argument used in showing Theorem 1.1(c), we have $\hat{W}(j) \leq W'(j)$ for $j \geq b_{i+1}$. Thus, if b' is the minimum value of $\nu \leq l_i$ such that when GUC is applied to $\mathcal{I}'$, the set of clauses at stage ν contains two clauses $\{y\}, \{\bar{y}\}$ for some y, then it is easy to see that $b_{i+1} \leq b'$ in distribution. We apply this idea with Z', Y', W' obtained by applying GUC to a random instance $\mathcal{I}'$ of 3-SAT with $\lfloor c'n \rfloor$ clauses of size 3, where $c' \in (c, c_3)$. Note that by definitions of l_i and c', we have $l_i \geq n_J \geq \alpha_0' n$ (where α_0' is defined as α_0 but with c replaced by c'). Thus, we apply Lemmas 5.1 and 5.2 to obtain that with high probability, the numbers of size three, size two and size one clauses with respect to $\mathcal{I}'$ satisfy that for $i \leq \log^5 n$,

$$\begin{aligned} Z'(l_i) &= c' {l_i}^3/n^2 + O(n^{3/4}\log^6 n), \\ Y'(l_i) &= l_i g(l_i/n) + O(n^{3/4}\log^{11} n), \\ W'(l_i) &= O(\log^2 n), \end{aligned}$$

where $g(x) = 3c'(1-x^2)/4 + \log x$. Let N' be the number of stages ν before n_J such that in applying GUC to $\mathcal{I}'$, the set of clauses at stage ν contains two clauses $\{y\}, \{\bar{y}\}$ for some y. Since $Z'(l_i) \geq \hat{Z}(l_i)$, $Y'(l_i) \geq \hat{Y}(l_i)$ and $W'(l_i) \geq \hat{W}(l_i)$ with high probability, it follows (by considering the waiting times b' defined above) that

$$\begin{aligned} & \mathbf{Pr}(\text{GUCB backtracks at least } \log^5 n \text{ times before stage } n_J) \\ \leq \ & \mathbf{Pr}(N' \geq \log^5 n) + o(1). \end{aligned}$$

Using (5.13), we see that when GUC is applied to $\mathcal{I}'$, we have with high probability that for all $j \geq n_J$, the number $W'(j)$ of size one clauses at stage j is $O(\log^2 n)$. Therefore, the probability that there is a contribution to N' at stage j equals $O(\mathbf{E}[W'(j)^2/j])$. Since $W'(j) = O(n)$, we have $\mathbf{E}[W'(j)] = O(\log^2 n)$, and hence

$$\mathbf{E}[N'] = O(\log^4 n).$$

It therefore follows that

$$\mathbf{Pr}(N' \geq \log^5 n) = O(1/\log n).$$

This shows (6.22).

To show (6.23), we have from (6.8 - 6.10) again that with high probability

$$\begin{aligned} \hat{Z}(n_J) &= c{n_J}^3/n^2 + O(n^{3/4}\log^6 n), \\ \hat{Y}(n_J) &= n_J f(n_J/n) + O(n^{3/4}\log^{11} n), \\ \hat{W}(n_J) &= O(\log^4 n). \end{aligned}$$

(Note that in the unlikely event where $n_J \in [b_i - 1, l_i]$ for some i, we may apply (6.1 - 6.3) and (6.6 - 6.7) to obtain the above estimates at stage n_J.) These values of $\hat{Z}, \hat{Y}, \hat{W}$ satisfy the hypotheses of Lemma 5.4. Thus, we obtain (6.23) from Lemma 5.4. Our proof of Theorem 1.2 is thus complete.

7 Proof of Theorem 1.3

We shall only give a sketch proof here. Consider SC when applied to a random instance of k-SAT with n variables and $m = \lfloor cn \rfloor$ clauses. We restrict our attention to

$$c > \left(\frac{k-1}{k-3}\right)\frac{k-1}{k-2}\frac{2^{k-3}}{k},$$

for otherwise SC succeeds with probability $1 - o(1)$ (see Chvátal and Reed [4]). Let $q_i(\nu)$ be the probability that a randomly selected clause from $\mathcal{C}_k(V_n)$ is of size i immediately before stage ν. It is not difficult to check that for $i = 3, \ldots, k$,

$$q_i(\nu) = \frac{\binom{n-\nu}{k-i}\binom{\nu}{i}}{\binom{n}{k}}2^{-(k-i)}.$$

Let $N_i'(\nu)$ be the number of size i clauses at stage ν. The above equation implies that with high probability, we have for $i = 3, \ldots, k$ that

$$N_i'(\nu) = \binom{k}{i}\frac{cn}{2^{k-i}}(\nu/n)^i(1-\nu/n)^{k-i} + O(n^{1/2}\log n), \tag{7.1}$$

whenever $\nu = \Omega(n)$. This gives a fairly accurate estimate for $N_3'(\nu)$ in particular.

Fix a (small) constant $\epsilon > 0$. Recall that β_1 is the largest root of the equation

$$p_3(x) = \binom{k}{3}cx^2(1-x)^{k-3}2^{-(k-3)} = 2/3.$$

Let $\beta_1' = \beta_1 + \epsilon$ and $\beta_1'' = \beta_1 - \epsilon$. Note that $N_1'(\nu-1) + N_2'(\nu-1) - N_1'(\nu) - N_2'(\nu)$ is bounded above by

$$\begin{cases} \Delta_{3,1}'(\nu), & \text{if } N_1'(\nu) + N_2'(\nu) = 0, \\ \Delta_{3,1}'(\nu) - 1, & \text{otherwise,} \end{cases}$$

where $\Delta_{3,1}'(\nu)$, defined in Section 2, is the number of new size 2 clauses created at stage ν. Since $\Delta_{3,1}'(\nu)$ is a binomial variable with parameters $N_3'(\nu)$ and $3/2\nu$ and since for $\nu \geq \beta_1' n$,

$$N_3'(\nu)\frac{3}{2\nu} = \frac{3c}{2}\binom{k}{3}(\nu/n)^2(1-\nu/n)^{k-3}2^{-(k-3)} + O(n^{-1/2}\log n) < 1$$

with high probability, it follows from Lemma 3.4 (see also proof of (5.13)) that for $\nu \geq \beta_1' n$,

$$N_1'(\nu) + N_2'(\nu) = O(\log^2 n)$$

with high probability. This gives an upper bound for $N_2'(\nu)$ which in turn gives that with high probability,

$$\sum_{n \geq \nu \geq \beta_1' n} N_1'(\nu) = O(\log^2 n).$$

The expected number of empty clauses created before stage $\beta_1'n$ thus equals $O(\log^2 n/n) = o(1)$. Hence

$$\lim_{n\to\infty} \mathbf{Pr}(\text{SC fails at or before stage } \beta_1'n) = 0. \tag{7.2}$$

Furthermore, for ν between $\beta_1''n$ and $\beta_1'n$, it is not difficult to obtain that there is $\gamma_1(\epsilon)$ which tends to 0 as $\epsilon \to 0$ such that

$$N_2'(\nu) \le \gamma_1(\epsilon)n \tag{7.3}$$

with high probability. This gives an upper bound for $N_1'(\nu)$ and it is not difficult to obtain in a similar (but simpler) fashion as our proof of Theorem 1.1(b) that there is $\gamma_2(\epsilon)$ where $\gamma_2(\epsilon) \to 0$ as $\epsilon \to 0$ such that

$$\lim_{n\to\infty} \mathbf{Pr}(\text{SC fails at a stage between } \beta_1''n \text{ and } \beta_1'n) \le \gamma_2(\epsilon). \tag{7.4}$$

Suppose we allow SC to have limited backtracking (as in GUC described in the previous section). Then in view of (7.2) and (7.4), the theorem follows from the following lemma.

Lemma 7.1 *For all small $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{Pr}(\text{SCB fails at or after stage } \beta_1''n) = 0.$$

We do not prove Lemma 7.1. Instead, we give a sketch proof of Lemma 7.2 below. ($c < c_k$ means that Lemmas 7.1 and 7.2 can be proved similarly.)

Lemma 7.2 *Let $n_0 = \lfloor \beta_1 n \rfloor$ and V be a set of n_0 variables. Let $\mathcal{I}$ be a random formula with $\hat{N}_i(n_0)$ clauses of size i, where for $i = 3, \ldots, k$,*

$$\hat{N}_i(n_0) = \binom{k}{i} cn(n_0/n)^i(1-n_0/n)^{k-i} + O(n^{1/2}\log n)$$

and $\hat{N}_i(n_0) = 0$ for $i = 0, 1, 2$. Each size i clause in $\mathcal{I}$ is chosen at random (with equal probability) and independently from $\mathcal{C}_i(V)$. Then

$$\lim_{n\to\infty} \mathbf{Pr}(\text{SCB, applied to } \mathcal{I}, \text{ fails at or after stage } n_0) = 0.$$

This lemma can be proved in a way similar to our proof of Theorem 1.2. The key point is that when SC (without backtracking) is applied to $\mathcal{I}$, we can follow our proof of (5.5) to obtain an estimate for the number $N_2'(\nu)$ of clauses of size two. Indeed, if $h = \lfloor n^{1/2} \rfloor$, $n_i = n_0 - ih$, $I_i = \{n_i + 1, \ldots, n_{i-1}\}$ and J is the greatest integer such that $n_0 - Jh \ge \alpha n + n^{0.76}$ where α is defined later, then we have with high probability that

$$\begin{aligned} N_2'(n_i) &= \frac{kcn_i}{2^{k-1}} \left[(1+(k-2)n_i/n)(1-n_i/n)^{k-2} - (1+(k-2)\beta_1)(1-\beta_1)^{k-2}\right] \\ &\quad + n_i \log(n_i/(\beta_1 n)) + O(in^{1/4}\log n), \end{aligned} \tag{7.5}$$

which can be proved using induction and difference equations as in Lemma 5.1. Intuitively, the above equation can be obtained as follows. Let

$$p_2(x) = \frac{kc}{2^{k-1}}(1+(k-2)x)(1-x)^{k-2} + \log x.$$

Note that $p_2(x) - p_2(\beta_1)$ is an approximation to $N_2'(\lfloor xn \rfloor)/\lfloor xn \rfloor$ according to (7.5). We define $\alpha < \beta_0$ as the smallest number so that $p_2(x) - p_2(\beta_1) = 0$. Note also that

$$\frac{dp_2}{dx} = \frac{1}{x}\left(-\frac{3}{2}p_3(x) + 1\right). \tag{7.6}$$

Thus $p_2(x)$ is maximized when $x = \beta_0$. Note that

$$p_2(\beta_0) - p_2(\beta_1) = \frac{1}{(k-1)(k-2)}\left(\frac{1}{\beta_0^2} + \frac{k-3}{\beta_0} - \frac{1}{\beta_1^2} - \frac{k-3}{\beta_1}\right) + \ln(\beta_0/\beta_1),$$

which is less than 1 according to the hypothesis of the theorem. Thus, taking (7.5) as induction hypothesis, we see that $N_2'(n_i)/n_i$ is, with high probability, at most a constant which is less than 1. This means that we can apply the results in Section 2 to approximate $N_1'(\nu)$, and in particular obtain that (see π_0 before (3.2))

$$\mathbf{Pr}(N_1'(\nu) = 0) \approx 1 - N_2'(\nu)/\nu.$$

This shows that

$$\begin{aligned} \mathbf{E}[N_2'(\nu-1) - N_2'(\nu)] &\approx \mathbf{E}[\Delta_{3,1}'(\nu) - \Delta_{2,0}'(\nu)] - \mathbf{Pr}(N_1'(\nu) = 0) \\ &\approx \frac{3}{2\nu}\mathbf{E}[N_3'(\nu)] - \frac{1}{\nu}\mathbf{E}[N_2'(\nu)] - 1. \end{aligned}$$

Putting $\varphi(x) = \mathbf{E}[N_2'(\lfloor xn \rfloor)/\lfloor xn \rfloor)$, we have for small $h > 0$ that

$$\begin{aligned} & \varphi(x-h) - \varphi(x) \\ \approx\ & (1 + h/x + O(h^2))\frac{1}{xn}\mathbf{E}[N_2'(\lfloor xn - hn \rfloor)] - \frac{1}{xn}\mathbf{E}[N_2'(\lfloor xn \rfloor)] \\ \approx\ & \frac{1}{xn}\left(\mathbf{E}[N_2'(\lfloor xn - hn \rfloor) - \mathbf{E}[N_2'(\lfloor xn \rfloor)]\right) + \frac{h}{x^2 n}\mathbf{E}[N_2'(\lfloor xn \rfloor)] + O(h^2) \\ \approx\ & \frac{h}{x}\left(\frac{3}{2}p_3(x) - 1\right). \end{aligned}$$

So $\varphi(x)$ should stay close to the solution of the differential equation (7.6). The induction proof of (7.5) is completed by showing that $N_2'(n_{i+1}) - N_2'(n_i)$ is close to its mean.

It can be shown that the Claim in Section 5 remains true for SCB when applied to $\mathcal{I}$. That is, the set of clauses after each (limited) backtracking remains uniformly random. Therefore, our proof of (6.22) and the statement before it can be extended to show that

$$\mathbf{Pr}(\text{SCB, applied to } \mathcal{I}, \text{ fails at a stage between } n_J \text{ and } n_0) = o(1). \tag{7.7}$$

It therefore remains to show that

$$\mathbf{Pr}(\text{SCB, applied to } \mathcal{I}, \text{ backtracks at and after stage } n_J) = o(1). \tag{7.8}$$

Proving (7.8) requires a result similar to Lemma 5.4. Since the backtracking in SCB does not change $N_i'(\nu)$ by much, we have in particular estimates for $\hat{N}_i(n_J)$ (similar to those given in (7.1) and (7.5)). Thus as in the proof of Lemma 5.4, there is (with high probability) $n^* \approx \alpha n$ such that $\hat{N}_1(n^*) = \hat{N}_2(n^*) = 0$ and that for $i = 3, \ldots, k$ and for $\nu \leq n^*$, $\hat{N}_i(\nu)$ can be approximated by estimates similar to those given in (7.1). Note that for $\nu < n^*$, $N_3(\nu)/\nu$ is less than a constant which is less than 2/3. Thus similar to (7.2) and (7.4), we have (7.8).

8 Other Models

We observe that repacing $m = \lfloor cn \rfloor$ by $m = \lfloor (c + o(1))n \rfloor$ yields exactly the same results above.

(a) Suppose we allow $x, \bar{x}$ in the same clause. Remove such clauses as they are always satisfied. With high probability there are $o(n)$ such clauses and what is left is random.

(b) Suppose we do not allow repetition of the same clause. Remove repetitions and argue as in (a).

(c) Suppose clauses are distinct but unordered, as are the literals in a clause. This follows from (b) as each instance in this model gives rise to the same number $m!(k!)^m$ instances of Model (b).

(d) If we allow a clause to have a repeated literal then this is the same as starting with a few clauses of size $k-1$ (with high probability no smaller clauses will occur). Nothing significant will happen, but one has to check that the analysis is essentially unaffected.

Acknowledgement: We thank Boris Pittel for pointing out errors and providing help on an earlier version of this paper.

References

[1] A.Z. Broder, A.M. Frieze and E. Upfal, On the satisfiability and maximum satisfiability of random 3-CNF formulas, to appear in *SODA 1993*.

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

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

[4] V. Chvátal and B. Reed, Mick gets his (the odds are on his side), *Proceedings of the 33rd IEEE Symposium on Foundations of Computer Science*, (1992) 620-627.

[5] V.Chvátal and E.Szemerédi, *Many hard examples for resolution*,

[6] M. Davis and H. Putnam, A computing procedure for quantification theory, *Journal of the ACM* 7 (1960) 201-215.

[7] A. Goerdt, A threshold for unsatisfiability, to appear in *17th International Symposium on Mathematical Foundations of Computer Science*, Prague, Czechoslovakia, August 1992.

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

[9] A.Kamath, R.Motwani, K.Palem and P.Spirakis, *Why Mick doesn't get any: thresholds for (un)satisfiability*, to appear.

[10] D.Knuth, R.Motwhani and B.Pittel, *stable marriage*

[11] T. Larabee, Evidence for the satisfiability threshold for random 3CNF formulas.

[12] A.El Maftouhi and W.Fernandez de la Vega, *On Random 3-sat*, to appear.

[13] D. Mitchell, B. Selman and H. Levesque, Hard and easy distributions of SAT problems.