Phase Transition in a Random NK Landscape Model

Sung-Soon Choi*
Seoul National University
Shillim-dong, Kwanak-gu
Seoul, 151-742 Korea
sschoi@soar.snu.ac.kr

Kyomin Jung*
MIT
77 Massachusetts Avenue
Cambridge, MA02139, USA
kmjung@mit.edu

Jeong Han Kim
Microsoft Research
One Microsoft Way
Redmond, WA98052, USA
jehkim@microsoft.com

Abstract

An analysis for the phase transition in a random NK landscape model, NK(n, k, z), is given. This model is motivated from population genetics and the solubility problem for the model is equivalent to a random $(k+1)$-SAT problem. Gao and Culberson [20] showed that a random instance generated by NK$(n, 2, z)$ with $z > z_0 = \frac{27-7\sqrt{5}}{4}$ is asymptotically insoluble. Based on empirical results, they conjectured that the phase transition occurs around the value $z = z_0$. We prove that an instance generated by NK$(n, 2, z)$ with $z < z_0$ is soluble with positive probability by providing a polynomial time algorithm. Using branching process arguments, we also reprove that an instance generated by NK$(n, 2, z)$ with $z > z_0$ is asymptotically insoluble. The results show the phase transition around $z = z_0$ for NK$(n, 2, z)$. In the course of the analysis, we introduce a generalized random 2-SAT formula, which is of self interest, and show its phase transition phenomenon.

1 Introduction

1.1 NK Landscape Models

A fitness landscape is a function that assigns each genetic composition (genotype) with the fitness of the expression (phenotype) of the genetic composition in an environment. The fitness landscape sometimes refers to its graphical representation as the word "landscape" indicates. The notion of fitness landscape was first introduced by Wright [43] for the analysis of population genetics. Afterwards, mathematical models to study the evolution on fitness landscape have been proposed by many researchers including Franklin and Lewontin [18], Lewontin [31], Ewens [15], Kauffman and Weinberger [26], and Macken and Perelson [32]. Among them, the *NK model* proposed by Kauffman [25] has attracted considerable attention. The NK model generates fitness landscapes with correlation structures in which we can control the degree of interactions between genes and so, indirectly, the ruggedness and correlation degrees of the landscapes.

An NK landscape is a real-valued function defined on the set of binary n-tuples, $\{0,1\}^n$, which is of the form

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)).$$

It is a summation of local fitness functions f_i's, where each f_i depends on its main variable x_i and the variables in the neighborhood of x_i. Here the neighborhood $\Pi(x_i)$ is a subset of the set

*This work was partially carried in Microsoft Research and partially supported by Institute of Theory and Education for Computing (ITEC) at Seoul National University.

$\{x_1, x_2, \ldots, x_n\} \setminus \{x_i\}$ and its size $|\Pi(x_i)|$ is k. There are two ways to choose the variables in the neighborhood $\Pi(x_i)$, adjacent neighborhood and random neighborhood. In the NK models with adjacent neighborhood, $\Pi(x_i)$ consists of the closest k variables (with a certain tie-break) to the main variable x_i with respect to the indices modulo n. In the NK models with random neighborhood, $\Pi(x_i)$ is composed of the k variables chosen uniformly at random from $\{x_1, x_2, \ldots, x_n\} \setminus \{x_i\}$. Local fitness functions are constructed independently of each other. For each local fitness function, a random value from a probability distribution is assigned for each input. In general, it is independently (or nearly independently) assigned for each of 2^{k+1} inputs and its expectation has small absolute value. In the Kauffman's original model, the uniform distribution between zero and one was used as the underlying distribution for local fitness functions. Later, it has been replaced with various probability distributions in the contexts of analysis and applications, inducing variants of the NK model.

The name, "NK landscape", is after two parameters n and k: n represents the number of genes an organism has and k stands for the number of other genes that affect the contribution of each gene to the overall fitness value of the organism. Generally, the parameter k plays a role in controlling the degree of interactions between genes. The larger the value of k is, the more genes interact one another in constructing the fitness landscape. Consider the case that k is small. Given two genotypes (or assignments) with the identical values for most of the genes, most of f_i's produce the same values for the genotypes. Since the values of f_i's are small relatively to the overall fitness f in absolute value, the two genotypes have similar fitnesses, which implies that the landscape has strong correlation structure. On the other hand, if k is $n-1$, each f_i has (nearly) independent values for the two genotypes, which induces the landscape consisting of 2^n (nearly) independent random values. Through experiments in the original NK model, Kauffman suggested that the ruggedness of the landscape generally increases as k increases [25].

Kauffman [25] further analyzed various features of the original NK model in terms of adaptive walks. Weinberger [39] and Fontana *et al.* [17] carried out more detailed analysis of such walks. The asymptotic properties of the global and local optima in NK landscapes were analyzed in various random NK landscape models. The differences between models are mainly due to the underlying distributions for local fitness functions. Evans and Steinsaltz [14], Durrett and Limic [12], Skellett *et al.* [36], and Kaul and Jacobson [28] [29] used the exponential, negative exponential, uniform, and both of normal and uniform distributions in their works, respectively. Weinberger [40] and, later, Wright *et al.* [42] studied the computational complexities of problems related to NK landscapes. Gao and Culberson [21] showed a treewidth result for NK landscapes in a probabilistic way.

NK models have been used in biology, physics, and so on. In biology, NK models explain evolutions of biological objects including amino acid sequences [26] [27] [32], protein or RNA sequences [38] [5] [16] [17] [34], and molecular quasispecies [13]. NK models have been served as a reference point for understanding the properties of those biological objects. In statistical physics, models of spin-glasses are investigated from the viewpoint of NK models in [39]. The evolution of organizations in a business environment is modeled based on an NK model [30]. NK models have been used as a benchmark for evaluating various encoding schemes and genetic operators on the evolutionary algorithm and

comparing them in the evolutionary computation area [4] [23] [33]. They have been also served as a basis for the design of problem difficulty measures for evolutionary algorithms [24] [37] and the design of epistasis measures [35].

1.2 Previous Results in Phase Transition

Recently, Gao and Culberson [20] proposed two random NK landscape models and provided results about the phase transition in the models. A phase transition in probabilistic combinatorial theory refers to the phenomenon that the probability of a property being satisfied in the random model rapidly changes as the order parameter changes around a certain value. Before describing the two random NK landscape models, we present a decision problem related to the models. Given an NK landscape f, the *solubility problem* is to ask whether there exists an assignment that makes all local fitness function values equal to one. An NK landscape f is called *insoluble* if there is no such assignment. Note that for the problem it is enough to consider the case that local fitness functions have the values of zero or one, since replacing all non-one values with zero yields the same solubility result. Weinberger [40] and Wright *et al.* [42] proved that, while the solubility problem for the NK landscapes with adjacent neighborhood can be solved in polynomial time for a fixed k, the problem for the NK landscapes with arbitrary neighborhood is NP-complete for $k \geq 2$.

To investigate the difficulties of the solubility problems for typical NK landscapes with random neighborhood, Gao and Culberson proposed two random models of NK landscapes, the *uniform probability model* and the *fixed ratio model* inspired by the two random graph models of Erdős-Rényi type, $G(n, p)$ and $G(n, m)$, respectively. In the uniform probability model, the fitness value of each input for a local fitness function is independently assigned to zero with probability p and one with probability $1 - p$. This process is independently repeated for each local fitness function. It was shown that an instance generated by this model is asymptotically insoluble or, if it is soluble, a solution can be found in polynomial time with high probability. However, unless p decreases very rapidly with n, it is easy to see that, with high probability, a random instance has a local fitness function that takes zero values for all inputs. This makes the random instance insoluble with high probability. For this reason, the model is not desirable as a model for representing typical instances.

The fixed ratio model overcomes the drawback of the uniform probability model by fixing the ratio of zero values for each local fitness function. The fixed ratio model $\mathrm{NK}(n, k, z)$ is as follows. The value of z ranges in $[0, 2^{k+1}]$. If z is an integer, for each local fitness function f_i, we choose z tuples of 2^{k+1} possible assignments uniformly at random and independently of other f_j's. Then $f_i = 0$ for those tuples and $f_i = 1$ for the other tuples. If z is not an integer so that $z = \lfloor z \rfloor + h$ $(0 < h < 1)$, we specify the fitness values of $\lfloor (1-h)n \rfloor$ local fitness functions as if they were local fitness functions in $\mathrm{NK}(n, k, \lfloor z \rfloor)$ and those of the rest of the local fitness functions as if they were in $\mathrm{NK}(n, k, \lfloor z \rfloor + 1)$. Another way to specify the fitness values of local fitness functions is that we regard each local fitness function as if it were a local fitness function in $\mathrm{NK}(n, k, \lfloor z \rfloor)$ with probability $1 - h$ and in $\mathrm{NK}(n, k, \lfloor z \rfloor + 1)$ with probability h, independently of all others. For example, if $z = 2 + h$, then each local fitness function has zero values for two random assignments with probability $1 - h$ and for three random assignments

with probability h. This new model is denoted by $\overline{\mathrm{NK}}(n, 2, z)$. It is easy to see that $\overline{\mathrm{NK}}(n, 2, z)$ is essentially the same as $\mathrm{NK}(n, 2, z)$.

For $k = 2$, it was proved [20] that an instance generated by the fixed ratio model with $z > z_0 = \frac{27-7\sqrt{5}}{4} \approx 2.837$ is almost always insoluble, where "an event A_n *almost always* occurs" means that $\lim_{n\to\infty} \Pr[A_n] = 1$. And it was empirically suggested that the instances generated by the model with $z < z_0$ are soluble and the solutions are found in polynomial time with probability close to one. From these, Gao and Culberson conjectured that the phase transition takes place around $z = z_0$ in the fixed ratio model with $k = 2$.

1.3 Contribution and Approach

In this paper, we prove that an instance generated by the fixed ratio model with $z < z_0$ is soluble with positive probability by providing a polynomial time algorithm. This settles the conjecture in an affirmative way. Using branching process arguments, we also reprove that an instance generated by the model with $z > z_0$ is almost always insoluble:

Theorem 1 *If $0 < z < z_0$, then there exists $\alpha > 0$ depending on z such that the probability of $NK(n, 2, z)$ being soluble is at least α as n goes to infinity. If $z > z_0$, then $NK(n, 2, z)$ is almost always insoluble.*

Though it is a very interesting question, we have no idea whether α can be arbitrarily close to 1 or not.

To prove Theorem 1, we reduce the solubility problem of an NK landscape to the $(k + 1)$-SAT problem as in [20]. For given Boolean variables, the variables and their complements are called *literals*. Two literals are *strictly distinct* if their underlying variables are different. A *k-clause* is a disjunction of k strictly distinct literals and a *k-SAT formula* is a conjunction of k-clauses. Given a k-SAT formula F, the k-SAT problem is to ask whether there is a truth assignment satisfying F.

Let an NK landscape $f = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$. For each local fitness function f_i, we construct the $(k + 1)$-clauses with the literals of the main variable and neighborhood variables of f_i such that f_i is equal to zero only for the assignments that do not satisfy one of the clauses. For example, suppose that a local fitness function $f_i(x_i, x_j, x_k)$ has zero value only when (x_i, x_j, x_k) is one of $(0, 0, 0)$, $(0, 1, 0)$, and $(1, 1, 0)$. Then, we construct three 3-clauses $(x_i \vee x_j \vee x_k)$, $(x_i \vee \overline{x_j} \vee x_k)$, and $(\overline{x_i} \vee \overline{x_j} \vee x_k)$ for $f_i(x_i, x_j, x_k)$. We take the conjunction of all the $(k+1)$-clauses obtained from all the f_i's to construct a $(k+1)$-SAT formula F. It is easy to check that f is soluble if and only if F is satisfiable. Thus, it is sufficient to consider the phase transition for the satisfiability of the 3-SAT formula F.

There have been many studies for the phase transition of the satisfiability of the random 3-SAT formula, in which the 3-clauses are chosen independently and uniformly at random [1] [2] [11] [19]. In verifying lower bounds of the threshold, many results were obtained by applying variants of the unit clause algorithm that were first analyzed by Chao and Franco [7] [8]. We will apply a variant of the unit clause algorithm to the 3-SAT formula reduced from the random NK landscape $\mathrm{NK}(n, 2, z)$ in the subcritical region of the phase transition. In Section 2, we describe the unit clause algorithm and

investigate some properties of the reduced 3-SAT formula that should be considered when the unit clause algorithm is applied to it. The properties suggest that it is useful to consider four types of random 2-clauses or random equalities (of truth values of variables).

In Section 3, we introduce a generalized random 2-SAT formula consisting of the random 2-clauses and the random equalities presented in Section 2. It generalizes the well-known random 2-SAT formula in which the 2-clauses are chosen independently and uniformly at random [9] [6]. After a parameter D is introduced, a threshold phenomenon result is obtained: A random 2-SAT formula generated by the model is satisfiable with positive probability if $D < 1$ and almost always unsatisfiable if $D > 1$. It turns out that the threshold is not sharp.

In Section 4, we provide the threshold phenomenon result for the satisfiability of the reduced 3-SAT formula, or equivalently, the proof of Theorem 1. To obtain the result for the subcritical region, we use similar approaches developed in Section 3. For the supercritical region, we introduce another random 2-SAT model, which is similar to the generalized random 2-SAT model presented in Section 2. The formula generated according to the model consists of random 2-clauses resolved from the 3-SAT formula reduced from $\mathrm{NK}(n, 2, z)$.

2 The Unit Clause Algorithm

In the subcritical region, we will apply a variant of unit clause algorithm to the 3-SAT formula $F(n, 2, z)$ reduced from a random instance of $\mathrm{NK}(n, 2, z)$, and show that the algorithm finds a satisfying assignment with positive probability. The 3-clauses in the formula are to be regarded as ordered 3-tuples and (copies of) literals came from main variables are placed in the first coordinate of the corresponding 3-clauses. Those (copies of) literals are called main (copies of) literals.

Now we consider unit clause algorithm (UC). UC takes as input a formula F over n variables and outputs a satisfying assignment of F, or outputs "Cannot determine." UC consists of one loop of n iterations and in each iteration of the loop, UC chooses a literal l contained in a unit clause chosen uniformly at random among all the unit clauses. If there is no unit clause, it chooses a literals l uniformly at random among all the literals not assigned truth values. And it sets l to be true. Then all the clauses containing l are satisfied and all the clauses containing $\overline{l}$ are shortened to the clauses without $\overline{l}$. UC fails to produce a satisfying assignment if and only if a 0-clause, a clause with no literal, is created.

Figure 1 describes the pseudo code of UC. For a literal l, let $var(l)$ be the underlying variable of l. For a set $V = \{x_1, \ldots, x_n\}$ of Boolean variables, let $L(V)$ denote the set of $2|V|$ literals on the variables of V. For $i \geq 0$, let $C_i(t)$ denote the collection of all the i-clauses of F at the end of the t^{th} iteration. When $F = F(n, 2, z)$, $C_3(0)$ is the collection of all the clauses of F and the other $C_i(0)$'s are empty. In general, it is easy to see that

$$C_i(t+1) = \{c |\ \left(c \in C_i(t),\ l \notin c, \text{ and } \overline{l} \notin c\right) \text{ or } (c \wedge \overline{l}) \in C_{i+1}(t)\}.$$

When we apply UC to $F(n, 2, z)$, there are three main distinctive properties to be considered. First, there may be a pair of 3-clauses of the form $(l_1 \vee l_2 \vee l_3)$ and $(\overline{l_1} \vee l_2 \vee l_3)$. If $\bar{l}_3$ is set to be 1, then

UC(F):
  $V \leftarrow \{x_1, x_2, \ldots x_n\}$;
  $S \leftarrow \varnothing$;
  For $t = 0, 1, \ldots, n-1$
    If $|C_1(t)| \neq 0$
      Choose a literal l uniformly at random from $C_1(t)$;
      $V \leftarrow V - \{var(l)\}$;
    Else
      Choose a literal l uniformly at random from $L(V)$;
      $V \leftarrow V - \{var(l)\}$;
    Satisfy all clauses of F containing l;
    Remove $\overline{l}$ from all clauses of F;
    $S \leftarrow S \cup \{l\}$;
  If $C_0(n) = \varnothing$ Output solution S;
  Else Output "Cannot determine.";

Figure 1: Pseudo code of the unit clause algorithm

two clauses $(l_1 \vee l_2)$ and $(\overline{l_1} \vee l_2)$ would be created. The conjunction of the two clauses is equivalent to the unit clause (l_2). So we will regard it as the unit clause. This property is called *sublimation*. If we apply UC without sublimation, UC almost always fails to satisfy F. Note that there are $\Theta(n)$ number of pairs of 3-clauses of the form $(l_1 \vee l_2 \vee l_3)$ and $(\overline{l_1} \vee l_2 \vee l_3)$ in $C_3(0)$. In the process of UC if $\overline{l_2}$ is set to be true, then two clauses $(l_1 \vee l_3)$ and $(\overline{l_1} \vee l_3)$ would be created. Then if $\overline{l_3}$ is set to be true, they would be reduced to a pair of unit clauses (l_1) and $(\overline{l}_1)$. Second, there may be a pair of 3-clauses of the form $(l_1 \vee \overline{l_2} \vee l_3)$ and $(\overline{l_1} \vee l_2 \vee l_3)$. Again, if $\bar{l}_3$ is set to be true, then two clauses $(l_1 \vee \overline{l_2})$ and $(\overline{l_1} \vee l_2)$ would be produced. The conjunction of the two clauses is equivalent to the *equality* $l_1 = l_2$. Third, the main (copies of) literals from different local fitness functions are strictly distinct. This fact turns out to increase the threshold value compared to the case without main literals.

In the process of UC, 2-clauses are produced from the 3-clauses of F. Some pair of 2-clauses will become equalities by the second property. Due to the third property, 2-clauses with main variables will appear so that the literals in their first places are strictly distinct. Similarly, equalities with main variables will appear too. Pairs of two clauses like $(l_1 \vee l_2)$ and $(\overline{l_1} \vee l_2)$ do not appear because of the sublimation property. Motivated by these facts, we will separately consider a generalized random 2-SAT formula consisting of random 2-clauses and equalities, both with and without main variables.

As explained in Section 3, unit clauses consisting of main (copies of) literals and unit clauses consisting of other (copies of) literals have different properties. So we will consider two types of unit clauses. Unit clauses consisting of main literals and the (copies of) literals therein are to be colored red. The other unit clauses and the (copies of) literals therein are to be colored blue. Then Figure 2 is the flow diagram of clauses in the process of UC.

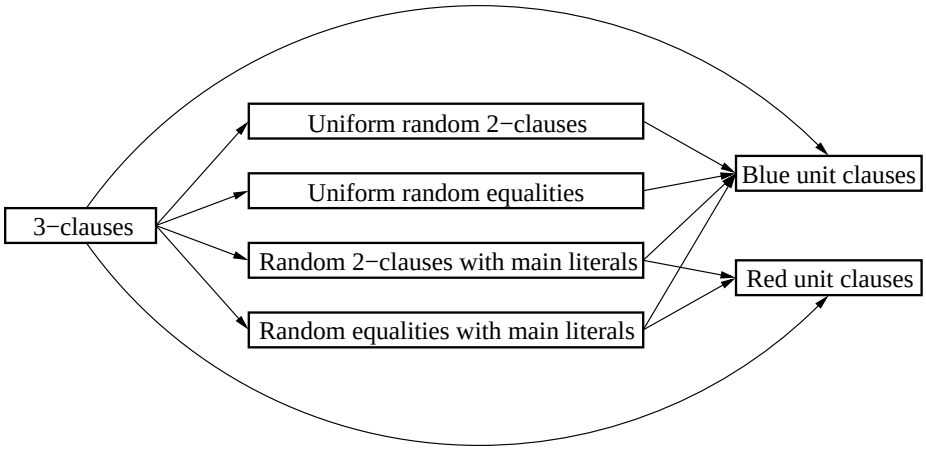

Figure 2: Flow diagram of clauses in the process of UC

3 A Generalized Random 2-SAT Formula

In this section, we define a generalized random 2-SAT formula and examine its satisfiability. As mentioned in Section 2, the generalized random 2-SAT formula has four types of random 2-clauses or equalities. Here 2-clauses and equalities are to be regarded as ordered pairs. The first type consists of typical uniform random clauses, that is, clauses chosen uniformly at random among all the 2-clauses. The second type consists of uniform random equalities over all the literals. The third and fourth types are the same as the first and the second types, respectively, except that the copy of literals in the first places of the clauses or the equalities are pairwise strictly distinct. Those copies of literals are called *main literals*. Let c_1, c_2, c_3 and c_4 be non-negative real numbers with $c_3 + c_4 \leq 1$. Denote $F_i = F_i(n, c_i)$ the conjunction of $c_i n$ 2-clauses or equalities of type i with repetition, $1 \leq i \leq 4$. Denoted by $F(n, c_1, c_2, c_3, c_4)$ is the conjunction of the four random formulae with pairwise strictly distinct main literals.

If $c_2 = c_3 = c_4 = 0$, it is well known [9, 10, 22] that $F(n, c_1, 0, 0, 0)$ is almost always satisfiable if $c_1 < 1$ and almost always unsatisfiable if $c_1 > 1$. It turns out that the parameter

$$D = c_1 + 2c_2 + c_3 + 2c_4 - \frac{(c_3 + 2c_4)^2}{4}$$

plays a similar role in the general case, as D essentially determines the branching ratio. Roughly speaking, the branching ratio is the expected number of unit clauses produced when a literal is set to be true. This is why a variant of UC succeeds with positive probability if $D < 1$, and it almost always fails if $D > 1$. More precisely, we have the following theorem.

Theorem 2 *If $D < 1$, then there exists $\alpha > 0$ depending on c_i's so that the probability of $F(n, c_1, c_2, c_3, c_4)$ being satisfiable is at least α as n goes to infinity. If $D > 1$, then the random formula is almost always unsatisfiable.*

It is worth to note that α may not be close to 1 if $c_2 > 0$ or $c_4 > 0$ as it is not hard to see that $(c_2 + c_4)n$ equalities imply $l = \bar{l}$ for a literal l with positive probability.

Theorem 2, in particular, says that the existence of main literals makes the random formula easier to be satisfied. For example, if there are $0.1n$ uniform random 2-clauses and n random 2-clauses with main literals, then the random formula is satisfiable with positive probability. On the other hand, if there are $1.1n$ uniform random 2-clauses, the random formula is almost always unsatisfiable. If one equality is regarded as its corresponding two 2-clauses then $c_1 + 2c_2 + c_3 + 2c_4$ represents the total number of 2-clauses. The extra term $-(c_3 + 2c_4)^2/4$ is the effect of the existence of main literals.

3.1 Subcritical Region

Now we prove the first part of Theorem 2. Without loss of generality, we may assume $c_1 + c_2 > 0$ and $c_3 + c_4 > 0$. Otherwise, some uniform random 2-clauses or random 2-clauses with main literals might be added to F while the conditions $D < 1$ and $c_3 + c_4 \leq 1$ are kept. Here we define some notations. "At time t" means after t times of iteration of UC, or equivalently, after t literals have been set. Let $V(t)$ denote the set of variables not assigned truth values at time t. For $1 \leq i \leq 4$, let $F_i(t)$ denote the conjunction of remaining 2-clauses or equalities of F_i at time t. Define $|F_i(t)|$ to be the number of 2-clauses or equalities in $F_i(t)$. Let $F(t) = F_1(t) \wedge F_2(t) \wedge F_3(t) \wedge F_4(t)$.

As in Section 2, unit clauses consisting of main (copies of) literals and the main (copies of) literals themselves are colored red. The other unit clauses and the (copies of) literals therein are colored blue. Let $B(t)$ and $R(t)$ denote the set of blue unit clauses and red unit clauses at time t, respectively. Let $V_M(t)$ denote the set of the underlying variables of the main literals of $F_3(t)$ and $F_4(t)$.

As mentioned, we apply a variant of UC that uses a different literal selection policy. Think of UC as an imaginary server whose task is satisfying one unit clause, if any, at each time. We regard $B(t)$ and $R(t)$ as two task queues that the server works for. The server will work for one queue at a time and the queue selection is made randomly with a given probability p, which will be specified later. We call this modified UC *UC with switching server policy (UCS)*. Figure 3 describes the pseudo code of UCS. Note that if $c_3 + c_4 = 1$ and $p < 1$, then UCS may encounter the case that a literal in $L(V(t) - V_M(t))$ must be chosen while $V(t) - V_M(t)$ is empty, which is, of course, impossible. We first consider the case that $c_3 + c_4 < 1$. Let

$$p = \frac{c_1 + 2c_2 + \sqrt{(c_1 + 2c_2)^2 + 2(c_1 + 2c_2)(c_3 + 2c_4)}}{c_1 + 2c_2 + c_3 + 2c_4 + \sqrt{(c_1 + 2c_2)^2 + 2(c_1 + 2c_2)(c_3 + 2c_4)}}. \tag{1}$$

Note that $0 < p < 1$. We defined p so that the expected number of blue unit clauses produced at each time is less than p and the expected number of red unit clauses produced at each time is less than $1 - p$. Using these facts and by a coupling argument, we will show that, with positive probability, no 0-clause is produced until $(1 - \epsilon)n$ variables are assigned truth values, for a small constant $\epsilon > 0$. When $(1 - \epsilon)n$ variables are assigned truth values, the remaining formula is very sparse and it is easy to show that the formula is satisfiable with positive probability.

Note that at each time t, $F(t)$ has the same distribution as $F(n - t, c_1(t), c_2(t), c_3(t), c_4(t))$, where $c_i(t) = |F_i(t)|/(n-t)$. This is a crucial property used in the analysis. The distributions of the numbers of blue and red unit clauses produced at each time highly depend on the sizes of $F_i(t)$'s. So, we first show that $|F_i(t)|$'s are highly predictable using Wormald's theorem [41]. Let $H(t)$ denote the history

UC with switching server policy (F):
For $t = 0, \ldots, n-1$
$\chi(t) \leftarrow 1$ with probability p, $\chi(t) \leftarrow 0$ otherwise;
If $\chi(t) = 1$
If $B(t) \neq \varnothing$
Choose a unit clause (l) uniformly at random from $B(t)$;
Else
Choose a literal l uniformly at random from $L(V(t))$;
If $\chi(t) = 0$
If $R(t) \neq \varnothing$
Choose a unit clause (l) uniformly at random from $R(t)$;
Else
If $V(t) - V_M(t) = \varnothing$ **Exit**;
Choose a literal l uniformly at random from $L(V(t) - V_M(t))$;
Satisfy clauses of F containing l;
Remove all the copies of $\overline{l}$ and sublimate if possible;

Figure 3: Pseudo code of UCS

of $F_i(t)$'s, i.e., the matrix $< \vec{F}(0), \ldots, \vec{F}(t) >$, where $\vec{F}(t) = (F_1(t), \ldots, F_4(t))$. The distributions of $|F_i(t+1)| - |F_i(t)|$ $(1 \leq i \leq 4)$ conditioned on $H(t)$ play important role here.

Lemma 1 *For any small $\epsilon > 0$, we have for all $0 \leq t \leq (1-\epsilon)n$,*

$$\begin{aligned}
\mathrm{E}[|F_1(t+1)| - |F_1(t)||H(t)] &= -\frac{2|F_1(t)|}{n-t}, \\
\mathrm{E}[|F_2(t+1)| - |F_2(t)||H(t)] &= -\frac{2|F_2(t)|}{n-t}, \\
\mathrm{E}[|F_3(t+1)| - |F_3(t)||H(t)] &= -(1+p)\frac{|F_3(t)|}{n-t} + o(1), \\
\mathrm{E}[|F_4(t+1)| - |F_4(t)||H(t)] &= -(1+p)\frac{|F_4(t)|}{n-t} + o(1).
\end{aligned}$$

Proof: Suppose that UCS sets a literal l to be true at time t. Then $|F_1(t+1)| - |F_1(t)| = -X_1$, where X_1 is the number of 2-clauses of $F_1(t)$ that contain l or $\overline{l}$. For a 2-clause $(l_1 \vee l_2)$ of $F_1(t)$, $\Pr[l_1 = l \text{ or } \overline{l_1} = l \text{ or } l_2 = l \text{ or } \overline{l_2} = l] = \frac{2}{n-t}$. And the 2-clauses of $F_1(t)$ are independent from one another. So X_1 has a binomial distribution $\mathrm{Bin}[|F_1(t)|, \frac{2}{n-t}]$. The same argument can be applied to see that $|F_2(t+1)| - |F_2(t)| = -X_2$, where X_2 has a binomial distribution $\mathrm{Bin}[|F_2(t)|, \frac{2}{n-t}]$. Also, $|F_3(t+1)| - |F_3(t)| = -Y_3 - Z_3$, where Y_3 is the number of 2-clauses of $F_3(t)$ whose main literals are equal to l or $\overline{l}$, and Z is the number of 2-clauses of $F_3(t)$ whose second literals are equal to l or $\overline{l}$. Here we divide into two cases according to the value of $\chi(t)$. First, suppose that $\chi(t) = 1$. Then Y_3 has Bernoulli distribution with density $\frac{|F_3(t)|}{n-t}$ since l and $\overline{l}$ are equal to at most one of the main literals of $F_3(t)$. And Z_3 has a binomial distribution $\mathrm{Bin}[|F_3(t)|, \frac{1}{(n-t-1)}]$ or $\mathrm{Bin}[|F_3(t)|-1, \frac{1}{(n-t-1)}]$ according to whether $Y_3 = 0$ or $Y_3 = 1$. Suppose that $\chi(t) = 0$. Then $Y_3 = 0$ because l and $\overline{l}$ cannot be equal to any of main literals of $F_3(t)$. And Z_3 has a binomial distribution $\mathrm{Bin}[|F_3(t)|, \frac{1}{(n-t-1)}]$. The distribution and the expectation of $|F_4(t+1)| - |F_4(t)|$ are obtained in the same way. So the lemma follows. □

Now we state Wormald theorem. A function g is said to satisfy a *Lipschitz condition* on an open set $D_0 \subset \mathbb{R}^{k+1}$ if there exists a constant $L > 0$ such that $|g(u_1, \ldots, u_{k+1}) - g(v_1, \ldots, v_{k+1})| \leq$

$L\sum_{i=1}^{k+1}|u_i - v_i|$, for all $(u_1, \ldots, u_{k+1})$ and $(v_1, \ldots, v_{k+1})$ in D_0.

Theorem 3 (Wormald) *For $1 \le j \le m$, where m is a fixed number, let $Y_j(t)$ (which also depends on n) be a sequence of real-valued random variables such that for all j, all t with $0 \le t \le t_0 = t_0(n)$, and n, $|Y_j(t)| \le C_0 n$ for some constant C_0. Let $H(t)$ denote the history of sequences, i.e. the matrix $< \vec{Y}(0), \ldots, \vec{Y}(t) >$, where $\vec{Y}(t) = (Y_1(t), \ldots, Y_m(t))$. Let D_0 be some bounded connected open set of $\mathbb{R}^{m+1}$ containing the closure of $\{(0, z_1, \ldots, z_m) | z_j = \frac{Y_j(0)}{n}, 1 \le j \le m, \text{ for some } n\}$. Let $g_j : \mathbb{R}^{m+1} \to \mathbb{R}$, $1 \le j \le m$, and suppose that the followings are true for some $t_0 = t_0(n)$.*
(i) For all j and uniformly over all $0 \le t < t_0$,

$$\mathrm{E}[Y_j(t+1) - Y_j(t) | H(t)] = g_j(t/n, Y_1(t)/n, \ldots, Y_m(t)/n) + o(1).$$

(ii) For all j and uniformly over all $0 \le t < t_0$,

$$\Pr[|Y_j(t+1) - Y_j(t)| > n^{\frac{1}{5}} | H(t)] = o(n^{-3}).$$

(iii) For each j, g_j is continuous and satisfies a Lipschitz condition on D_0.
Then the followings hold.
(a) For $(0, \hat{z}^{(1)}, \ldots, \hat{z}^{(m)}) \in D_0$ the system of differential equations

$$\frac{dz_j}{ds} = g_j(s, z_1, \ldots, z_m), 1 \le j \le m$$

has a unique solution in D_0 for $z_j : \mathbb{R} \to \mathbb{R}$ passing through $z_j(0) = \hat{z}^{(j)}$, $1 \le j \le m$, and which extends to points arbitrarily close to the boundary of D_0.
(b) Almost always $Y_j(t) = z_j(\frac{t}{n}) \cdot n + o(n)$ uniformly for $0 \le t \le \min\{\sigma n, t_0\}$ and for each i, where $z_j(s)$ is the solution in (a) with $\hat{z}^{(j)} = \frac{Y_j(0)}{n}$, and $\sigma = \sigma(n)$ is the supremum of those s to which the solution can be extended.

Using Lemma 1 and applying Wormald theorem, we may have

Lemma 2 *For any small $\epsilon > 0$, almost always we have uniformly for all $0 \le t \le (1-\epsilon)n$,*

$$\begin{aligned} \frac{|F_1(t)|}{n-t} &= c_1\left(1 - \frac{t}{n}\right) + o(1), \qquad \frac{|F_2(t)|}{n-t} = c_2\left(1 - \frac{t}{n}\right) + o(1), \\ \frac{|F_3(t)|}{n-t} &= c_3\left(1 - \frac{t}{n}\right)^p + o(1), \qquad \frac{|F_4(t)|}{n-t} = c_4\left(1 - \frac{t}{n}\right)^p + o(1). \end{aligned} \tag{2}$$

Proof: To apply Wormald theorem to our situation, let $m = 4$, $Y_i(t) = |F_i(t)|$ $(1 \le i \le 4)$, $C_0 = c_1 + c_2 + c_3 + c_4$, and $t_0 = (1-\epsilon)n$. Let

$$D_0 = \{(s, z_1, z_2, z_3, z_4) | -\epsilon < s < 1, -\epsilon < z_i < c_i + \epsilon\},$$

and

$$\begin{aligned} g_1(s, z_1, z_2, z_3, z_4) &= -\frac{2z_1}{1-s}, \qquad g_2(s, z_1, z_2, z_3, z_4) = -\frac{2z_2}{1-s}, \\ g_3(s, z_1, z_2, z_3, z_4) &= -(1+p)\frac{z_3}{1-s}, \qquad g_4(s, z_1, z_2, z_3, z_4) = -(1+p)\frac{z_4}{1-s}. \end{aligned}$$

By the expectations and distributions of $|F_{i+1}(t)| - |F_i(t)|$, the conditions in Wormald theorem are easily satisfied directly. Then we get $\varphi_i : [0, 1-\epsilon] \to \mathbb{R}$, the solution of the following system of differential equations,

$$\begin{cases} \frac{d\varphi_1}{dx} &= -\frac{2\varphi_1(x)}{1-x} & \varphi_1(0) = c_1 \\ \frac{d\varphi_2}{dx} &= -\frac{2\varphi_2(x)}{1-x} & \varphi_2(0) = c_2 \\ \frac{d\varphi_3}{dx} &= -(1+p)\frac{\varphi_3(x)}{1-x} & \varphi_3(0) = c_3 \\ \frac{d\varphi_4}{dx} &= -(1+p)\frac{\varphi_4(x)}{1-x} & \varphi_4(0) = c_4 \end{cases}$$

such that, almost always $|F_i(t)| = \varphi_i(\frac{t}{n}) \cdot n + o(n)$ holds uniformly for $0 \le t \le (1-\epsilon)n$ and $1 \le i \le 4$. Note that $\varphi_1(x) = c_1(1-x)^2$, $\varphi_2(x) = c_2(1-x)^2$, $\varphi_3(x) = c_3(1-x)^{1+p}$, and $\varphi_4(x) = c_4(1-x)^{1+p}$. □

As $\lim_{x\to 1} \sum_{i=1}^{4} \frac{\varphi_i(x)}{1-x} = 0$, we may choose $\epsilon > 0$ so that almost always the total number of 2-clauses and equalities remaining at time $t = (1-\epsilon)n$ is less than $0.01\epsilon n$. From Lemma 2, almost always the following holds:

$$|V_M(t)| = |F_3(t)| + |F_4(t)| \le (c_3 + c_4)(n-t) + o(n) < n - t = |V(t)|,$$

uniformly for all $0 \le t \le (1-\epsilon)n$. So, for $0 \le t \le (1-\epsilon)n$, almost always UCS does not encounter the case that $\chi(t) = 0$ but $V(t) - V_M(t)$ is empty.

For $1 \le i \le 4$, let $b_i(t)$ be the number of blue unit clauses coming from $F_i(t)$ at time t and let $r_i(t)$ be the number of red unit clauses coming from $F_i(t)$ at time t. We will obtain the expectations and distributions of $b_i(t)$'s and $r_i(t)$'s, conditioned on $|F_i(t)|$'s. Suppose that UCS sets a literal l to be true at time t. Then $b_1(t)$ is the number of 2-clauses in $F_1(t)$ that contain $\overline{l}$. And $r_1(t) = 0$ since there is no main literal in $F_1(t)$. Note that, for each 2-clause $(l_1 \vee l_2) \in F_1(t)$, $\Pr[\overline{l} = l_1 \text{ or } \overline{l} = l_2] = \frac{1}{n-t}$. And the 2-clauses in $F_1(t)$ are independent from one another. So $b_1(t)$ has a binomial distribution $\mathrm{Bin}[|F_1(t)|, \frac{1}{(n-t)}]$. The same argument can be applied to have that $r_2(t) = 0$ and $b_2(t)$ has a binomial distribution $\mathrm{Bin}[|F_2(t)|, \frac{2}{(n-t)}]$. For $b_3(t)$, observe that $b_3(t)$ is the number of 2-clauses in $F_3(t)$ whose main literals are $\overline{l}$. Here we consider two cases according to the value of $\chi(t)$. First, suppose that $\chi(t) = 1$. Then $b_3(t)$ has a Bernoulli distribution with density $\frac{|F_3(t)|}{2(n-t)}$ since $\overline{l}$ may be equal to at most one of the main literals in $F_3(t)$. When $\chi(t) = 0$, $b_3(t) = 0$ since $\overline{l}$ cannot be equal to any of the main literals in $F_3(t)$ and hence only unit clauses with main literals are produced. For $r_3(t)$, observe that $r_3(t)$ is the number of 2-clauses in $F_3(t)$ whose second literals are equal to $\overline{l}$. Suppose that l is not strictly distinct with one of main variables of $F_3(t)$. Then, since exactly one 2-clause in $F_3(t)$ has l or $\overline{l}$ as main literal, $r_3(t)$ has a binomial distribution $\mathrm{Bin}[|F_3(t)| - 1, \frac{1}{2(n-t-1)}]$. Otherwise, $r_3(t)$ has a binomial distribution $\mathrm{Bin}[|F_3(t)|, \frac{1}{2(n-t-1)}]$. Thus, the distribution of $r_3(t)$ is a linear combination of $\mathrm{Bin}[|F_3(t)| - 1, \frac{1}{2(n-t-1)}]$ and $\mathrm{Bin}[|F_3(t)|, \frac{1}{2(n-t-1)}]$. The same argument can be applied to obtain the distributions of $b_4(t)$ and $r_4(t)$. If $\chi(t) = 1$, $b_4(t)$ has a Bernoulli distribution with density $\frac{|F_4(t)|}{(n-t)}$ and if $\chi(t) = 0$, then $b_4(t) = 0$. And $r_4(t)$ has a binomial distribution $\mathrm{Bin}[|F_4(t)| - 1, \frac{1}{(n-t-1)}]$ if l is not strictly distinct with one of the main variables of $F_4(t)$. Otherwise, $r_4(t)$ has a binomial distribution $\mathrm{Bin}[|F_4(t)|, \frac{1}{(n-t-1)}]$. Thus, the distribution of $r_4(t)$ is a linear combination of $\mathrm{Bin}[|F_4(t)| - 1, \frac{1}{(n-t-1)}]$ and $\mathrm{Bin}[|F_4(t)|, \frac{1}{(n-t-1)}]$.

Observe that $b_i(t)$'s and $r_i(t)$'s (over t) are dependent random variables, which makes our analysis difficult. Fortunately, the dependency is weak and it is possible to bypass this obstacle using couplings. It is not hard to see that we may take mutually independent random variables $b_i^*(t)$'s (and $r_i^*(t)$'s) as follows: The random variable $b_1^*(t)$ has a binomial distribution $\mathrm{Bin}[(1+\delta)c_1 n(1-\frac{t}{n})^2, \frac{1}{(n-t)}]$ and $r_1^*(t) = 0$, where $\delta > 0$ is a small constant defined later. Though $b_i^*(t)$'s and $r_i^*(t)$'s, $i = 2, 3, 4$, may be similarly defined, we just list their distributions for completeness. The random variable $b_2^*(t)$ has a binomial distribution $\mathrm{Bin}[(1+\delta)c_1 n(1-\frac{t}{n})^2, \frac{2}{(n-t)}]$ and $r_2^*(t) = 0$. And, $b_3^*(t)$ is χ times a random variable that has a Bernoulli distribution with density $(1+\delta)\frac{c_3}{2}(1-\frac{t}{n})^p$, and $r_3^*(t)$ has a binomial distribution $\mathrm{Bin}[(1+\delta)c_3 n(1-\frac{t}{n})^{1+p}, \frac{1}{2(n-t-1)}]$. Finally, $b_4^*(t)$ is χ times a random variable that has a Bernoulli distribution with density $(1+\delta)\frac{c_3}{2}(1-\frac{t}{n})^p$, and $r_4^*(t)$ has a binomial distribution $\mathrm{Bin}[(1+\delta)c_4 n(1-\frac{t}{n})^{1+p}, \frac{1}{(n-t-1)}]$. We couple $b_i(t)$ and $b_i^*(t)$, (and $r_i(t)$ and $r_i^*(t)$) so that if $|F_i(t)|$'s satisfy Equation (2), then $b_i(t) \le b_i^*(t)$ (and $r_i(t) \le r_i^*(t)$). In particular, almost always

$$b_i(t) \le b_i^*(t), \quad r_i(t) \le r_i^*(t)$$

uniformly for all $0 \le t \le (1-\epsilon)n$. The expectations of $b_i^*(t)$ and $r_i^*(t)$ are as follows:

$$\begin{bmatrix} \mathrm{E}[b_i^*(t)] \\ \mathrm{E}[r_i^*(t)] \end{bmatrix} = T_i^*(t) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix} + o(1),$$

where

$$T_1^*(t) = c_1(1+\delta)(1-\frac{t}{n})\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \qquad T_2^*(t) = c_2(1+\delta)(1-\frac{t}{n})\begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix},$$
$$T_3^*(t) = c_3(1+\delta)(1-\frac{t}{n})^p\begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad T_4^*(t) = c_4(1+\delta)(1-\frac{t}{n})^p\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Then for $b(t) = \sum_{i=1}^4 b_i(t)$, $r(t) = \sum_{i=1}^4 r_i(t)$, $b^*(t) = \sum_{i=1}^4 b_i^*(t)$, and $r^*(t) = \sum_{i=1}^4 r_i^*(t)$, it is clear that, almost always, $b(t) \le b^*(t)$ and $r(t) \le r^*(t)$ for all $0 \le t \le (1-\epsilon)n$. By letting $T^*(t) = \sum_{i=1}^4 T_i^*(t)$, we have that

$$\begin{bmatrix} \mathrm{E}[b^*(t)] \\ \mathrm{E}[r^*(t)] \end{bmatrix} = T^*(t) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix} + o(1).$$

Lemma 3 *There exists a constant $\delta > 0$ such that, $\mathrm{E}[b^*(t)] < p - \delta$ and $\mathrm{E}[r^*(t)] < 1 - p - \delta$, almost always and uniformly for all $0 \le t \le (1-\epsilon)n$.*

Proof: Since $T^*(t) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix} \le T^*(0) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix}$, where the inequality for the vectors means that the inequality holds for each pair of the entries, it suffices to show that

$$T^*(0) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix} < \begin{bmatrix} p \\ 1-p \end{bmatrix}.$$

Clearly, $T^*(0) = (1+\delta)\begin{bmatrix} c_1 + 2c_2 + \frac{1}{2}c_3 + c_4 & c_1 + 2c_2 \\ \frac{1}{2}c_3 + c_4 & \frac{1}{2}c_3 + c_4 \end{bmatrix}$ has two nonnegative eigenvalues,

$$(1+\delta)\frac{c_1 + 2c_2 + c_3 + 2c_4 \pm \sqrt{(c_1+2c_2)^2 + 2(c_1+2c_2)(c_3+2c_4)}}{2}.$$

Let $\lambda(\delta)$ be the larger one. Since $D < 1$ implies $\lambda(0) < 1$, we may choose a small constant δ so that $\lambda(\delta) < 1$. Note that $\begin{bmatrix} p \\ 1-p \end{bmatrix}$ is an eigenvector of $T^*(0)$ corresponding to $\lambda(\delta)$. So,

$$T^*(0) \cdot \begin{bmatrix} p \\ 1-p \end{bmatrix} = \lambda(\delta) \begin{bmatrix} p \\ 1-p \end{bmatrix} < \begin{bmatrix} p \\ 1-p \end{bmatrix},$$

as desired. □

Then using these facts, we show that almost always the sizes of $\sum |B(t)|$ and $\sum |R(t)|$ are $O(n)$. In the course of that, we use a simplified version of *Lazy-server lemma*, which was introduced by Achlioptas [1]. Suppose that there is a server so that the probability that the server would work at time t is $w(t)$ and, if it works, it can handle one task per unit time. And the expected number of tasks that arrive to the server at time t is $z(t)$. Then Lazy-server lemma says that if $z(t)$ is bounded above by $w(t)$ uniformly for all t, then almost always the sum of sizes of the task queue over all t would be bounded linearly.

Lemma 4 (Lazy-server lemma) *Let $Z(0), Z(1), \ldots$ be a sequence of random variables and denote $z(t) = E[Z(t)]$. Let $W(0), W(1), \ldots$ be a sequence of independent Bernoulli random variables with density $w(t)$, i.e. $W(t) = 1$ with probability $w(t)$, and 0 otherwise. Let $Q(0), Q(1), \ldots$ be a sequence of random variables defined by $Q(0) = 0$ and $Q(t+1) = \max(Q(t) - W(t), 0) + Z(t)$. Assume that*
(i) there exist a constant $\rho > 0$ such that for all $t \geq 0$,

$$z(t) < w(t) - \rho.$$

(ii) there exist constants $a, b, c > 0$ such that for any fixed $0 \leq j_1 \leq j_2$ and $\beta > 0$,

$$\Pr[\sum_{t=j_1}^{j_2} Z(t) > (1+\beta) \sum_{t=j_1}^{j_2} z(t)] < \exp\left(-a\beta^b \left(\sum_{t=j_1}^{j_2} z(t)\right)^c\right).$$

Then there exists constant C and K depending on ρ, a, b, c such that for every $m \geq 1$,

$$\Pr[\sum_{t=0}^{m-1} Q(t) > Cm] = O(m^{-2}),$$

$$\Pr[\max_{0 \leq t < m} Q(t) > \log^K m] = O(m^{-2}).$$

From Lazy-server lemma, we have

Lemma 5 *We almost always have*

$$\sum_{t=0}^{(1-\epsilon)n} |B(t)| < Cn, \quad \max_{0 \leq t \leq (1-\epsilon)n} |B(t)| < \log^K n,$$

$$\sum_{t=0}^{(1-\epsilon)n} |R(t)| < Cn, \quad \max_{0 \leq t \leq (1-\epsilon)n} |R(t)| < \log^K n,$$

for some constants C, K.

Proof: We apply Lazy-server lemma with $Z(t) = b^*(t)$, $W(t) = \chi(t)$, $w(t) = p$ for $0 \leq t \leq (1-\epsilon)n$. Condition (i) follows by Lemma 3. And condition (ii) follows by Chernoff bound. So almost always for some constants $C_1 = C_1(\epsilon)$ and $K_1 = K_1(\epsilon)$,

$$\sum_{t=0}^{(1-\epsilon)n} Q(t) < C_1 n \quad \text{and} \quad \max_{0 \leq t \leq (1-\epsilon)n} Q(t) < \log^{K_1} n.$$

Note that $|B(t+1)| = \max(|B(t)| - \chi(t), 0) + b(t)$, and almost always $b(t) \leq b^*(t)$ for all $0 \leq t \leq (1-\epsilon)n$. So, by induction, almost always $|B(t)| \leq Q(t)$,

$$\sum_{t=0}^{(1-\epsilon)n} |B(t)| < C_1 n, \quad \text{and} \quad \max_{0 \leq t \leq (1-\epsilon)n} |Q(t)| < \log^{K_1} n.$$

Same argument can be applied to obtain that almost always there exist constants $C_2 = C_2(\epsilon)$ and $K_2 = K_2(\epsilon)$ such that

$$\sum_{t=0}^{(1-\epsilon)n} |R(t)| < C_2 n \quad \text{and} \quad \max_{0 \leq t \leq (1-\epsilon)n} |R(t)| < \log^{K_2} n.$$

Then for $C = C(\epsilon) = C_1 + C_2$ and $K = K(\epsilon) = \max\{K_1 + 1, K_2 + 1\}$, almost always

$$\sum_{t=0}^{(1-\epsilon)n} (|B(t)| + |R(t)|) < Cn \quad \text{and} \quad \max_{0 \leq t < (1-\epsilon)n} (|B(t)| + |R(t)|) < \log^{K} n.$$

□

Now we prove that with positive probability no 0-clause is produced until $t = (1-\epsilon)n$. Under the condition that no 0-clause is produced until time $t-1$, the probability that the same holds until time t is at least

$$\left(1 - \frac{1}{2(n-t-1)}\right)^{|B(t)|} \left(1 - \frac{|R(t)|}{2(n-t-1)}\right) \geq \left(1 - \frac{2}{\epsilon n}\right)^{|B(t)|+|R(t)|}$$

So the probability that no 0-clause is produced until $t = (1-\epsilon)n$ is at least

$$\left(1 - \frac{2}{\epsilon n}\right)^{\sum_{t=0}^{n-\epsilon n} (|B(t)|+|R(t)|)} \geq \left(1 - \frac{2}{\epsilon n}\right)^{Cn} = e^{-\frac{2C}{\epsilon}} + o(1).$$

We complete the proof by showing that when no 0-clause is produced until $t = (1-\epsilon)n$, the remaining formula at $t = (1-\epsilon)n$ is satisfiable with positive probability. Remind that the total number of 2-clauses and equalities remaining at the time is almost always less than $0.01\epsilon n$. Think of a multi graph G defined by the following rules. The vertices represent the Boolean variables in $V((1-\epsilon)n)$. For each 2-clause or equality of $F((1-\epsilon)n)$, we set an edge between the two vertices that appear in the 2-clause or equality. Notice that if G is acyclic and no pair of unit clauses at time $t = (1-\epsilon)n$ belongs to the same connected component, then there exists a truth assignment of $V((1-\epsilon)n)$ which satisfies $F((1-\epsilon)n)$.

Note that

$$\Pr[G \text{ has a cycle}] \leq \mathrm{E}[\text{number of cycles of } G] = \sum_{\text{possible cycles } \mathrm{C_G}} \Pr[C_G \text{ appears in } G].$$

The last summand is at most

$$\sum_{k=2}^{\epsilon n}\sum_{j=0}^{k}\binom{k}{j}|2V_M\left((1-\epsilon)n\right)|^j(\epsilon n)^{k-j}\left(\frac{1}{\epsilon n-1}\right)^j\left(\frac{|F_1(n-\epsilon n)|+|F_2(n-\epsilon n)|}{\binom{\epsilon n}{2}}\right)^{k-j},$$

where k is the length of cycle C_G, and j is the number of edges from $F_3((1-\epsilon)n)$ or $F_4((1-\epsilon)n)$. This summand is at most

$$\sum_{k=2}^{\infty}\sum_{j=0}^{k}\binom{k}{j}\left(0.02\frac{\epsilon n}{\epsilon n-1}\right)^j\left(0.02\frac{\epsilon n}{\epsilon n-1}\right)^{k-j}<0.1.$$

For each pair of the existing unit clauses (v_1) and (v_2) at time $t=(1-\epsilon)n$, it is not difficult to see that the probability that v_1 and v_2 are in the same connected component is $O(1/n)$. So the probability that there is a pair of unit clauses at time $t=(1-\epsilon)n$ that belongs to the same connected component is at most $\binom{\log^K n}{2}O(1/n)=o(1)$. Hence with probability higher than 0.9 the remaining formula at the time is satisfiable.

Now consider the case that $c_3+c_4=1$. Since $V(0)-V_M(0)$ is empty in this case as mentioned above, UCS may encounter the case that $\chi(t)=0$ but $V(t)-V_M(t)$ is empty unless $p=1$. However, when we set p as in (1), UCS may not encounter the case that $\chi(t)=0$ but $V(t)-V_M(t)$ is empty: Initially, $|V(0)-V_M(0)|=0$. At each step t of the first δn steps, if $\chi(t)=1$, then the expected change of $|V(t)-V_M(t)|$ is $1+O(\delta)$, as one uniform random literal eliminates $2+O(\delta)$ 2-clauses or equalities with main literals, in expectation. The other effects are small enough if δ is small enough. If $\chi(t)=0$, then the expected change is $O(\delta)$, as a non-main literal eliminates $1+O(\delta)$ 2-clause or equality with main literal, in average. Thus, at each step, $|V(t)-V_M(t)|$ increases by $p+O(\delta)$, in average, and hence $|V(t)-V_M(t)|>0$ for $t\geq 1$ with positive probability. Notice that UCS produces 0-clause in the first δn steps with probability $O(\delta)$ (cf. Lazy-server lemma). Therefore, with positive probability, UCS proceeds to the first δn steps without encountering $\chi(t)=0$ and $V(t)-V_M(t)=\emptyset$. After $t=\delta n$ steps, it is easy to see that $c_3(t)+c_4(t)\leq(1-\epsilon_0)(n-t)$ for some constant $\epsilon_0>0$, which is covered in the previous case.

3.2 Supercritical Region

For a 2-SAT formula F, setting a literal x to be true produces the unit clause (l) after $\overline{x}$ is removed from each clause $(\overline{x}\vee l)$ in F. Again, setting the literal l to be true yields other unit clauses and the process repeats for other unit clauses, if any. The process terminates if there is no more unit clause. This process is called an *implication process* starting from x, or simply an implication process if the identity of x is clear in the context. Strictly speaking, an implication process depends on the order of unit clauses chosen. We assume that there is a fixed order among all the unit clauses, as our argument below does not depend on a particular order. For the generalized random 2-SAT formula we are dealing with, there are two types of unit clauses colored blue and red as explained in the proof for the subcritical region. So we call the process in which blue and red unit clauses are distinguished an *implication process with two types*. If the implication process starting from x produces a 0-clause, there does not exist a satisfying assignment setting x to be true for the formula F. In addition, if the implication

process starting from $\overline{x}$ also produces a 0-clause, there does not exist a satisfying assignment setting $\overline{x}$ to be true (equivalently setting x to be false) for F and, consequently, the formula F is unsatisfiable. In the following, we will prove the unsatisfiability of $F(n, c_1, c_2, c_3, c_4)$ with $D > 1$ by showing that there almost always exists such a variable x that both of x and $\overline{x}$ produce 0-clauses in the random formula. To maintain the independence among the implications in the implication process, once a 2-clause or an equality is used in the process, we remove it from the formula and do not consider it in the subsequent process.

Before investigating the implication process with two types, we consider a branching process with two types, a simpler than but very close to the implication process. In the branching process, there are two types of organisms, say colored blue and red, and each type produces both types of organisms according to a certain probability distribution independently of the other. For an organism x, an organism is in the first generation if it is produced by x. In general, an organism is in the k^{th} generation if it is produced by an organism in the $(k-1)^{\text{th}}$ generation. We use the term "after k generations" when all organisms of k^{th} or less generation are exposed. Suppose that a blue organism produces a blue and c red organisms in expectation and a red organism produces b blue and d red organisms in expectation. Then, the branching ratios of the process are represented by a 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

We have the following result.

Lemma 6 *Suppose a two-type branching process with the branching ratio matrix*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

with $a \geq d$, and the larger eigenvalue of A is larger than 1. Then, there is a constant κ such that the expected number of the blue organisms that are produced after κ generations, starting from one blue organism, is larger than 1.

Proof: Let $B_b(k)$ ($R_b(k)$, resp.) be the expected numbers of blue (red, resp.) organisms produced after k generations, starting from one blue organism and $B_r(k)$ ($R_r(k)$, resp.) be the expected numbers of blue (red, resp.) organisms produced after k generations, starting from one red organism. Then, by induction on k,

$$\begin{bmatrix} B_b(k) \\ R_b(k) \end{bmatrix} = A^k \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} B_r(k) \\ R_r(k) \end{bmatrix} = A^k \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Let λ_A be the larger eigenvalue of A and $[u, v]$ be the corresponding eigenvector with $u + v = 1$, that is,

$$\lambda_A = \frac{a + d + \sqrt{(a-d)^2 + 4bc}}{2},$$

and

$$\begin{bmatrix} u \\ v \end{bmatrix} = \mu \begin{bmatrix} a - d + \sqrt{(a-d)^2 + 4bc} \\ 2c \end{bmatrix},$$

where μ is a positive constant.

If $b = 0$ or $c = 0$, $a \geq d$ and $\lambda_A > 1$ imply that $B_b(1) = a > 1$. Thus we may take $\kappa = 1$. Suppose that $b > 0$ and $c > 0$. We first choose the minimum integer K such that $(\lambda_A)^K \min\{u, v\} > 2$, i.e., set

$$K = \lceil \frac{\log \frac{2}{\min\{u,v\}}}{\log \lambda_A} \rceil + 1.$$

If $B_b(K) > 1$, we can set $\kappa = K$. If $R_r(K) > 1$, it is easy to see that there is a constant α so that $R_r(\alpha K) > 1.1/(bc)$. As, in expectation, one blue organism produces c red organisms, and one red organisms produces $R_r(\alpha K)$ red organisms after αK generations, and then each of those red organism produces b blue organisms, we have $bcR_r(\alpha K)$, which is at least 1.1, blue organisms in expectation after $\alpha K + 2$ generations. Suppose now $B_b(K) \leq 1$ and $R_r(K) \leq 1$. Note that

$$A^K \begin{bmatrix} u \\ v \end{bmatrix} = uA^K \begin{bmatrix} 1 \\ 0 \end{bmatrix} + vA^K \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} u \cdot B_b(K) + v \cdot R_r(K) \\ u \cdot R_b(K) + v \cdot R_r(K) \end{bmatrix}.$$

On the other hand, as $[u, v]$ is an eigenvector of A,

$$A^K \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} (\lambda_A)^K u \\ (\lambda_A)^K v \end{bmatrix}.$$

As $(\lambda_A)^K \min\{u, v\} > 2$ and all of u, v, $B_b(K)$, and $R_r(K)$ are less than or equal to 1, $B_r(K) > 1$ and $R_b(K) > 1$ and hence

$$\begin{bmatrix} B_b(2K) \\ R_b(2K) \end{bmatrix} = A^{2K} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = A^K A^K \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} B_b(K)^2 + B_r(K) \cdot R_b(K) \\ B_b(K) \cdot R_b(K) + R_b(K) \cdot R_r(K) \end{bmatrix}.$$

Therefore, $B_b(2K) \geq B_r(K)R_b(K) > 1$ and we can set $\kappa = 2K$. □

It is well known that a branching process with branching ratio larger than one continues forever with positive probability [3]. From Lemma 6, we have an analogy for branching processes with two types.

Corollary 1 *Suppose a two-type branching process with the branching ratio matrix satisfying the condition in Lemma 6. Then, the branching process, starting from one blue organism, continues forever with positive probability.*

Now we prove the second part of Theorem 2. Consider a random formula $F = F(n, c_1, c_2, c_3, c_4)$ with $D > 1$ and the implication process in F. Involved in the implication process, a *κ-generation implication process* (shortly, *κ-implication process*) is defined as follows. The κ-implication process consists of *ordinary rounds* followed by one *supplemental round*. It starts from a literal chosen uniformly at random (or equivalently, the literal in a blue unit clause). In the first ordinary round, the unit clauses in the κ^{th} or less generations from the literal are all exposed and the literals in them are assigned truth values. In each subsequent ordinary round, if there remain blue unit clauses, one of them is chosen, the literal in it is set, and the literals in the unit clauses in the κ^{th} or less generations are all assigned their truth values. Otherwise, the process is terminated. If the process proceeds by the first $n^{\frac{2}{3}}$ ordinary rounds, all the literals in the existing blue unit clauses are assinged their truth values in the supplemental round and the process is terminated.

Lemma 7 *Given $F = F(n, c_1, c_2, c_3, c_4)$ with $D > 1$, for some constant κ, the κ-implication process in F starting from a variable x chosen uniformly at random proceeds to the supplemental round with positive probability.*

Proof: Consider the implication process in F starting from a variable x chosen uniformly at random. A blue or red unit clause being assigned at time t in the implication process may be regarded as if UCS sets a literal with $\chi(t) = 1$ or 0. As in proving Lemma 2, we see that, almost always,

$$|F_i(t)| = |F_i(0)| + o(n) = c_i n + o(n) \tag{3}$$

$(1 \leq i \leq 4)$ uniformly for all $0 \leq t \leq n^{1-\epsilon}$ for arbitrarily small $\epsilon > 0$. For the implication ratios in the process, let $\hat{a}(t)$ and $\hat{c}(t)$ ($\hat{b}(t)$ and $\hat{d}(t)$, resp.) be the numbers of blue and red unit clauses produced by a blue (red, resp.) unit clause at time t. Recall that $b(t)$ and $r(t)$, the numbers of blue and red unit clauses coming from $F(t)$ at time t in the process of UCS, were investigated in the previous section. In fact, $\hat{a}(t)$ and $\hat{c}(t)$ have the same distributions as $b(t)$ and $r(t)$ with $\chi(t) = 1$, respectively. And, $\hat{b}(t)$ and $\hat{d}(t)$ have the same distributions as $b(t)$ and $r(t)$ with $\chi(t) = 0$, respectively. For example, $\hat{a}(t) = \sum_{i=1}^{4} \hat{a_i}(t)$, where $\hat{a_1}(t)$ and $\hat{a_2}(t)$ have binomial distributions $\mathrm{Bin}[|F_1(t)|, \frac{1}{n-t}]$ and $\mathrm{Bin}[|F_2(t)|, \frac{2}{n-t}]$, respectively, and $\hat{a_3}(t)$ and $\hat{a_4}(t)$ have Bernoulli distributions with densities $\frac{|F_3(t)|}{2(n-t)}$ and $\frac{|F_4(t)|}{n-t}$, respectively. Similarly, the distributions of $\hat{b}(t)$, $\hat{c}(t)$, and $\hat{d}(t)$ can be obtained from the results in the previous section. Combined with Equation (3), we see that the implication ratio matrix at time t is almost always

$$T(t) = \begin{bmatrix} c_1 + 2c_2 + \frac{c_3+2c_4}{2} & c_1 + 2c_2 \\ \frac{c_3+2c_4}{2} & \frac{c_3+2c_4}{2} \end{bmatrix} + o(1)$$

uniformly for all $1 \leq t \leq n^{1-\epsilon}$.

Consider the branching process with two types whose branching ratio matrix is

$$A = (1-\delta) \begin{bmatrix} c_1 + 2c_2 + \frac{c_3+2c_4}{2} & c_1 + 2c_2 \\ \frac{c_3+2c_4}{2} & \frac{c_3+2c_4}{2} \end{bmatrix}.$$

The eigenvalues of A are

$$(1-\delta)\frac{c_1 + 2c_2 + c_3 + 2c_4 \pm \sqrt{(c_1 + 2c_2)^2 + 2(c_1 + 2c_2)(c_3 + 2c_4)}}{2}.$$

Let $\hat{\lambda}(\delta)$ be the larger one. Since $D > 1$ implies $\hat{\lambda}(0) > 1$, we choose $\delta > 0$ so that $\hat{\lambda}(\delta) > 1$. Then, by Lemma 6, there is a constant κ such that the expected number of the blue organisms produced after κ generations, starting from one blue organism, is larger than one in the branching process. Note that, for sufficiently large n, the entries of $T(t)$ are larger than or equal to the corresponding entries of A for all $1 \leq t \leq n^{1-\epsilon}$. By coupling the implication process with the branching process and using Corollary 1, we see that the κ-implication process in F proceeds to the supplemental round with positive probability. □

Consider the κ-implication process in F starting from x chosen uniformly at random, where κ is specified as in Lemma 7. Now condition that the κ-implication process proceeds to the supplemental round. Let the strictly distinct literals in the blue unit clauses existing in the beginning of the supplemental round be $y_1, \ldots, y_L$, where L is the number of the literals. Since the process almost always

produces in expectation more than one blue unit clause in each of the $n^{\frac{2}{3}}$ ordinary rounds, by the large deviation result [3], L is almost always $\Theta(n^{\frac{2}{3}})$. And let the variables that have not occurred in the ordinary rounds be $z_1, \ldots, z_M$, where M is the number of such variables and almost always $n + o(n)$. Note that $D > 1$ implies $c_1 > 0$ or $c_2 > 0$. Suppose that $c_1 > 0$. Denote by $\hat{F}_1$ the formula consisting of the 2-clauses remaining in F_1 in the beginning of the supplemental round. Then, $|\hat{F}_1|$ is almost always $c_1 n + o(n)$. Define the random variable X_i for $1 \leq i \leq M$ such that $X_i = 1$ if the unit clauses (z_i) and $(\overline{z_i})$ are produced from the clauses in $\hat{F}_1$ by setting y_j's and $X_i = 0$ otherwise. And define the random variable $X_{i,1}$ ($X_{i,2}$, resp.) for $1 \leq i \leq M$ such that $X_{i,1} = 1$ ($X_{i,2} = 1$, resp.) if the unit clause (z_i) ($(\overline{z_i})$, resp.) is produced from the clauses in $\hat{F}_1$ by setting y_j's and $X_{i,1} = 0$ ($X_{i,2} = 0$, resp.) otherwise. Note that $\hat{F}_1$ consists of 2-clauses chosen uniformly at random among all the 2-clauses over the literals y_i's, $\overline{y_i}$'s, z_i's, and $\overline{z_i}$'s. Thus,

$$\begin{aligned}
\Pr[X_{i,1} = 0] &= \Pr[\text{neither } (\overline{y_j} \vee z_i) \text{ nor } (z_i \vee \overline{y_j}) \text{ are not in } \hat{F}_1 \text{ for all } 1 \leq j \leq L] \\
&= \left(1 - \frac{2L}{U}\right)^{|\hat{F}_1|},
\end{aligned}$$

where $U = (2M+2L)(2M+2L-2)$. Similarly, $\Pr[X_{i,2} = 0] = \left(1 - \frac{2L}{U}\right)^{|\hat{F}_1|}$ and $\Pr[X_{i,1} = 0 \text{ and } X_{i,2} = 0] = \left(1 - \frac{4L}{U}\right)^{|\hat{F}_1|}$. Hence,

$$\begin{aligned}
\Pr[X_i = 1] &= \Pr[X_{i,1} = 1 \text{ and } X_{i,2} = 1] \\
&= 1 - \Pr[X_{i,1} = 0 \text{ or } X_{i,2} = 0] \\
&= 1 - \left(2\left(1 - \frac{2L}{U}\right)^{|\hat{F}_1|} - \left(1 - \frac{4L}{U}\right)^{|\hat{F}_1|}\right). \qquad (4)
\end{aligned}$$

Again, condition that $L = \Theta(n^{\frac{2}{3}})$, $M = n + o(n)$, and $|\hat{F}_1| = c_1 n + o(n)$, which hold almost always. Let $X = \sum_{i=1}^{M} X_i$. Since $\mathrm{E}[X_i] = \Pr[X_i = 1] = 8\binom{|\hat{F}_1|}{2}\left(\frac{L}{U}\right)^2 - \Theta(n^{-1}) = \Theta(n^{-\frac{2}{3}})$ by the binomial expansion with Equation (4),

$$\mathrm{E}[X] = \Theta(M \cdot n^{-\frac{2}{3}}) = \Theta(n^{\frac{1}{3}}).$$

And, we see that $\mathrm{Var}[X_i] = \mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2 = \Theta(n^{-\frac{2}{3}}) - \Theta(n^{-\frac{4}{3}}) = \Theta(n^{-\frac{2}{3}})$. Since

$$\mathrm{E}[X_i X_j] = \Pr[X_i = 1 \text{ and } X_j = 1] = 1 - \Pr[X_{i,1} = 0 \text{ or } X_{i,2} = 0 \text{ or } X_{j,1} = 0 \text{ or } X_{j,2} = 0],$$

which is

$$1 - \left(\binom{4}{1}\left(1 - \frac{2L}{U}\right)^{|\hat{F}_1|} - \binom{4}{2}\left(1 - \frac{4L}{U}\right)^{|\hat{F}_1|} + \binom{4}{3}\left(1 - \frac{6L}{U}\right)^{|\hat{F}_1|} - \binom{4}{4}\left(1 - \frac{8L}{U}\right)^{|\hat{F}_1|}\right),$$

and

$$\mathrm{E}[X_i] = \Pr[X_i = 1] = 1 - \left(2\left(1 - \frac{2L}{U}\right)^{|\hat{F}_1|} - \left(1 - \frac{4L}{U}\right)^{|\hat{F}_1|}\right),$$

the binomial expansion says that

$$\mathrm{Cov}[X_i, X_j] = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\mathrm{E}[X_j] = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]^2 = -\Theta(n^{-\frac{7}{3}}).$$

Thus,

$$\mathrm{Var}[X] = \sum_{i=1}^{M} \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}[X_i, X_j] = \Theta(M \cdot n^{-\frac{2}{3}}) - \Theta(M^2 \cdot n^{-\frac{7}{3}}) = \Theta(n^{\frac{1}{3}}).$$

By Chebyshev inequality [3],

$$\Pr[X = 0] \leq \frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} = \Theta(n^{-\frac{1}{3}}).$$

Hence, a 0-clause is almost always produced in the supplemental round. For the case that $c_1 = 0$ and $c_2 > 0$, this fact can be obtained in a similar way by considering the equalities remaining in F_2.

Now we have that the κ-implication process in F starting from a variable x chosen uniformly at random produces a 0-clause with positive probability. Suppose that the process produces a 0-clause. In the case that F consists of equalities only, the fact that setting x to be true produces a 0-clause implies that setting $\overline{x}$ to be true also produces a 0-clause. In the other cases, we remove all the 2-clauses, which contain the underlying variables of the literals occurred in the κ-implication process, from F, except for the 2-clauses that contain the literal x. Since there remain $\Theta(n)$ uniform random 2-clauses or uniform random equalities almost always, setting $\overline{x}$ to be true produces a blue unit clause in at most two generations with positive probability and we have another κ-implication process starting from the blue unit clause. Note that, for the new process, the distributions mentioned in the above argument almost always do not change asymptotically. Hence, by the same argument, the new process produces a 0-clause with positive probability. In summary, with positive probability, 0-clauses are produced from both x and $\overline{x}$ for a variable x.

In the case that the κ-implication process starting from a variable x does not produce any 0-clause, we remove all the 2-clauses and equalities that contain the underlying variables of the literals occurred in the process. Then, we choose one of the remaining variables uniformly at random and consider the κ-implication processes starting from the variable and its negation. Since the expectation of the number of the exposed variables in a κ-implication process is $O(n^{\frac{2}{3}})$, the probability that the number of the exposed variables is greater than $n^{\frac{3}{4}}$ is $O(n^{-\frac{1}{12}})$ by Markov inequality [3]. This means that, when we consider the successive κ-implication processes starting from $\Theta(\log n)$ different literals, the distributions mentioned in the above argument almost always do not change asymptotically. So the probability, that there does not occur a variable such that 0-clauses are produced from the variable and its negation until $\log n$ variables are successively chosen, approaches to zero as n goes to infinity. This means that the random formula F is almost always unsatisfiable.

4 Solubility of $\mathbf{NK}(n, 2, z)$

In this section, we prove Theorem 1 for the model $\overline{\mathrm{NK}}(n, 2, z)$. This is enough as $\overline{\mathrm{NK}}(n, 2, z)$ is essentially the same as $\mathrm{NK}(n, 2, z)$. Recall $z_0 = \frac{27-7\sqrt{5}}{4} \approx 2.837$. In the first subsection, the result for the subcritical region $z < z_0$ is proven. The next subsection is for another proof for the supercritical region. By the monotonicity of the solubility of $\overline{\mathrm{NK}}(n, 2, z)$, it is enough to consider cases $2 < z < z_0$ and $z_0 < z < 3$.

4.1 Subcritical Region

As in NK$(n, 2, z)$, a 3-SAT formula F can be reduced from a random instance f of $\overline{\text{NK}}(n, 2, z)$. More precisely, a 3-SAT formula L_j is reduced from each local fitness function f_j of f and F is the conjunction of L_j's. We call L_j a *local formula*. Then each local formula consists of two 3-clauses with probability $1-h$ and three 3-clauses with probability h where $z = 2+h$, independently of all other local formulae. Main variables or its negations appeared in a local formula are called main (copies of) literals of the local formula. Note that any pair of main literals from different L_j's are strictly distinct. As in the generalized random 2-SAT problem, we apply UCS to F and show that UCS succeeds to find a satisfying assignment of F with positive probability. In the process of UCS, there appear four types of 2-clauses or equalities as presented in the generalized 2-SAT problem. Denoted by $F_i(t)$ $(1 \le i \le 4)$ is the 2-SAT formula consisting of the 2-clauses or equalities of type i at time t. Denoted by $F_5(t)$ is the 3-SAT formula consisting of the remaining local formulae at time t. Let $|F_i(t)|$ be the number of the 2-clauses or equalities in $F_i(t)$ and $|F_5(t)|$ be the number of the local formulae in $F_5(t)$ at time t. Then it is clear that $F_i(0)$ is empty for $1 \le i \le 4$ and $|F_5(0)| = n$. Let $V_M(t)$ be the set of the underlying variables of the main literals of $F_3(t)$, $F_4(t)$ and $F_5(t)$. We consider that in the process of UCS, unit clauses consisting of main literals and the copies of literals therein are colored red and other unit clauses and the copies of literals therein are colored blue. As in Section 3, we let $B(t)$ ($R(t)$, respectively) be the set of blue (red, respectively) unit clauses at time t.

For this problem, we run UCS with

$$p = p(t) = p_0 - \frac{t}{10n},$$

where $p_0 = \frac{\sqrt{5}-1}{2} \approx 0.618$. As one can see later, we defined $p(t)$ so that the expected number of blue (red, respectively) unit clauses produced at each time t is uniformly bounded above by $p(t)$ $(1-p(t)$, respectively) for $1 \le t \le (1-\epsilon)n$, where ϵ is a small constant. Then by a coupling argument and Lazy-server lemma, we obtain that the sizes of $\sum_{t=0}^{(1-\epsilon)n} |B(t)|$ and $\sum_{t=0}^{(1-\epsilon)n} |R(t)|$ are $O(n)$. Then we show that, with positive probability, no 0-clause is produced until $t = (1-\epsilon)n$. At $t = (1-\epsilon)n$, the remaining formula is sparse enough that it is satisfiable with positive probability.

Note that at each time t, $F_1(t)$ consists of uniform random 2-clauses over $V(t)$, and $F_2(t)$ consists of uniform random equalities over $V(t)$. The formula $F_3(t)$ consists of random 2-clauses with main literals over $V(t)$, and $F_4(t)$ consists of random equalities with main literals over $V(t)$, where the main literals in $F_3(t)$ and $F_4(t)$ and $F_5(t)$ are pairwise strictly distinct. Let $b_i(t)$ and $r_i(t)$ be the numbers of blue and red unit clauses coming from $F_i(t)$ at time t, respectively. As mentioned in Section 2, during the execution of UCS, there occurs some sublimations. So we also consider $b_5(t)$ ($r_5(t)$, respectively), the number of blue (red, respectively) unit clauses produced by sublimations at time t.

As in the generalized 2-SAT problem, we investigate $\mathrm{E}[|F_i(t+1)|-|F_i(t)|]$, $\mathrm{E}[b_i(t)]$ and $\mathrm{E}[r_i(t)]$ $(1 \le i \le 5)$ and use Wormald theorem to obtain approximations of $|F_i(t)|$, and then obtain approximations of $\mathrm{E}[b_i(t)]$, and $\mathrm{E}[r_i(t)]$. For $1 \le i \le 4$, let $u_i(t)$ be the number of 2-clauses or equalities that come from $F_5(t)$ to $F_i(t)$ at time t, and $d_i(t)$ be the number of 2-clauses or equalities that are removed from $F_i(t)$ at time t. Then for $1 \le i \le 4$, $\mathrm{E}[|F_i(t+1)|-|F_i(t)|] = \mathrm{E}[u_i(t)]-\mathrm{E}[d_i(t)]$. As we already obtained

$\mathrm{E}[d_i(t)]$, $\mathrm{E}[b_i(t)]$ and $\mathrm{E}[r_i(t)]$ $(1 \le i \le 4)$ in Section 3, we only need to obtain $\mathrm{E}[u_i(t)]$ $(1 \le i \le 4)$, $\mathrm{E}[b_5(t)]$, $\mathrm{E}[r_5(t)]$ and $\mathrm{E}[|F_5(t+1)| - |F_5(t)|]$.

Lemma 8 *For any small $\epsilon > 0$, we have for all $0 \le t \le (1-\epsilon)n$,*

$$\begin{aligned}
\mathrm{E}[|F_1(t+1)| - |F_1(t)|] &= -\frac{2|F_1(t)|}{n-t} + p(t)(\frac{4}{7} - \frac{h}{7})\frac{|F_5(t)|}{n-t},\\
\mathrm{E}[|F_2(t+1)| - |F_2(t)|] &= -\frac{2|F_2(t)|}{n-t} + p(t)(\frac{1}{14} + \frac{h}{14})\frac{|F_5(t)|}{n-t},\\
\mathrm{E}[|F_3(t+1)| - |F_3(t)|] &= -(1+p(t))\frac{|F_3(t)|}{n-t} + (\frac{8}{7} - \frac{2h}{7})\frac{|F_5(t)|}{n-t} + o(1),\\
\mathrm{E}[|F_4(t+1)| - |F_4(t)|] &= -(1+p(t))\frac{|F_4(t)|}{n-t} + (\frac{1}{7} + \frac{h}{7})\frac{|F_5(t)|}{n-t} + o(1),\\
\mathrm{E}[|F_5(t+1)| - |F_5(t)|] &= -(2+p(t))\frac{|F_5(t)|}{n-t} + o(1).
\end{aligned}$$

Proof: Let L be a local formula of $F_5(t)$. Recall that the probability that L is a conjunction of two (or three) 3-clauses is $1-h$ (or h) independently of all other local formulae. Denoted by $|L|$ is the number of 3-clauses of L. We write $m(L)$ for the main variable of L. Suppose a literal l is set to be true at time t. For convenience, we define an index $I(L,l)$ to be an ordered pair (a,b), where $a = 1(2,$ respectively) means that l and the main (a neighborhood) variable of L are not strictly distinct, and b indicates how many copies of l appears in the 3-clauses of L. For example, $I(L,l) = (1,2)$ means that $m(L)$ and l are not strictly distinct and l appears in two 3-clauses of L.

Now we obtain $\mathrm{E}[u_1(t)]$. Note that one uniform random 2-clause is produced from L only when $|L| = 2, \chi(t) = 1, I(L,l) = (1,1)$ or $|L| = 3, \chi(t) = 1, I(L,l) = (1,2)$. So the corresponding expected numbers of new uniform random 2-clauses that come from $F_5(t)$ are

$$(1-h) \times p(t) \times \frac{|F_5(t)|}{n-t} \times \frac{4 \cdot 4}{28} \quad \text{and} \quad h \times p(t) \times \frac{|F_5(t)|}{n-t} \times \frac{4\binom{4}{2}}{56},$$

and hence

$$\mathrm{E}[u_1(t)] = p(t)(\frac{4}{7} - \frac{h}{7})\frac{|F_5(t)|}{n-t}.$$

Similarly, contributions for $u_2(t)$ from L are made only when $|L| = 2, \chi = 1, I(L,l) = (1,0)$ or $|L| = 3, \chi = 1, I(L,l) = (1,1)$. So the corresponding expected numbers of new uniform random equalities that come from $F_5(t)$ are

$$(1-h) \times p(t) \times \frac{|F_5(t)|}{n-t} \times \frac{2}{28} \quad \text{and} \quad h \times p(t) \times \frac{|F_5(t)|}{n-t} \times \frac{2 \cdot 4}{56},$$

and hence

$$\mathrm{E}[u_2(t)] = p(t)(\frac{1}{14} + \frac{h}{14})\frac{|F_5(t)|}{n-t}.$$

For $u_3(t)$, the cases are $|L| = 2, \chi = 1, I(L,l) = (2,1)$ and $|L| = 2, \chi = 0, I(L,l) = (2,1)$ and $|L| = 3, \chi = 1, I(L,l) = (2,2)$ and $|L| = 3, \chi = 0, I(L,l) = (2,2)$. The corresponding expected numbers of new random 2-clauses with main literals that come from $F_5(t)$ are

$$(1-h) \times p(t) \times \frac{2|F_5(t)|}{n-t} \times \frac{4 \cdot 4}{28}, \quad (1-h) \times (1-p(t)) \times \frac{2|F_5(t)|}{n-t-1} \times \frac{4 \cdot 4}{28}$$

and

$$h \times p(t) \times \frac{2|F_5(t)|}{n-t} \times \frac{4 \cdot \binom{4}{2}}{56}, \quad h \times (1-p(t)) \times \frac{2|F_5(t)|}{n-t-1} \times \frac{4 \cdot \binom{4}{2}}{56},$$

and hence

$$\mathrm{E}[u_3(t)] = (\frac{8}{7} - \frac{2h}{7})\frac{|F_5(t)|}{n-t} + o(1).$$

For $u_4(t)$, the cases are $|L| = 2, \chi = 1, I(L,l) = (2,0))$ and $|L| = 2, \chi = 0, I(L,l) = (2,0)$ and $|L| = 3, \chi = 1, I(L,l) = (2,1)$ and $|L| = 3, \chi = 0, I(L,l) = (2,1)$. The corresponding expected numbers are

$$(1-h) \times p(t) \times \frac{2|F_5(t)|}{n-t} \times \frac{2}{28}, \quad (1-h) \times (1-p(t)) \times \frac{2|F_5(t)|}{n-t-1} \times \frac{2}{28}$$

and

$$h \times p(t) \times \frac{2|F_5(t)|}{n-t} \times \frac{2 \cdot 4}{56}, \quad h \times (1-p(t)) \times \frac{2|F_5(t)|}{n-t-1} \times \frac{2 \cdot 4}{56},$$

and hence

$$\mathrm{E}[u_4(t)] = (\frac{1}{7} - \frac{h}{7})\frac{|F_5(t)|}{n-t} + o(1).$$

Similarly,

$$\begin{aligned} \mathrm{E}[b_5(t)] &= (1+p(t))(\frac{1}{7} + \frac{2h}{7})\frac{|F_5(t)|}{n-t} + o(1), \\ \mathrm{E}[r_5(t)] &= (\frac{1}{7} + \frac{2h}{7})\frac{|F_5(t)|}{n-t} + o(1), \end{aligned}$$

and

$$\mathrm{E}[|F_5(t+1)| - |F_5(t)|] = -(2+p(t))\frac{|F_5(t)|}{n-t} + o(1).$$

□

For any small constant $\epsilon > 0$ we can apply Wormald theorem to approximate $|F_i(t)|$ uniformly for $1 \le t \le (1-\epsilon)n$. For the solution $\varphi_i(x) : [0, 1-\epsilon] \to \mathbb{R}$ of the following system of differential equations,

$$\begin{cases} \frac{d\varphi_1}{dx} &= -\frac{2\varphi_1(x)}{1-x} + (p_0 - 0.1x)(\frac{4}{7} - \frac{h}{7})\frac{\varphi_5(x)}{1-x} & \varphi_1(0) = 0 \\ \frac{d\varphi_2}{dx} &= -\frac{2\varphi_2(x)}{1-x} + (p_0 - 0.1x)(\frac{1}{14} + \frac{h}{14})\frac{\varphi_5(x)}{1-x} & \varphi_2(0) = 0 \\ \frac{d\varphi_3}{dx} &= -(1+p_0-0.1x)\frac{\varphi_3(x)}{1-x} + (\frac{8}{7} - \frac{2h}{7})\frac{\varphi_5(x)}{1-x} & \varphi_3(0) = 0 \\ \frac{d\varphi_4}{dx} &= -(1+p_0-0.1x)\frac{\varphi_4(x)}{1-x} + (\frac{1}{7} + \frac{h}{7})\frac{\varphi_5(x)}{1-x} & \varphi_4(0) = 0 \\ \frac{d\varphi_5}{dx} &= -(2+p_0-0.1x)\frac{\varphi_5(x)}{1-x} & \varphi_5(0) = 1, \end{cases}$$

almost always

$$|F_i(t)| = \varphi_i(\frac{t}{n}) \cdot n + o(n)$$

uniformly for all $1 \le t \le (1-\epsilon)n$ and $1 \le i \le 5$.

Now we choose $\epsilon > 0$ so that $\sum_{i=1}^{5} \frac{\varphi_i(1-\epsilon)}{\epsilon} < 0.01$, i.e., the total number of 2-clauses, equalities, and local formulae remaining at time $t = (1-\epsilon)n$ is almost always less than $0.01\epsilon n$.

Lemma 9 *There exists a constant $\delta > 0$ such that, $\mathrm{E}[b(t)] < p(t) - \delta$ and $\mathrm{E}[r(t)] < 1 - p(t) - \delta$, almost always and uniformly for all $0 \le t \le (1-\epsilon)n$.*

Proof: The expectations of $b_i(t)$ and $r_i(t)$ are as follows.

$$\begin{bmatrix} \mathrm{E}[b_i(t)] \\ \mathrm{E}[r_i(t)] \end{bmatrix} = T_i(t) \cdot \begin{bmatrix} p(t) \\ 1-p(t) \end{bmatrix} + o(1),$$

where

$$T_1(t) = \frac{|F_1(t)|}{(n-t)} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad T_2(t) = \frac{|F_2(t)|}{(n-t)} \begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix},$$
$$T_3(t) = \frac{|F_3(t)|}{(n-t)} \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad T_4(t) = \frac{|F_4(t)|}{(n-t)} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$
$$T_5(t) = \frac{(\frac{1}{7} + \frac{2h}{7})|F_5(t)|}{(n-t)} \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

Let $b(t) = \sum_{i=1}^{5} b_i(t)$ and $r(t) = \sum_{i=1}^{5} r_i(t)$. Then for $T(t) = \sum_{i=1}^{5} T_i(t)$,

$$\begin{bmatrix} \mathrm{E}[b(t)] \\ \mathrm{E}[r(t)] \end{bmatrix} = T(t) \cdot \begin{bmatrix} p(t) \\ 1-p(t) \end{bmatrix} + o(1).$$

So

$$\begin{aligned} \mathrm{E}[b(t)] &= \frac{1}{n-t}[|F_1(t)| + 2|F_2(t)| + (\frac{1}{7} + \frac{2h}{7})|F_5(t)| \\ &\quad + p(t)(\frac{|F_3(t)|}{2} + |F_4(t)| + (\frac{1}{7} + \frac{2h}{7})|F_5(t)|)] + o(1), \\ \mathrm{E}[r(t)] &= \frac{1}{n-t}(|F_3(t)| + 2|F_4(t)| + (\frac{1}{7} + \frac{2h}{7})|F_5(t)|) + o(1). \end{aligned} \tag{5}$$

Note that $\begin{bmatrix} p_0 \\ 1-p_0 \end{bmatrix} = \begin{bmatrix} \frac{(\sqrt{5}-1)}{2} \\ 1 - \frac{(\sqrt{5}-1)}{2} \end{bmatrix}$ is an eigenvector of $T(0) = (\frac{1}{7} + \frac{2h}{7}) \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and the corresponding eigenvalue is less than one if and only if $z < z_0$.

From (5), by letting

$$\begin{aligned} \varphi_b(x) &= \frac{1}{1-x}(\varphi_1(x) + 2\varphi_2(x) + (\frac{1}{7} + \frac{2h}{7})\varphi_5(x) \\ &\quad + \frac{p_0 - 0.1x}{2}(\varphi_3(x) + 2\varphi_4(x) + (\frac{1}{7} + \frac{2h}{7})\varphi_5(x))) \end{aligned}$$

and

$$\varphi_r(x) = \frac{1}{1-x}(\varphi_3(x) + 2\varphi_4(x) + (\frac{1}{7} + \frac{2h}{7})\varphi_5(x)),$$

we obtain that $\mathrm{E}[b(t)] = \varphi_b(t/n) + o(1)$ and $\mathrm{E}[r(t)] = \varphi_r(t/n) + o(1)$. So, if we show that $\varphi_b(x) < p_0 - 0.1x$ and $\varphi_r(x) < 1 - (p_0 - 0.1x)$ for $0 \le x < 1$, we obtain the required result. By inserting the formulae for $\frac{d\varphi_i}{dx}$'s and by the fact that $\varphi_i(x)$'s are nonnegative, we obtain that $\varphi_b(0) < p_0$, $\varphi_b'(0) < -0.1$ and $\varphi_b''(x) < 0$ for $0 \le x < 1$. So $\varphi_b(x) < p_0 - 0.1x$. Similarly, $\varphi_r(0) < 1 - p_0$, $\varphi_r'(0) < 0.1$ and $\varphi_r''(x) < 0$. So $\varphi_r(x) < 1 - (p_0 - 0.1x)$. It is worth to note that intuitively we have defined $p(t)$ so that $p(nx)$ is a line between the curves $\varphi_b(x)$ and $1 - \varphi_r(x)$, using the facts that $\varphi_b(x)$ and $\varphi_r(x)$ are convex. □

Now as in the generalized random 2-SAT problem with $D < 1$ and $c_3 + c_4 = 1$, for some small constant $\delta > 0$, at each step t of the first δn steps $|V(t) - V_M(t)|$ increases in average. Hence by

the same argument, with positive probability UCS proceeds to the first δn steps without encountering $\chi(t) = 0$ and $V(t) - V_M(t) = \emptyset$ and without producing any 0-clauses.

Note that $|V(t)| = n - t$ and almost always, $|V_M(t)| = |F_3(t)| + |F_4(t)| + |F_5(t)| = n(\varphi_3(t/n) + \varphi_4(t/n) + \varphi_5(t/n)) + o(n)$. And we can obtain that for all $0 < x < 1 - \epsilon$, $\varphi_3(x) + \varphi_4(x) + \varphi_5(x) < 1 - x$, by using the formulae for $\frac{d\varphi_i}{dx}$'s $(i = 3, 4, 5)$ and the fact that $\varphi_i(x)$'s are nonnegative. So under the condition that UCS proceeds to the first δn steps, almost always UCS does not encounter the case that $\chi(t) = 0$ and $V(t) - V_M(t) = \emptyset$ for $\delta n \leq t \leq (1 - \epsilon)n$.

Then as in the generalized random 2-SAT problem with $D < 1$, using coupling argument and Lazy-server lemma, we obtain that almost always $\sum_{t=1}^{(1-\epsilon)n} |B(t)|$ and $\sum_{t=1}^{(1-\epsilon)n} |R(t)|$ are bounded by $O(n)$. Then with positive probability, no 0-clause is produced until $t = (1 - \epsilon)n$. And under the condition that no 0-clause is produced until $t = (1 - ep)n$, the remaining formula at $t = (1 - \epsilon)n$ is satisfiable with positive probability.

4.2 Supercritical Region

As in the previous section, let F be the 3-SAT formula reduced from a random instance f of $\overline{\mathrm{NK}}(n, 2, z)$. For $z_0 < z < 3$, local formulae in F consist of two 3-clauses or three 3-clauses. A local formula with two 3-clauses may have the form of $(l_1 \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee l_2 \vee l_3)$, which is equivalent to the 2-clause $(l_2 \vee l_3)$. According to the position of the negation, the 2-clause is a uniform random 2-clause or a random 2-clause with main literal. A local formula with three 3-clauses may have the form of $(l_1 \vee l_2 \vee l_3) \wedge (l_1 \vee l_2 \vee \overline{l_3}) \wedge (l_1 \vee \overline{l_2} \vee l_3)$, which is equivalent to $(l_1 \vee l_2) \wedge (l_1 \vee l_3)$. We call such a conjunction of two 2-clauses a *cherry* as the shape of the relations among literals looks like it. According to the positions of the negations, the literal appearing twice in a cherry may be a main literal or not. If it is a main literal, the cherry is called a *symmetric cherry*. Otherwise, it is called an *asymmetric cherry*. A local formula with three 3-clauses may have the form of $(l_1 \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee \overline{l_2} \vee \overline{l_3})$, which implies the 2-clause $(l_2 \vee l_3)$. (The converse does not hold.) The 2-clause is a uniform random 2-clause or a random 2-clause with main literal according to the positions of the negations.

This way, we resolve the local formulae in F into four types of 2-clauses and cherries: uniform random 2-clauses, asymmetric cherries, random 2-clauses with main literals, and symmetric cherries. Note that the unsatisfiability of the 2-SAT formula consisting of the 2-clauses and cherries obtained from the resolution implies the unsatisfiability of F. In fact, we prove the unsatisfiability of F by showing the unsatisfiability of the 2-SAT formula. To do this, we define a generalized random 2-SAT formula and examine its satisfiability. The second part of Theorem 1 is obtained as a corollary.

The generalized random 2-SAT formula has four types of random 2-clauses and cherries. The first type consists of uniform random 2-clauses. The second type consists of random asymmetric cherries of the form $(r_i \vee u_{i1}) \wedge (u_{i1} \vee u_{i2})$, where u_{i1} and u_{i2} being strictly distinct with r_i are chosen uniformly at random. The third type consists of random 2-clauses with main literals. The fourth type consists of random symmetric cherries of the form $(r_i \vee u_{i1}) \wedge (r_i \vee u_{i2})$, where u_{i1} and u_{i2} being strictly distinct with r_i are chosen uniformly at random. The copies of the literals r_i's in cherries and the literals in the first places in random 2-clauses with main literals are pairwise strictly distinct and they are called

main literals in the random 2-SAT formula. Let c_1, c_2, c_3 and c_4 be non-negative real numbers with $c_2+c_3+c_4 \leq 1$. Denote $\overline{F_i} = \overline{F_i}(n, c_i)$ the conjunction of $c_i n$ 2-clauses or cherries of type i $(1 \leq i \leq 4)$. Denoted by $\overline{F}(n, c_1, c_2, c_3, c_4)$ is the conjunction of the four random formulae with pairwise strictly distinct main literals. The parameter

$$\overline{D} = (c_1 + c_2) + (c_2 + c_3 + 2c_4) - \frac{(c_2 + c_3 + 2c_4)^2}{4}$$

essentially determines the branching ratio in $\overline{F}(n, c_1, c_2, c_3, c_4)$. The satisfiability of $\overline{F}(n, c_1, c_2, c_3, c_4)$ can be described in terms of $\overline{D}$ as follows.

Theorem 4 *If $\overline{D} < 1$, then there exists $\alpha > 0$ depending on c_i's so that the probability of $\overline{F}(n, c_1, c_2, c_3, c_4)$ being satisfiable is at least α as n goes to infinity. If $\overline{D} > 1$, then the random formula is almost always unsatisfiable.*

Theorem 4 can be proved in a similar way as the proof of Theorem 2. We present the proof for the second part of Theorem 4 in Appendix for the completeness of the proof in this section.

Now we prove the second part of Theorem 1. As mentioned above, a local formula in F is resolved into a uniform random 2-clause $(l_2 \vee l_3)$ when it has the form of $(l_1 \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee l_2 \vee l_3)$ or $(l_1 \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee l_2 \vee l_3) \wedge (\overline{l_1} \vee \overline{l_2} \vee \overline{l_3})$. If we let $z = 2 + h$ $(z_0 - 2 < h < 1)$, this takes place with probability $\frac{1}{7}h + \frac{1}{7}(1-h) = \frac{1}{7}$. So, the probability that a local formula is resolved into a uniform random 2-clause is $\frac{1}{7}$. In a similar way, we see the following: The probability of a local formula being resolved into an asymmetric cherry is $\frac{2}{7}h$, the probability for a 2-clause with main literal is $\frac{2}{7}$, and the probability for a symmetric cherry is $\frac{1}{7}h$. Then, the expected numbers of uniform random 2-claueses, asymmetric cherries, 2-clauses with main literals, and symmetric cherries obtained from F are $\frac{1}{7}n$, $\frac{2}{7}hn$, $\frac{2}{7}n$, and $\frac{1}{7}hn$, respectively.

Since each local formula is created independently of other local formulae, the numbers of 2-clauses and cherries obtained from F are highly concentrated around their expectations. Let $c_1 = \frac{1}{7}$, $c_2 = \frac{2}{7}h$, $c_3 = \frac{2}{7}$, $c_4 = \frac{1}{7}h$, and $c_i(\delta) = c_i - \delta$ for $1 \leq i \leq 4$. Then, by the large deviation result [3], the numbers of 2-clauses and cherries of four types are almost always larger than $c_1(\delta)n$, $c_2(\delta)n$, $c_3(\delta)n$, and $c_4(\delta)n$, respectively, for arbitrarily small $\delta > 0$. Since $z_0 - 2 < h < 1$ implies that $c_2 + c_3 + c_4 \leq 1$ and $\overline{D} > 1$, we may choose $\delta > 0$ so that $c_2(\delta) + c_3(\delta) + c_4(\delta) \leq 1$ and

$$(c_1(\delta) + c_2(\delta)) + (c_2(\delta) + c_3(\delta) + 2c_4(\delta)) - \frac{(c_2(\delta) + c_3(\delta) + 2c_4(\delta))^2}{4} > 1.$$

Then, the second part of Theorem 4 implies that the random 2-SAT formula resolved from F is almost always unsatisfiable and the proof completes.

In that the 2-SAT formula resolved from F is exploited in the proof, our approach is similar to that of Gao and Culberson [20]. However, they counted the number of unsatisfiable subformulae in the resolved 2-SAT formula, which leads to complex computation. We used branching process arguments to derive the proof in more intuitive and natural way.

5 Conclusion

In this paper, we analyzed the phase transition in NK landscape on the fixed ratio model, NK$(n, 2, z)$. We also proposed a generalized random 2-SAT model and introduced a corresponding parameter D. Then a phase transition result for the model is obtained, that is, if $D < 1$, the formula is satisfiable with positive probability, and if $D > 1$, the formula is almost always unsatisfiable. For the proof of the subcritical region, we provided a variant of the unit clause algorithm, the unit clause algorithm with switching server policy, and analyzed it. For the supercritical region, a branching process argument was used.

Using a similar argument as in the generalized random 2-SAT model, it was proved that a random instance generated by NK$(n, 2, z)$ with $z < z_0 = \frac{27-7\sqrt{5}}{4}$ is soluble with positive probability. To the best of our knowledge, this is the first mathematical result that describes the behavior of NK$(n, 2, z)$ with $z < z_0$. We also reproved that a random instance generated by NK$(n, 2, z)$ with $z > z_0$ is almost always insoluble using a branching process argument. This approach is a novel one and is simpler than that of Gao and Culberson. From these results, we established the threshold value, z_0, of the phase transition in NK$(n, 2, z)$.

We believe that our approach used for NK(n, k, z) with $k = 2$ works for general $k \geq 3$ to obtain at least partial results for the phase transition phenomenon.

References

[1] D. Achlioptas. Setting two variables at a time yields a new lower bound for random 3-SAT. In *32nd Annual ACM Symposium on Theory of Computing*, pages 28–37, 2000.

[2] D. Achlioptas. Lower bounds for random 3-SAT via differential equations. *Theoretical Computer Science*, 265(1–2):159–185, 2001.

[3] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley, New York, 1992.

[4] L. Altenberg. NK fitness landscapes. In T. Bäck, D. Fogel, and Z. Michalewicz, editors, *Handbook of Evolutionary Computation*. Oxford University Press, 1997.

[5] C. Amitrano, L. Peliti, and M. Saber. Population dynamics in a spin-glass model of chemical evolution. *Journal of Molecular Evolution*, 29:513–525, 1989.

[6] B. Bollobás, Christian Borgs, Jennifer Chayes, J.H. Kim, and D.B. Wilson. The scaling window of the 2-SAT transition. *Random Structures and Algorithms*, 18:201–256, 2001.

[7] M. T. Chao and J. Franco. Probabilistic analysis of two heuristics for the 3-satisfiability problem. *SIAM Journal on Computing*, 15(4):1106–1118, 1986.

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

[9] V. Chvátal and B. Reed. Mick gets some (the odds are on his side). In *33th Annual Symposium on Foundations of Computer Science*, pages 620–627, 1992.

[10] W. Fernandez de la Vega. On random 2-SAT. Unpublished manuscript, 1992.

[11] O. Dubois, Y. Boufkhad, and J. Mandler. Typical random 3-SAT formulae and the satisfiability threshold. In *11th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 126–127, 2000.

[12] R. Durrett and V. Limic. Rigorous results for the NK model. *Annals of Probability*, 31(4):1713–1753, 2003.

[13] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Advanced Chemical Physics*, 75:149–263, 1989.

[14] S. N. Evans and D. Steinsaltz. Estimating some features of NK fitness landscapes. *Annals of Applied Probability*, 12:1299–1321, 2002.

[15] W. Ewens. *Mathematical Population Genetics*. Springer Verlag, 1979.

[16] H. Flyvbjerg and B. Lautrup. Evolution in a rugged fitness landscape. *Physical Review A*, 46:6714–6723, 1992.

[17] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Physical Review E*, 47:2083–2099, 1993.

[18] I. Franklin and R. Lewontin. Is the gene the unit of selection ? *Genetics*, 65:707–734, 1970.

[19] A. M. Frieze and S. Suen. Analysis of two simple heuristics on a random instance of k-SAT. *Journal of Algorithms*, 20(2):312–355, 1996.

[20] Y. Gao and J. Culberson. An analysis of phase transition in NK lanscapes. *Journal of Artificial Intelligence Research*, 17:309–332, 2002.

[21] Y. Gao and J. Culberson. On the treewidth of NK landscapes. In *Genetic and Evolutionary Computation Conference*, pages 948–954, 2003.

[22] A. Goerdt. A threshold for unsatisfiability. *Journal of Computer and System Sciences*, 53(3):469–486, 1996.

[23] W. Hordijk. A measure of landscapes. *Evolutionary Computation*, 4(4):335–360, 1997.

[24] T. Jones and S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *Sixth International Conference on Genetic Algorithms*, pages 184–192, 1995.

[25] S. A. Kauffman. Adaptation on rugged fitness landscapes. In D. Stein, editor, *Lectures in the Sciences of Complexity*, pages 527–618. Addison Wesley, 1989. Santa Fe Institute Studies in the Sciences of Complexity.

[26] S. A. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128:11–45, 1987.

[27] S. A. Kauffman, E. D. Weinberger, and A. S. Perelson. Maturation of the immune response via adaptive walks on affinity landscapes. In A. S. Perelson, editor, *Theoretical Immunology I*. Addison Wesley, 1988. Santa Fe Institute Studies in the Sciences of Complexity.

[28] H. Kaul and S. H. Jacobson. Global optima results for the Kauffman NK model. *Mathematical Programming*, 106(2):319–338, 2006.

[29] H. Kaul and S. H. Jacobson. New global optima results for the Kauffman NK model: Handling dependency. *Mathematical Programming*, 2006. Accepted.

[30] D. A. Levinthal. Adaptation on rugged landscapes. *Management Science*, 43:934–950, 1997.

[31] R. Lewontin. *The Genetic Basis of Evolutionary Change*. Columbia University Press, 1974.

[32] C. A. Macken and A. S. Perelson. Protein evolution on rugged landscapes. In *National Academic Science USA*, volume 86, pages 6191–6195, 1989.

[33] P. Merz and B. Freisleben. On the effectiveness of evolutionary search in high-dimensional NK-landscapes. In *IEEE International Conference on Evolutionary Computation*, pages 741–745, 1998.

[34] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computational Chemistry*, 18:295–324, 1994.

[35] D. I. Seo, Y. H. Kim, and B. R. Moon. New entropy-based measures of gene significance and epistasis. In *Genetic and Evolutionary Computation Conference*, pages 1345–1356, 2003.

[36] B. Skellet, B. Cairns, N. Geard, B. Tonkes, and J. Wiles. Maximally rugged NK landscapes contain the highest peaks. In *Genetic and Evolutionary Computation Conference*, pages 579–584, 2005.

[37] T. Smith, P. Husbands, P. Layzell, and M. O'Shea. Fitness landscapes and evolvability. *Evolutionary Computation*, 10(1):1–34, 2002.

[38] E. D. Weinberger. A more rigorous derivation of some properties of uncorrelated fitness landscapes. *Journal of Theoretical Biology*, 134:125–129, 1988.

[39] E. D. Weinberger. Local properties of Kauffman's NK model, a tuneably rugged energy landscape. *Physical Review A*, 44(10):6399–6413, 1991.

[40] E. D. Weinberger. NP completeness of Kauffman's NK model, a tuneably rugged fitness landscape. Technical Report 96-02-003, Santa Fe Institute, Santa Fe, 1996.

[41] N. C. Wormald. Differential equations for random processes and random graphs. *Annals of Applied Probability*, 5(4):1217–1235, 1995.

[42] A. H. Wright, R. K. Thompson, and J. Zhang. The computational complexity of NK fitness functions. *IEEE Trans. on Evolutionary Computation*, 4(4):373–379, 2000.

[43] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

Appendix: The Proof of the Second Part of Theorem 4

Here, we prove the second part of Theorem 4. Consider a random formula $\overline{F} = \overline{F}(n, c_1, c_2, c_3, c_4)$ with $\overline{D} > 1$ and the implication process in $\overline{F}$ starting from a literal x chosen uniformly at random. For $1 \le i \le 4$, let $\overline{F_i}(t)$ denote the conjunction of remaining 2-clauses or cherries of $\overline{F_i}$ at time t. Define $|\overline{F_i}(t)|$ to be the number of 2-clauses or cherries in $\overline{F_i}(t)$. Let $\overline{F}(t) = \overline{F_1}(t) \wedge \overline{F_2}(t) \wedge \overline{F_3}(t) \wedge \overline{F_4}(t)$. We first investigate the implication ratios in the process conditioned on $\overline{F_i}(t)$'s. Let $\overline{a}(t)$ and $\overline{c}(t)$ ($\overline{b}(t)$ and $\overline{d}(t)$, resp.) be the numbers of blue and red unit clauses produced by a blue (red, resp.) unit clause at time t. For $1 \le i \le 4$, let $\overline{b_{i,b}}(t)$ and $\overline{r_{i,b}}(t)$ ($\overline{b_{i,r}}(t)$ and $\overline{r_{i,r}}(t)$, resp.) be the numbers of blue and red unit clauses produced from $\overline{F_i}(t)$ at time t by a blue (red, resp.) unit clause. Then, $\overline{a}(t) = \sum \overline{b_{i,b}}(t)$, $\overline{c}(t) = \sum \overline{r_{i,b}}(t)$, $\overline{b}(t) = \sum \overline{b_{i,r}}(t)$, and $\overline{d}(t) = \sum \overline{r_{i,r}}(t)$.

For $\overline{b_{i,b}}(t)$'s, we see that $\overline{b_{1,b}}(t)$ has a binomial distribution $\mathrm{Bin}[|\overline{F_1}(t)|, \frac{1}{n-t}]$, $\overline{b_{2,b}}(t)$ is a linear combination of a random variable with a Bernoulli distribution with density $\frac{|\overline{F_2}(t)|}{2(n-t)}$ and a random variable with a binomial distribution $\mathrm{Bin}[|\overline{F_2}(t)|, \frac{1}{n-t}]$, $\overline{b_{3,b}}(t)$ has a Bernoulli distribution with density $\frac{|\overline{F_3}(t)|}{2(n-t)}$, and $\overline{b_{4,b}}(t)$ is two times a random variable that has a Bernoulli distribution with density $\frac{|\overline{F_4}(t)|}{2(n-t)}$. For $\overline{r_{i,b}}(t)$'s, $\overline{r_{1,b}}(t) = 0$ and $\overline{r_{2,b}}(t)$, $\overline{r_{3,b}}(t)$, and $\overline{r_{4,b}}(t)$ have binomial distributions $\mathrm{Bin}[|\overline{F_2}(t)|, \frac{1}{2(n-t)}]$, $\mathrm{Bin}[|\overline{F_3}(t)|, \frac{1}{2(n-t)}]$, and $\mathrm{Bin}[|\overline{F_4}(t)|, \frac{1}{n-t}]$, respectively. For $\overline{b_{i,r}}(t)$'s, $\overline{b_{1,r}}(t)$ and $\overline{b_{2,r}}(t)$ have binomial distributions $\mathrm{Bin}[|\overline{F_1}(t)|, \frac{1}{n-t}]$ and $\mathrm{Bin}[|\overline{F_2}(t)|, \frac{1}{n-t-1}]$, respectively, and $\overline{b_{3,r}}(t) = \overline{b_{4,r}}(t) = 0$. For $\overline{r_{i,r}}(t)$'s, $\overline{r_{1,r}}(t) = 0$ and $\overline{r_{2,r}}(t)$, $\overline{r_{3,r}}(t)$, and $\overline{r_{4,r}}(t)$ have binomial distributions $\mathrm{Bin}[|\overline{F_2}(t)|, \frac{1}{2(n-t-1)}]$, $\mathrm{Bin}[|\overline{F_3}(t)|, \frac{1}{2(n-t-1)}]$, and $\mathrm{Bin}[|\overline{F_4}(t)|, \frac{1}{n-t-1}]$, respectively.

Using similar arguments as in the proof of Theorem 2, we see that the implication ratio matrix at time t is almost always

$$\overline{T}(t) = \begin{bmatrix} c_1 + c_2 + \frac{c_2+c_3+2c_4}{2} & c_1 + c_2 \\ \frac{c_2+c_3+2c_4}{2} & \frac{c_2+c_3+2c_4}{2} \end{bmatrix} + o(1)$$

uniformly for all $1 \le t \le n^{1-\epsilon}$ for arbitrarily small $\epsilon > 0$. Consider the branching process with two types whose branching ratio matrix is

$$\overline{A} = (1-\delta) \begin{bmatrix} c_1 + c_2 + \frac{c_2+c_3+2c_4}{2} & c_1 + c_2 \\ \frac{c_2+c_3+2c_4}{2} & \frac{c_2+c_3+2c_4}{2} \end{bmatrix}.$$

The eigenvalues of $\overline{A}$ are

$$(1-\delta)\frac{c_1 + 2c_2 + c_3 + 2c_4 \pm \sqrt{(c_1+c_2)^2 + 2(c_1+c_2)(c_2+c_3+2c_4)}}{2}.$$

Let $\overline{\lambda}(\delta)$ be the larger one. Since $\overline{D} > 1$ implies $\overline{\lambda}(0) > 1$, we choose $\delta > 0$ so that $\overline{\lambda}(\delta) > 1$. Then, by coupling the implication process with the branching process and using Corollary 1, we see that,

for some constant κ, the κ-implication process starting from x proceeds to the supplemental round with positive probability. Conditioned that the κ-implication process proceeds to the supplemental round, the second moment method says that a 0-clause is almost always produced in the supplemental round. Hence, we have that the κ-implication process starting from x produces a 0-clause with positive probability. By considering successive κ-implication processes starting from $\Theta(\log n)$ different literals, we can show that a 0-clause is almost always produced from $\overline{F}$ in the same way as the proof of Theorem 2.