University of Alberta

Threshold Phenomena in NK Landscapes

by

Yong Gao ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Computing Science

Edmonton, Alberta
Spring 2001

0-612-60428-4

University of Alberta

Library Release Form

Name of Author: Yong Gao

Title of Thesis: Threshold Phenomena in NK Landscapes

Degree: Master of Science

Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Yong Gao

Yong Gao
Michener Park,
Edmonton, Alberta
Canada, T6H 5A1

Date: Jan 3, 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Threshold Phenomena in NK Landscapes** submitted by Yong Gao in partial fulfillment of the requirements for the degree of **Master of Science**.

Joseph Culberson
Joseph Culberson

Russell Greiner
Russell Greiner

Ryan B. Hayward
Ryan Hayward

M. Shirvani
Maziar Shirvani

Date: Dec/15/00

Abstract

In this thesis, we study the threshold phenomena in the NK landscape, a combinatorial model widely used in the study of genetic algorithms and population genetic dynamics. We establish two random models for the decision problem of the NK landscape model, called the *uniform probability model* and the *fixed ratio model* respectively.

The aim of the study is to investigate the hardness of the NK landscape model in terms of the theory of threshold phenomena and phase transitions. We show theoretically that the uniform probability model is trivially insoluble as the problem size tends to infinity. For the fixed ratio model, we establish two upper bounds of insolubility on the control parameter of the model above which the problems are asymptotically insoluble with probability 1. We show that instances with parameters above the upper bounds contain some easy subproblems such as 2-SAT, and hence can be solved by polynomial algorithms.

The fixed ratio model is also studied empirically. The experimental results show that there is a threshold phenomenon in the model and our upper bound on the threshold is tight. From the experiments, we also observe that random instances of the fixed ratio model are also typically easy in the soluble region and phase transition region.

Acknowledgements

There are three buttons on the mouse, and you can try things out by clicking on them.

I would like to thank my supervisor Joe Culberson who drew my attention to the phase transition aspect of the NK landscape model. His deep involvement, insightful comments and encouragement have been invaluable during the preparation of this thesis. Many parts of the study came from our weekly meeting.

I am grateful to my joint supervisor Russ Greiner. I have worked with him on the problems of optimal strategies for decision trees and the learning of Bayesian networks. I have learnt a lot from that experience.

Thanks are also given to my classmates in the algorithm group, Adam Beacham and Wenbin Ma, and to all of my friends at the department of computing science.

Finally, I would like to thank my daughter, my wife, and my parents for their patience and understanding. Year after year, it is them who share every piece of pressure, anxiety, and joy that comes to me.

Yong Gao

October, 2000
Edmonton, Canada

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Combinatorial search and optimization are fundamental problems in theoretical computer science. Many practical tasks in Artificial Intelligence(AI), computer architecture design, machine vision, database, and computer networks, can be formulated as search problems on combinatorial structures. The most popular and extensively studied combinatorial problems are Boolean Satisfiability(SAT), Graph Coloring(GC), Hamiltonian Cycle(HC), and Constraint Satisfaction(CSP).

From the perspective of computational complexity, all of these problems(and many other problems) are known to be in a class of decision problems called NP-complete problems(NPC), and it is commonly believed that this class of problems cannot be solved efficiently by any algorithms. However, there has been much debate about the implications of these NP-complete results. Researchers from AI and many other fields argue that the pessimistic NP-complete analysis is irrelevant at least to many classes of practical problems due to the worst-case-based nature of the analysis, and that what we really need is an average case study which shows the typical hardness of a class of problems and predicts for which part of the problem space an algorithm will perform well.

In recent years, much progress has been made in our understanding of the nature of the combinatorial search problems and the average/typical performance of the search algorithms. A common approach to the average-case analysis is to establish a random model for the problem and study the statistical properties of the instances randomly generated from the model. One of the most exciting findings is the discovery of the threshold phenomena and phase transitions that occur in many combinatorial search problems in an algorithmically independent manner. Roughly speaking, a phase transition in combinatorial search refers to the phenomenon that the probability that a random instance of the problem has a solution drops abruptly from 1 to 0 as the order parameter of the random model crosses a critical value called the *threshold*. Closely related to this phase transition in solvability is the hardness of solving the problems. There has been strong empirical evidence and theoretical arguments showing that the hardest instances of the problems usually occur around the threshold and instances generated with parameters far away from the threshold are relatively easy. The study of threshold phenomena and phase transitions is emerging as a new research field that attracts the interest of people from theoretical computer science, discrete mathematics, and statistical mechanics.

The study of threshold phenomena in combinatorial structures can be traced back to the work of Erdös and Renyi[ERR60]. Since then, it has been shown that many graph properties such as connectivity, Hamiltonicity, and colourability exhibit threshold behavior; i.e., the probability of their appearance changes dramatically from 1 to 0 in a very small interval of the parameter of the random graph model. It has also been proved that some NP-complete problems such as edge-colouring and Hamiltonian cycle can be solved in polynomial time with the probability asymptotic to 1 as the problem size tends to infinity.

Theoretical computer scientists are more interested in the correlation between problem hardness and the solvability threshold phenomena. Since the seminal work of Cheeseman et al. [CKT91], many NP-complete combinatorial search problems have been shown to have an easy-hard-easy pattern of hardness and the hardest instances usually occur around the solvability threshold [CMI97, FRE96, GEW96, MLE96, MSL92, SMH96]. In the past few years, researchers have developed theories to explain the occurrence of the easy-hard-easy patterns (see, for example, the work of Manasson et al on the *Backbone* and (2+p)-SAT [MOZ96, MOZ97, MZK99a, MZK99b] and Culberson and Gent's work on frozen development [CUL99a, CUL99b]). These theories provide a new way to characterize the nature of the computational complexity of NP-hard problems. The theories and results in the study of phase transitions in search have also been used to generate benchmark problems that can be used to evaluate newly designed algorithms [AGK00], including various random local search algorithms such as GSAT [SML92], simulated annealling [KGV83], and genetic algorithms [HOL92].

As a class of random search algorithms inspired by Darwin's theory of natural evolution, *genetic algorithms* gained their popularity in the past two decades. An important approach to studying genetic algorithms is *fitness landscape analysis*, which has long been used in the field of evolutionary biology since the work of Wright in 1932 [WRI32]. Roughly speaking, a fitness landscape is an objective function defined on a data structure together with the neighborhood structure introduced by some operators. The objective function is usually called the *fitness function*. Fitness landscape analysis deals with the interaction between the fitness values under a given neighborhood structure. This interaction is called the correlation structure (or epistasis).

The *NK landscape* is a fitness landscape model devised by Kauffman [KAU89] in which the "ruggedness" of the landscape can be tuned by changing some parameters of the correlation structure. In genetic algorithms, the NK landscape model has been used as a prototype in the analysis of the performance of different genetic operators and the effects of different encoding methods on the algorithm's performance. It has also been used as benchmark in the experimental evaluation of various new genetic algorithms.

In this thesis, we analyze the NK landscape model from the perspective of threshold phenomena and phase transitions. We establish two random models for the decision problem of NK landscapes and study the threshold phenomena and the associated hardness of the phase transitions in these two models, theoretically and experimentally. The thesis is organized as follows. In Chapter 2, basic concepts of random graphs and the propositional satisfiability problem are introduced. The study of threshold phenomena and phase transitions in random graphs and propositional satisfiability problem and NK landscapes is reviewed. In Chapter 3, we first establish two random models for the decision problem of NK landscapes in section 3.1, and then analyze the threshold phenomenon in one of the random models, the uniform probability model, and study the performance of several polynomial algorithms for this model in section 3.2. In section 3.3, we establish upper bounds on the threshold of the second random model, the fixed ratio model, and prove that random instances generated with the parameters above these upper bounds can be solved polynomially with probability asymptotic to one. In section 3.4, the analyses of sections 3.2 and 3.3 are generalized to the *generalized NK landscape* model. In section 3.5, we report our experimental results on typical hardness of the fixed ratio model. In Chapter 4, we conclude our investigation and discuss future research directions on this topic.

Chapter 2

Preliminaries and Previous Work

In this chapter, we introduce basic notation in the theory of random graphs, SAT, and NK landscapes, and review previous work on the phase transition of the SAT problem and on the analysis of NK landscapes. The phenomena of the phase transition has been discovered and discussed in several combinatorial problems such as graph coloring, constraint satisfaction, and SAT. The reason we choose the phase transition in SAT to review is that the NK landscape model is closely related to SAT problems.

2.1 Theory of Random Graphs

The study of the theory of random graphs started with Erdös and Renyi's seminal work [ERR60]. After forty years' development, random graph theory has become a mature field and provides lots of powerful theoretical results and tools for the study of phase transition in combinatorial search.

2.1.1 Random Graphs and Threshold Functions

A random graph is a probability space $(\mathcal{G}, P)$ where $\mathcal{G}$ is a family of graphs and P is a probability distribution on $\mathcal{G}$ [KAR95]. In the theory of random graphs, the most widely used random model is the so called *binomial random graph* $G(n, p_e)$ in which $\mathcal{G}$ is the set of all graphs with the same set of vertices $V =$

$\{v_1 \cdots, v_n\}$, and each of the $\binom{n}{2}$ edges is in $G(n, p_e)$ with the probability p_e and independent of other edges. The parameter p_e is called the *edge probability*. Another random graph model closely related to the binomial random graph $G(n, p_e)$ is the *uniform random graph* $G(n, m)$ in which $\mathcal{G}$ is also the set of all graphs with the same set of vertices V and the m edges of $G(n, m)$ are selected without replacement from the $\binom{n}{2}$ edges.

A *graph property* $\mathcal{A}$ is a subset of all the graphs. Examples of interesting graph properties include connectivity, colorability, and Hamiltonicity. In random graph theory, we are concerned with establishing the probability that a random graph has certain properties and investigating the analytical behavior of the probability as a function of some control parameters when the size n of the graph tends to infinity. A class of graph properties that are of great interest are the monotone properties. A graph property $\mathcal{A}$ is said to be increasing(decreasing) if $G \in \mathcal{A}$ and G is a spanning subgraph(supergraph) of a graph H implies that $H \in \mathcal{A}$. A graph property is called monotone if it is increasing or decreasing. For example, the property of connectivity is increasing, while the property of colorability is decreasing.

An important topic in random graph theory is the study of the *threshold phenomena* for (monotone) graph properties, i.e., the probability that a random graph $G(n, p_e)$ has a property changes dramatically from values close to 0 to values close to 1 at a certain critical value of the parameter p_e.

Definition 2.1.1. ([KAR95], Section 5) Let $\mathcal{A}$ be a monotone graph property and assume that the edge probability $p_e = p_e(n)$ is a function of n . A function $f_{\mathcal{A}} = f_{\mathcal{A}}(n)$ is called a threshold function for the property $\mathcal{A}$ if $\lim_{n\to\infty} Pr\{G(n, p_e(n)) \in \mathcal{A}\} = 0$ when $p_e(n) = o(f_{\mathcal{A}})$, and $\lim_{n\to\infty} Pr\{G(n, p_e(n)) \in \mathcal{A}\} = 1$ when $f_{\mathcal{A}} = o(p_e(n))$.

In 1986, Bollobás and Thomason [BOT86] proved the existence of the threshold function for the monotone graph property.

Theorem 2.1.1. *[BOT86] Every non-trivial increasing property $\mathcal{A}$ has a threshold function in $G(n, p_e)$.*

A recent progress in the study of the threshold phenomena in random graphs is the work of Friedgut on the *sharp threshold* [FRI99]. For a monotone graph property, the *threshold interval* is an interval of the edge probability in $[0, 1]$ in which the probability that a random graph has the property changes quickly from 0 to 1. Roughly speaking, a threshold associated with a monotone property is sharp if the length of the threshold interval decreases much faster than the threshold itself does. As a result, for a monotone graph property with a sharp threshold, the limit probability that a random graph has the property will be discontinuous as a function of the edge probability. In [FRI99], Friedgut introduced the concept of the sharp threshold for a graph property and established a very general sufficient conditional for a threshold to be sharp. Friedgut also used his sufficient condition to show the existence of a sharp threshold for the random SAT problem. In his PhD thesis [ACH99b], Achlioptas used Friedgut's sufficient condition to show the existence of a sharp threshold for the k-colorability graph property.

Although the above results are interesting in their own right, they do not give explicit expressions for the threshold functions. Locating the exact thresholds for various graph properties is still a challenging task and has attracted much attention.

2.1.2 Thresholds for Specific Graph Properties

Over the years, much work has been done in locating and bounding exact thresholds for many interesting graph properties, including connectivity, colorability, and Hamiltonicity. In the following, we summarize the existing results in this respect.

Connectivity The threshold for connectivity was first established in [ERR60] in which the following result was proved:

Theorem 2.1.2. *Let $G(n, p_e)$ be the binomial random graph with $p_e(n) = \frac{d}{n}$. For any $\epsilon > 0$,*

(1). If $d = 1 - \epsilon$, then with probability asymptotic to 1 all connected components have at most one cycle and $O(logn)$ vertices;

(2). If $d = 1 + \epsilon$, then with probability asymptotic to 1 there exists a unique connected component with many cycles and $\Omega(n)$ vertices.

The above result has a vivid exposition in the framework of evolving graph processes, which shows how the threshold phenomena in random graphs is analogous to the phase transition undergone by water and other materials in the physical world. See [KAR95] for the details.

Hamiltonicity The existence of the threshold function for Hamiltonicity was proposed as an open problem by Erdös and Renyi. About 20 years later, using the methods of rotation and extensions, Komlos and Szemerédi [KOS83] were able to prove the following result and settle the open problem:

Theorem 2.1.3. *Let $G(n, m)$ be the uniform random graph with $m = n(\log n + \log\log n + c_n)/2$. Then,*

$$\lim_{n\to\infty} Pr\{G(n, m) is\ Hamiltonian\} = \begin{cases} 0, & if\ c_n \to -\infty, \\ e^{-e^{-c}}, & if\ c_n \to c, \\ 1, & if\ c_n \to \infty. \end{cases} \qquad (2.1.1)$$

Researchers have also come up with polynomial algorithms which solve the Hamiltonian Cycle problem with probability asymptotic to 1. See [FRM97] for an overview. The recent work of Vandegriend and Culberson [VCU98] further shows that the Hamiltonian cycle problem is not as hard as many people think even in the phase transition region—a surprise to people who get to know the theory of phase transitions from the work of Cheeseman, Kanefsky, and Taylor [CKT91]. (The main result of the current thesis is yet another surprise to people who are enthusiastic in solving NP-hard problem using genetic algorithms.)

Graph Colorability The graph coloring problem is one of the most important problems in graph theory. In random graph theory, the chromatic number of the random graph $G(,n,p_e)$ has been studied for more than 25 years. The problem of a threshold for colorability property can be stated as follows:

Question Let $k \geq 2$ be a positive integer and $G(n, p_e(n))$ be a random graph with $p_e(n) = \frac{d}{n}$ for some positive constant d. Define

$$d_k = sup\{d| \lim_{n \to \infty} Pr\{G(n, p_e(n)) \text{ is k-colorable}\}\}. \tag{2.1.2}$$

The problem is to determine the exact value of d_k or to bound it from above (and/or below).

It turns out that the threshold problem for graph colorability is much more difficult than those for connectivity and Hamiltonicity. We still do not know the exact value of d_k except that $d_2 = 1$. In recent years, researchers have come up with many approaches to establish upper and lower bounds for d_k.

In his PhD thesis [ACH99b], Achlioptas proved that the threshold for the k-colorability is sharp for $k \geq 3$ and showed that there exists a function $t_k(n)$ such that for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} Pr\{G(n, \frac{t_k(n) - \epsilon}{n}) \text{ is k-colorable}\} = 1 \text{ and}$$

$$\lim_{n \to \infty} Pr\{G(n, \frac{t_k(n) + \epsilon}{n}) \text{ is k-colorable}\} = 0.$$

The problem with Achlioptas's result is that $t_k(n)$ depends on n and its explicit expression remains unknown. Actually, we do not even know if the limit $\lim_{n \to \infty} t_k(n)$ exists.

In recent years, researchers have come up with many approaches to establish upper and lower bounds for d_k. Erdos and Renyi [ERR60] obtained the first lower bound $d_3 \geq 1$ based on the observation that all the connected components of $G(n, \frac{1-\epsilon}{n})$ asymptotically have at most one cycle. Luczak [LUC91] proved $d_3 \geq 1.0001$ by showing that $G(n, \frac{1.0001}{n})$ is 3-colorable. Chvátal improved the lower bound to $d_3 \geq 2.88$ in [CHV91]. In 1996, Pittel, et al.

[PSW96] obtained a better lower bound $d_3 \geq 3.35$. This lower bound was improved in [ACH99b] to $d_3 > 3.847$ which, to my knowledge, is the best result so far.

Using the first moment method, the upper bound $d_3 \leq 5.420$ can be established without much difficulty([ACH99b, CHV91]). Molloy and Reed in [MOL92] obtained the upper bound $d_3 < 5.15$. The best upper bound $d_3 < 5.044$ belongs to Achlioptas [ACH99b] and is established by using a generalized form of the first moment method of Kirousis, et al. [KKK98] which was developed to upper bound the threshold for SAT problem.

2.1.3 Probability Tools for the Study of Threshold Phenomena

Probability theory provides many powerful tools for the study of threshold phenomena in graphs and combinatorial search, including various probabilistic inequalities, martingale methods, and limit theorems. The most fundamental tools are the *first moment method* and the *second moment method.*

The first moment method is based on Markov's Inequality: For any non-negative random variable X, we have

$$Pr\{X \geq t\} \leq \frac{E[X]}{t}, \quad \text{for } t > 0. \qquad (2.1.3)$$

The first moment method is typically used to prove the non-existence of a special pattern in a random structure. For example, let X be the number of triangles contained in the binomial random graph $G(n, p_e)$. Since it is easy to see that $E[X] = \binom{n}{3} p_e^3 \sim (np_e)^3/6$, we know that $G(n, p_e)$ asymptotically contains no triangle if $np_e < 1$ for sufficient large n. As another example, consider a 3-SAT formula with the clause-to-variable ratio $r = m/n$. Letting X be the number of assignments to the variables that satisfy the 3-SAT formula, we have $E[X] = 2^n(7/8)^{rn}$. Therefore, for $r > log_{8/7}2 \approx 5.191$, the 3-SAT formula is asymptotically unsatisfiable with probability 1. This is actually the first upper bound for the 3-SAT phase transition in the literature [FRP83]. The following result on the existence of high girth and high chromatic number illustrates how powerful the first moment method is.

Theorem 2.1.4. *(See [MOL98]) For any $g, k \geq 1$, there exists graphs with no cycles of length at most g and with chromatic number greater than k.*

This result was first obtained by Erdos [ERD59]. The basic idea is to use the First Moment method to show that the probability for a random graph to have a large chromatic number and contain only short cycles is greater than 0. It took more than ten years for researchers to come up with a non-probabilistic construction to prove the above result [LOV68].

A problem with the first moment method is that the bounds obtained are usually not very tight. To see why this happens, observe that $E[X] = \sum_{k=1}^{\infty} Pr\{X \geq k\} = Pr\{X > 0\} + \sum_{k=2}^{\infty} Pr\{X \geq k\}$. The first moment method proves that $Pr\{X > 0\}$ tends to zero by showing that $E[X]$ tends to zero. This works because $Pr\{X > 0\} < E[X]$. But it is possible that $Pr\{X > 0\}$ itself tends to zero while the term $\sum_{k=2}^{\infty} Pr\{X \geq k\}$ does not. In [KKK98], a generalized first moment method is proposed and used to establish a tighter upper bound for the 3-SAT phase transition. The same approach was also used in [ACH99b] to obtain a tighter upper bound for the graph coloring phase transition.

The second moment method is based on Chebyschev's inequality: For a random variable X and positive number t,

$$Pr\{|X - E[X]| \geq t\} \leq \frac{var(X)}{t^2} \tag{2.1.4}$$

where $var(X) = E[(X - E[X])^2]$ is the variance of X.

We usually use the second moment method to prove the existence of a special kind of pattern in a random structure. This is done by letting X be the number of the patterns contained in the random structure and showing that $E[X] > 0$ and $var(X) = o(E[X])$.

2.2 Phase Transition in SAT

The propositional satisfiability (SAT) problem plays an important role in computational complexity. It was the first problem shown to be NP-complete and many of the NP-complete proofs of other problems are based on a reduction from SAT. There are lots of NP-hard problems that can be transformed efficiently to SAT. SAT problems are used as benchmarks in testing the efficiency of search and optimization algorithms. Random SAT is also one of the NP-complete problems in which the phenomenon of phase transition is observed and studied. In this section, we introduce the basic concepts of SAT problems and review existing results on the SAT phase transition.

2.2.1 Propositional Satisfiability Problem(SAT)

A SAT problem is the problem of determining if there is an assignment of truth values to the variables of a formula in conjunctive normal form. Throughout this section, let $X = (x_1, \cdots, x_n)$ be a set of Boolean variables and $L = (x_1, \bar{x}_1, \cdots, x_n, \bar{x}_n)$ be the set of positive and negative literals on X. Sometimes, we abuse the notation for the variable x_i to represent the positive literal and $\bar{x}_i$ for the negative literal. Generally, given a literal u in L, we use $\bar{u}$ to denote its negative. The variable underlying the literal u is denoted as $|u|$. Two literals u and v are said to be *variable-distinct* if $|u| \neq |v|$.

A clause is a disjunction of a set of literals and a formula ϕ is the conjunction of set of clauses. The size of a clause is the number of literals in it. A truth assignment is a binary mapping defined on the set of literals L. A formula is satisfiable if and only if there is a truth assignment such that for every clause, at least one of the literals is true under the assignment. The k-SAT problem can be stated as follows: Given a formula ϕ with the size of each of its clauses being k, is there a truth assignment that satisfies ϕ?

It is well-known that the k-SAT problem is NP-complete for $k \geq 3$, while the 2-SAT problem is polynomially solvable. Beside the 2-SAT problem, researchers have also identified many classes of k-SAT ($k \geq 3$) problems which are polynomially solvable. The details can be found in [FRG98] and the references therein. For the general NP-complete k-SAT problem, many algorithms have been developed. A good survey on the algorithms for SAT problems can be found in [GUJ96].

2.2.2 Phase Transition in SAT

(1) Random Models for SAT

In the study of the threshold phenomena and phase transition of the SAT problem, three random models are widely used. They are the *constant-density model*, the *fixed-length model*, and the $2+p$ model [FRP83, FRS95a, GEW96, MLE96, MZK99a, MZK99b, SMH96].

The constant-density model, $\overline{M}(n, m, p)$, has three parameters: the number of variables n, the number of clauses m, and the probability p. An instance contains l clauses, and each clause is generated in such a way that it contains each of the $2n$ literals with probability p. It can be seen that the size of clauses in the constant-density model is random and the average is $2np$. The fixed-length model, $M(n, m, k)$, also has three parameters: the number of variables n, the number of clauses m, and the size of the clauses k. In $M(n, m, k)$, the m clauses all have the size k and are chosen uniformly, independently and with replacement among all $2^k\binom{n}{k}$ non-trivial clauses of size k. The $2+p$ model, $M(n, m, 2+p), 0 \leq p \leq 1$, is introduced by Monasson et al [MZK99a, MZK99b] and has attracted lots of attention lately. $M(n, m, 2+p)$ can be viewed as a mixture of $M(n, m, 2)$ and $M(n, m, 3)$: pm clauses are chosen from the set of all clauses of size 3 and $(1-p)m$ clauses are chosen from the set of all clauses of size 2.

(2) Early Work on $\overline{M}(n, m, p)$

The study of the average-case complexity of SAT started with Goldberg's controversial thesis [GOL79]. Goldberg claimed that under a random model similar to $\overline{M}(n,m,p)$, SAT problems can be readily solved on average in $O(n^2)$ time. It is proved by Franco et al. (see [FRS95a]) that in $\overline{M}(n,m,p)$, "virtually the entire parameter space is covered by a collection of polynomial time algorithms that find solutions to random instances with probability tending to 1 as instance size increases". To put it in another way, this means that Goldberg's $O(n^2)$ result was not because he had a good algorithm, but because the random model used by Goldberg is very unlikely to be able to generate really hard SAT instances.

(3) Threshold Phenomena in Polynomial Classes

Due to the work of Goerdt [GOE92], Chvatal and Reed [CHR92], and Bollobäs et al. [BBC99], the threshold phenomenon in random 2-SAT $M(n,m,2)$ is well understood. The phase transition occurs at $\frac{m}{n} = 1$ [CHR92, GOE92], i.e.,

$$\lim_{n\to\infty} Pr\{M(n,m,2) \text{ is satisfiable}\} = \begin{cases} 0, & \text{if } \frac{m}{n} > 1 \\ 1, & \text{if } \frac{m}{n} < 1. \end{cases} \qquad (2.2.1)$$

In [BBC99], the exact finite-size scaling of the $M(n,m,2)$ is determined which gives us the asymptotic behavior of the random 2-SAT problem inside the phase transition interval. In [IST98], the exact scaling behavior of random Horn formula is derived. In [FRG98], threshold functions are established for a random k-SAT $M(n,m,k)$ to be Horn formula(q-Horn formula, matched formula, or SLUR formula).

(4) Phase Transition in $M(n,m,k)$

The study of the phase transition in $M(n,m,k)$ provides a real challenge. The problem is to prove the following conjecture and determine the threshold d_k: There is a constant d_k such that,

$$\lim_{n\to\infty} Pr\{M(n,m,k) \text{ is satisfiable}\} = \begin{cases} 0, & \text{if } \frac{m}{n} > d_k, \\ 1, & \text{if } \frac{m}{n} < d_k. \end{cases} \qquad (2.2.2)$$

[FRI99] proves the existence of a function $d_k(n)$ that satisfies the above equality and leaves it as an open question that 'Does $d_k(n)$ converge as n tends to infinity?'. Based on the empirical study carried out since the early 1990's [CMI97, MSL92], it is believed that the threshold d_k is around 4.20.

The first upper bound $d_3 \leq 5.191$ for the threshold of the random 3-SAT $M(n, m, 3)$ is obtained by Franco and Paul [FRP83]. In [BFU93], Franco and Paul's upper bound is improved to $d_3 \leq 5.191 - 10^{-7}$. New upper bounds were obtained two years later, independently by El-Maftouhi and Fernandez de la Vega [MAV95] ($d_3 \leq 5.08$) and Kamath et al. [KRP95] ($d_3 < 4.758$). It took another two years for Kirousis et al. to come up with a generalized first moment method and obtain a tighter upper bound $d_3 < 4.601$ [KKK98] in 1998. This upper bound was then improved to $d_3 < 4.596$ by Janson, Stamation and Vamvakari [JSV99]. The best upper bound $d_3 < 4.506$ belongs to Dubios, Boufkahad and Mandler [DBM00].

Lower bounding the threshold for $M(n, m, k)$ is even more difficult. Franco is also the first to obtain a lower bound. In [FRA84], he showed that for $\frac{m}{n} < 1$, the pure literal heuristic eventually sets all variables with probability asymptotic to 1, and thus obtained the lower bound $d_3 \geq 1$. Using the same approach, Broder, Frieze, and Upfal [BFU93] improved the lower bound to $d_3 \geq 1.63$. By introducing two new heuristics and analyzing their probabilities of success on $M(n, m, 3)$, Frieze and Suen [FSS96] proved that $d_3 \geq 3.003$. The best lower bound belongs to Achlioptas [ACH99a], who proved that $d_3 > 3.145$ by an improvement to the heuristics of Frieze and Suen [FSS96].

(5) The $(2+p)$ Model $M(n, m, 2+p)$ The $2+p$ model was introduced by Monasson et al. [MZK99a, MZK99b] in an effort to understand the difference between the phase transition in 2-SAT(polynomial problem) and 3-SAT (NP-complete problem). In a series of papers, Monasson et al. studied the threshold phenomenon of random SAT problem in the framework of statistical mechanics [MOZ96, MOZ97, MZK99a]. By introducing the concept of *backbone*—the subset of variables that are forced to be true/false in a formula, they were able to come up with a new characterization of the phase transition behavior in random SAT problems to help understand the relationship between satisfiable-to-unsatisfiable phase transitions and the associated easy-hard-easy patterns. They discovered that at the threshold, the ratio of the size of the backbone to the problem size n changes continuously for random 2-SAT problems, but discontinuously for random 3-SAT problems. A similar approach, called frozen development, has also been developed by Culberson and Gent [CUL99a, CUL99b] in their study on the phase transition of graph coloring.

In [MZK99a], it is proved that for each $0 \leq p \leq 1$, there is a critical value r_p around which $M(n, rn, 2+p)$ undergoes a satisfiable-to-unsatisfiable phase transition. An even more interesting thing to consider is the critical value p_c such that $M(n, m, 2+p)$ behaves like a random 2-SAT for $p < p_c$, and like a 3-SAT for $p \geq p_c$. This critical value can be defined as [ACH99b]:

$$p_c = sup\{p : \lim_{n \to \infty} Pr\{M(n, \lambda n/(1-p), 2+p) \text{ is satisfiable}\} = 1, \ \forall \lambda < 1\}.$$

In [MZK99a], a lower bound $p_c \geq 0.4$ was established based on statistical mechanics method and a numerical upper bound $p_c \leq 4.1$ was observed. In [ACH99b], Achlioptas found a theoretical upper bound $p_c \leq 0.695$. Achilioptas also conjectured that $p_c = 0.4(!)$ and provided a detailed discussion about the reasonableness of his conjecture [ACH99b]. It is hoped that studies along this line will shed light on the difference between P problems and NP-complete problems, and help us tackle the famous conjecture $P \neq NP$.

2.3 Analysis of NK Landscape

The notion of *fitness landscape* was first introduced by Wright in 1932 [WRI32], and since then has been used as a metaphor in the analysis of population genetic dynamics. Biological organisms can be viewed by their *genotype*, which is the genetic encoding of the organisms, and by their *phenotype* which represents the actual form and behavior of the organisms. For each phenotype, there is an associated fitness value which abstracts the phenotype's ability to survive and reproduce. The evolution and dynamics of a population can thus be viewed as a process that searches the landscape of fitness in order to find phenotypes with higher fitness values. A fitness landscape characterizes the interactions between genes and its effects on the overall fitness of a population. An interesting kind of interaction between genes is called *epistasis*, where the effect on fitness from altering one gene depends on the state of other genes.

In the combinatorial search and function optimization perspective, a fitness landscape can be viewed as an objective function defined on binary strings where the function values and variables have a specific *correlation structure*. Over the last two decades, the idea of fitness landscape analysis has been used as a prototype in the study of classes of search and optimization algorithms that are motivated from nature. Examples of such algorithms are simulated annealing [KGV83] and genetic algorithms [HOL92].

The NK landscape is a fitness landscape model, devised by Kauffman [KAU89], in which the "ruggedness" of the landscape can be tuned by changing some control parameters. NK landscapes have been used as a prototype and benchmark in the theoretical and empirical study of genetic algorithms. Before discussing previous work on the analysis of NK landscapes in genetic algorithms, let us first establish the formal definition of the NK landscape.

Definition 2.3.1. An NK landscape f is a real-valued function defined on binary string of fixed length,

$$f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)),$$

where $n > 0$ is a positive integer and $x = (x_1, \cdots, x_n) \in \{0,1\}^n$. The NK landscape f is the sum of n *local fitness functions* $f_i(x_i, \Pi(x_i))$, $1 \leq i \leq n$. Each local fitness function $f_i(x_i, \Pi(x_i))$ depends on its *main variable* x_i and its neighborhood $\Pi(x_i) \subset \{x_1, \cdots, x_n\} \backslash \{x_i\}$. The main parameters of a NK landscape are n, the number of variables, and the size of the neighborhood $k = |\Pi(x_i)|$.

In an NK landscape, the neighborhood $\Pi(x_i)$ can be chosen in two different ways: the *adjacent neighborhood*, where k variables with indices nearest to i (modulo n) are chosen, and the *random neighborhood*, where the k variables are randomly chosen from the set $\{x_1, \cdots, x_n\} \backslash \{x_i\}$. Once the variables in the neighborhood are determined, the local fitness function f_i is determined by a fitness lookup table which specifies the function value f_i for each of the 2^{k+1} possible assignments to the variables x_i and $\Pi(x_i)$. Usually, the function values $f_i(s), s \in \{0, 1\}^{k+1}$ are obtained by independently sampling some probability distributions such as the uniform distribution on $[0, 1]$ and the Bernoulli distribution on $\{0, 1\}$. See [ALT96] for a detailed discussion on the fitness lookup table.

NK landscapes have been studied from the perspectives of statistics and computational complexity. In the rest of this section, we discuss previous work on the analysis of the NK landscape as an optimization model.

The statistical analysis of the NK landscape characterizes the NK landscape in terms of the distribution of the local optima, the average distance between these local optima, and the way in which function values at different points correlate with each other.

It should be noted that these characteristics depend on how the topological structure of the space $\{0, 1\}^n$ is defined. In the point of view of local search algorithms, this means that the characteristics of an NK landscape depend on how the search algorithms generate new solutions and move in the search space. As an example, consider the point mutation and crossover operators used in genetic algorithms. The point mutation operator generates a new solution by flipping a randomly chosen bit in the solution's binary representation, and thus views the space $\{0, 1\}^n$ with the topology defined by the Hamming distance (the number of different bits of two binary strings). On the other hand, the crossover operator generates new solutions by exchanging segments of two binary strings which defines a totally different topology in the search space. We refer the readers to the work [JON95, CUL96] for detailed discussions.

Most of the theoretical analyses of the NK landscape are conducted on the topology induced by the point mutation [ALT96, WEI90, WEI91, WEI96]. The typical approach to analyzing the NK landscape under point mutation topology is to study the statistical characteristics of a random walk on the space $\{0, 1\}^n$.

Let $\{x(i), i \geq 1\}$ be the random walk on $\{0, 1\}^n$ and $y(i) = f(x(i))$. Then $\{y(i), i \geq 1\}$ defines a time series on $\{0, 1\}^n$. In his series of work [WEI90, WEI91, WEI96], Weinberger introduces the *auto-correlation function* and the *correlation length* of the time series $\{y(i), i \geq 1\}$ as the measures of the correlation structure of the NK landscape. Weinberger's work gives us a clear picture about the correlation structures of the NK landscape under the Hamming topology. For $k = 0$, each local fitness function only depends on the main variables and the NK landscape has one global optimum which can be found in the expected $n/2$ steps by hill-climbing algorithms under the Hamming topology. For $k = n - 1$, each local fitness function depends on all the n variables and the overall fitness values are statistically independent. This makes the NK landscape have many local optimum each of which has a small basin of attraction.

Weinberger's approach has also been used to examine the fitness landscapes under the topologies such as those generated by crossover operators [MDS91]. In [HOR97], Hordijk developed a complete time-series analysis on the fitness time series $\{y(i), i \geq 1\}$ in which the fitness time series is fitted to some well-known time series models such as the auto-regressive moving average(ARMA) model. The obtained model can then be used to make predictions or perform simulations. In his PhD thesis [JON95], Jones established a generalized fitness landscape model with which the algorithmic and statistical characteristics of the fitness landscapes can be studied under different topologies induced by operators in genetic algorithms. By using an algorithm called *reverse hillclimbing* and the proposed fitness landscape model, Jones examined the properties of different NK landscapes such as the number of peaks, the sizes of the basins of attraction, and the lengths of paths to discover peaks.

The computational complexity of the NK landscapes has also been investigated. In [WEI96], it is shown that NK landscape models with the adjacent neighborhood are polynomially solvable, while NK landscapes with the random neighborhood are NP complete for $k \geq 3$. The computational complexity of NK landscapes are further studied by Wright et al. In [WTZ99] which summarizes their research since 1995, Wright et al. show that the NK landscape with random neighborhood is NP-complete for $k \geq 2$ and polynomial solvable for $k = 1$. Besides the above theoretical studies, NK landscapes have been extensively used as a benchmark in the experimental investigation of the performance of genetic algorithms [EIB96, JON95, POT98]. In a certain sense, a class of widely used testing functions called deceptive functions in the genetic algorithms [WHI91] can be viewed as variations of the NK landscape model.

These NP-complete results bring up an interesting question with regard to the difference between NK landscapes with random neighborhood and adjacent neighborhood: Since NK landscapes with these two neighborhood structures are statistically similar [KAU93, WEI96], what is the real difference between the class of NP-complete NK landscapes and the class of polynomial NK landscapes? This thesis tries to answer this question by analyzing the threshold phenomena of different random models of NK landscapes.

Chapter 3

Threshold Phenomena in NK Landscapes

In this chapter, we study the threshold phenomena in NK landscapes. In section 3.1, we establish two random models for the decision problem of NK landscapes, called *the uniform probability model* and *the fixed ratio model.* In sections 3.2-3.4, we study the threshold phenomena and phase transitions under these two random models theoretically and empirically. It is proved that the phase transition of the uniform probability model is easy in the sense that there is a polynomial algorithm that can solve a random instance of the problem with the probability asymptotic to 1 as the problem size tends to infinity. For the fixed ratio model, we establish several upper bounds for the solvability threshold, and prove that random instances with parameters above these upper bounds can be solved linearly or polynomially. This, together with our empirical study for random instances generated below and in the phase transition region, suggests that the phase transition of the fixed ratio model is also easy.

3.1 Random Models for NK Landscapes

Throughout this chapter, we consider the NK landscape model

$$f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)), \quad x = (x_1, \cdots, x_n) \in \{0,1\}^n \tag{3.1.1}$$

with the size of the neighborhood $k = |\Pi(x_i)|$. As NK landscape problems with adjacent neighborhoods are in the class P [WEI96], we will concentrate on NK landscapes with random neighborhoods. To simplify the discussion, we further assume that the local fitness functions take on binary values, i.e., for each $1 \leq i \leq n$, we have $f_i(x_i, \Pi(x_i)) \in \{0, 1\}, \forall x \in \{0, 1\}^n$.

Definition 3.1.1. Let $f_i = f_i(x_i, \Pi(x_i))$ be the local fitness function for the variable x_i. For each assignment $(x_i, \Pi(x_i)) \in \{0, 1\}^{k+1}$, $f(x_i, \Pi(x_i))$ is a Bernoulli random variable. The fitness distribution of the NK landscape model is the joint probability distribution P_f on $\{0, 1\}^{k+1}$ of the 2^{k+1}-dimensional Bernoulli random vector $\{f_i(x_i, \Pi(x_i)), (x_i, \Pi(x_i)) \in \{0, 1\}^{k+1}\}$.

Remark 3.1.1. It is possible to study NK landscape models in which different local fitness functions have different fitness distributions. But as a common practice in the study of NK landscape models, we assume throughout this chapter that all the local fitness functions have the same fitness distribution. We therefore use the notation P_f in the above definition to indicate that all the local fitness functions have the same fitness distribution.

Definition 3.1.2. The decision problem of an NK Landscape model with random adjacent neighborhoods

$$f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$$

is defined as: Is the optimum of $f(x)$ equal to n? An NK landscape decision problem is insoluble if there is no solution for it.

It has been proved in [THO95, WEI96, WTZ99] that the decision problem of the NK landscape model is NP complete for $k \geq 2$. The proofs were based on a reduction from SAT to the decision problem of NK landscapes. To study the typical hardness of the NK landscape decision problems in the framework of thresholds and phase transitions, we introduce two random models that have different fitness distributions. In both of the models defined below, the neighborhood set $\Pi(x_i)$ of a variable x_i is selected by randomly choosing without replacement $k = |\Pi(x_i)|$ variables from $x \backslash \{x_i\}$.

Definition 3.1.3. (The Uniform Probability Model $\overline{N}(n, k, p)$) In this model, the fitness value of the local fitness function $f_i(x_i, \Pi(x_i))$ is determined as follows: For each assignment $y \in Dom(f_i) = \{0, 1\}^{k+1}$, let $f_i(y) = 0$ with the probability p and $f_i(y) = 1$ with the probability $1 - p$, and this is done for each possible assignment and each local fitness function independently.

Remark 3.1.2. In the uniform probability model, the fitness values $f_i(y), y \in \{0, 1\}^{k+1}$ are assigned independently. Thus, the fitness distribution P_f is an independent 2^{k+1}-dimensional Bernoulli distribution. It follows that each local fitness function has on average $2^{k+1}p$ zeroes.

Definition 3.1.4. (The Fixed Ratio Model $N(n, k, z)$) In this model, the parameter z takes on values from $[0, 2^{k+1}]$. If z is an integer, we specify the local fitness function $f_i(x_i, \Pi(x_i))$ by randomly choosing without replacement z tuples of possible assignments $Y = (y_1, \cdots, y_z)$ from $Dom(f_i) = \{0, 1\}^{k+1}$, and defining the local fitness function as follows:

$$f_i(y) = \begin{cases} 0, & \text{if } \ y \in Y; \\ 1, & \text{otherwise.} \end{cases}$$

For a non-integer $z = (1 - \alpha)[z] + \alpha[z + 1]$, we choose randomly without replacement $[(1 - \alpha)n]$ local fitness functions and determine their fitness values according to $N(n, k, [z])$. The rest of the local fitness functions are determined according to $N(n, k, [z] + 1)$.

Remark 3.1.3. In the fixed ratio model, the 2^{k+1}-dimensional Bernoulli random vector $\{f_i(x_i, \Pi(x_i)), (x_i, \Pi(x_i)) \in \{0, 1\}^{k+1}\}$ is not independent. For an integer z, the joint distribution of this Bernoulli random vector is a uniform distribution in the space

$$\{(r_i, 1 \leq i \leq 2^{k+1}) \in \{0,1\}^{2^{k+1}} : \sum_{i=1}^{2^{k+1}} r_i = z\}.$$

There are exactly $[\alpha n]$ local fitness functions which have $[z]+1$ randomly assigned zero values and $n-[\alpha n]$ local fitness functions which have $[z]$ randomly assigned zero values.

We conclude this section by establishing a relation between the decision problem of NK landscapes and the SAT problem. A decision problem of the NK landscape

$$f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)),$$

"is the optimum of $f(x)$ equal or greater than n?", can be reduced to a (k+1)-SAT problem as follows:

(1) For each local fitness function $f_i(x_i, \Pi(x_i))$, construct a conjunction $C_i = \bigwedge_{j=1}^{z} C_i^j$ of clauses with exactly $k+1$ variable-distinct literals from the set of variables $\{x_i, \Pi(x_i)\}$, where z is the number of zero values that f_i takes and C_i^j is such that for any assignment $y_j \in \{0,1\}^{k+1}$ that falsifies C_i^j, we have $f_i(y_j) = 0$.

(2) The (k+1)-SAT is the conjunction $\varphi = \bigwedge_{i=1}^{n} C_i$.

Table 3.1 shows an example of the fitness assignment of a local fitness function $f_i = f_i(x, y, z)$ and its associated equivalent 3-SAT clauses. It is easy to see that for any assignment s to the variables x, y, z, $f_i(s) = 1$ if and only if the assignment satisfies the formula

$$x \vee y \vee z,\ x \vee \bar{y} \vee \bar{z},\ \bar{x} \vee y \vee \bar{z},\ \bar{x} \vee \bar{y} \vee z.$$

Let $\widehat{M}(n,2,p)$ and $\widetilde{M}(n,2,z)$ denote respectively the random SAT problems derived from the NK models $\overline{N}(n,2,p)$ and $N(n,2,z)$. We have

(1)An instance of $\widehat{M}(n,2,p)$ is a sample from the probability space $(\prod_{i=1}^{n} C_i, \prod_{i-1}^{n} P_i)$, where $C_i = \bigcup_{i_1,i_2} L_i(i_1,i_2)$ with $L_i(i_1,i_2)$ being the set of all size-2 clauses with

x	y	z	f_i	Clauses
0	0	0	0	$x \vee y \vee z$
0	0	1	1	
0	1	0	1	
0	1	1	0	$x \vee \bar{y} \vee \bar{z}$
1	0	0	1	
1	0	1	0	$\bar{x} \vee y \vee \bar{z}$
1	1	0	0	$\bar{x} \vee \bar{y} \vee z$
1	1	1	1	

Table 3.1: A local fitness function and its equivalent 3-clauses.

exactly 3 variable-distinct literals from $\{x_i, x_{i_1}, x_{i_2}\}$, and P_i is a probability distribution on C_i such that for each CNF set $c \in C_i, |c| = z$, we have $P_i(c) = \frac{\binom{8}{z}p^z(1-p)^{8-z}}{\binom{n-1}{2}}$.

(2) An instance of $\widetilde{M}(n, 2, z)$ is a sample from the probability space $(\prod_{i=1}^{n} C_i, \prod_{i-1}^{n} P_i)$, where $C_i = \bigcup_{i_1, i_2} L_i(i_1, i_2)$ with $L_i(i_1, i_2)$ being the set of clauses with exactly 3 variable-distinct literals from $\{x_i, x_{i_1}, x_{i_2}\}$, and P_i is a uniform probability distribution on C_i. Note that $|C_i| = \binom{8}{z}\binom{n-1}{2}$.

3.2 Analysis of the Uniform Probability Model

In the uniform probability model $\overline{N}(n, k, p)$, the parameter p determines how many zero values a local fitness function can take. We are interested in how the solvability and hardness of the NK landscape decision problem change as the parameter p increases from 0 to 1. It turns out that for fixed p, the decision problem is asymptotically trivially insoluble. This is quite similar to the phenomena in the random models of the constraint satisfaction problem observed in [AKK97]. For the case that $p = p(n)$ is a function of the problem size n, we will derive threshold functions and polynomial algorithms, showing that the problem is still asymptotically trivial. Let us first look at the case where the parameter p is fixed.

Theorem 3.2.1. *For each fixed* $0 < p < 1$ *and* k*, we have*

$$Pr\{\overline{N}(n,k,p) is\ soluble\} \leq (1 - p^{2^{k+1}})^n,$$

and hence

$$\lim_{n\to\infty} Pr\{\overline{N}(n,k,p) \text{ is soluble}\} = 0.$$

Proof: Let A be the event that the random NK decision problem $\overline{N}(n,k,p)$ is soluble, and A_i be the event that $f_i(y) = 0$, for each possible assignment $y \in \{0,1\}^{k+1}$. Then we have $A \subset \bigcap_{i=1}^{n} A_i^c$, where A_i^c denotes the complement of the set A_i. Since the fitness values of the local fitness functions are assigned independently, we have

$$\begin{aligned} Pr\{A\} &\leq Pr\{\bigcap_{i=1}^{n} A_i^c\} = \prod_{i=1}^{n}(1 - Pr\{A_i\}) \\ &= (1 - p^{2^{k+1}})^n. \end{aligned}$$

Since $0 \leq p \leq 1$, it follows that $\lim_{n\to\infty} Pr\{\overline{N}(n,k,p)$ is soluble$\} = 0$. □

Theorem 3.2.1 implies that under the uniform probability model $\overline{N}(n,k,p)$, the decision problem of NK landscapes is almost surely trivial as n tends to infinity if p and k are fixed. Actually, the simple linear-time algorithm A1 in Table 3.2 will solve the decision problem with probability asymptotic to 1 as long as $p = p(n)$ does not decrease too fast.

Corollary 3.2.2. *Assume that in the random NK landscape model* $\overline{N}(n,k,p)$, k *is fixed,* $p = p(n)$ *is a function of* n*, and* $\lim_{n\to\infty} p(n)$ *exists. We have*

$$\lim_{n\to\infty} Pr\{Algorithm\ A1\ succeeds\} = 1$$

if $\lim_{n} p(n) n^{\frac{1}{2^{k+1}}} = +\infty$.

```
Algorithm A1
Input: An instance of $\overline{N}(n,k,p)$, $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$;
Output: INSOLUBLE (algorithm succeeds), ABORT(algorithm fails)


For each $1 \leq i \leq n$ {
    If $f_i(s) = 0$ for each $s \in \{0,1\}^{k+1}$
        return INSOLUBLE;
}
return ABORT;
```

Table 3.2: Algorithm A1, a linear time algorithm for the uniform probability model with fixed p and k.

Proof: From Theorem 3.2.1, we have

$$\begin{aligned}\lim_{n\to\infty} Pr\{\text{Algorithm A1 succeeds}\} &= 1 - \lim_{n\to\infty} Pr\{\overline{N}(n,k,p) \text{ is soluble}\} \\ &\geq 1 - \lim_{n\to\infty} (1 - p(n)^{2^{k+1}})^n.\end{aligned}$$

It is obvious that the right hand side of the above formula tends to 1 if $\lim_{n\to\infty} p(n) = p_0 > 0$. For the case $\lim_{n\to\infty} p(n) = 0$, write

$$\lim_{n\to\infty} \left(1 - p(n)^{2^{k+1}}\right)^n = \lim_{n\to\infty} \left(\left(1 - p(n)^{2^{k+1}}\right)^{\frac{1}{p(n)^{2^{k+1}}}}\right)^{np(n)^{2^{k+1}}},$$

and we have

$$\lim_{n\to\infty} \left(1 - p(n)^{2^{k+1}}\right)^{\frac{1}{p(n)^{2^{k+1}}}} = \frac{1}{e} < 1.$$

It follows that

$$\lim_{n\to\infty} \left(1 - p(n)^{2^{k+1}}\right)^n = 0$$

since k is fixed and

$$\lim_{n} np(n)^{2^{k+1}} = \lim_{n} (p(n)n^{\frac{1}{2^{k+1}}})^{2^{k+1}} = +\infty.$$

This proves the Corollary. □

Now, let us look at the situation where $\lim_{n} p(n)n^{\frac{1}{2^{k+1}}} \in [0, +\infty)$. We first introduce the concept of a connection graph for the NK landscape model and some results in the theory of random graphs.

Definition 3.2.1. The *connection graph* of an NK landscape instance $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$ is a graph $G = G(V, E)$ satisfying

(1) Each vertex $v \in V$ corresponds to a local fitness function; and

(2)There is an edge between v_i, v_j if and only if the corresponding local fitness functions f_i, f_j share variables and both of them have at least one zero value.

Definition 3.2.2. Let $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$ be an NK landscape instance with the connection graph $G = G(V, E)$. Let $G_1, \cdots, G_l$ be the connected components of G. Since the vertices of G correspond to local fitness functions, we can regard G_i as a set of local fitness functions. For each $1 \leq i \leq l$, define $U_i \subset x = (x_1, \cdots, x_n)$ as

$$U_i = \{y \in x : y \text{ appears in the definition of some local fitness functions in } G_i\}.$$

It's easy to see that $(U_1, \cdots U_l)$ forms a disjoint partition of the variables $x = (x_1, \cdots, x_n)$, and that the local fitness functions in G_i only depend on the variables in U_i. In fact, we have the following:

Theorem 3.2.3. *Let $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$ be an NK landscape instance with the connection graph $G = G(V, E)$. Let $G_1, \cdots, G_l$ be the connected components of G and $(U_1, \cdots U_l)$ be the associated disjoint partition of the variables $x = (x_1, \cdots, x_n)$. Then, the NK decision problem is soluble if and only if for each $1 \leq i \leq l$, there is an assignment $s_i \in \{0, 1\}^{|U_i|}$ to the variables in U_i such that for each local fitness function $g \in G_i$, $g(s) = 1$.*

Proof: The "only if" part of the theorem is obvious. To prove the "if" part, assume that for each $1 \leq i \leq l$, there is an assignment s_i to the variables in U_i such that $g(s_i) = 1$ for any local fitness function $g \in G_i$. Since $U_i, 1 \leq i \leq l$ is a disjoint partition of the variables $x = (x_1, \cdots, x_n)$ and the local fitness functions in G_i only depend on the variables in U_i, by combining $s_i, 1 \leq i \leq l$, we get an assignment to $x = (x_1, \cdots, x_n)$ which is a solution to the NK decision problem. □

Based on this result, we can design an algorithm A2 as shown in Table 3.3 which splits the NK decision problem into several sub-problems according to the connected components of the connection graph, and solves these sub-problems using brute-force search.

```
Algorithm A2
Input: An instance of $\overline{N}(n, k, p), f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$;
Output: SOLUBLE/INSOLUBLE(algorithm always succeeds)

1. Find the connected components $G_i, 1 \leq i \leq l$ of $G$;
2. For each $1 \leq i \leq l$ {
       If there is no $s \in \{0, 1\}^{|U_i|}$ such that $f(s) = 1$ for all $f \in G_i$
          return INSOLUBLE;
   }
   return SOLUBLE;
```

Table 3.3: Algorithm A2, an algorithm for the uniform probability model that splits the problem into sub-problems and solve the sub-problems by brute force.

Corollary 3.2.4. *Algorithm A2 is correct and its time complexity is* $O(n^2 + n * 2^M)$ *with* $M = \max(|U_i|, 1 \leq i \leq l)$ *being the maximum size of the subsets* $(U_i, 1 \leq i \leq l)$ *associated with the connected components of the connection graph.*

Proof. The correctness of Algorithm A2 follows directly from Theorem 3.2.3. It takes $O(n^2)$ time to find the connected components of a graph. Since the local fitness functions in G_i only depend on the set of variables U_i, a brute force search will finish in $O(2^{|U_i|})$ for the connected component G_i. The result then follows from the fact that there are at most n connected components in a graph of size n. □

It can be seen that if the maximum size of the connected components of the connection graph is $O(\log n)$, then Algorithm A2 is polynomial. In the following, we will show that this is true for the connection graph of the NK decision problem in the uniform probability model when $\lim_{n} p(n)n^{\frac{1}{2^k+1}} \in [0, +\infty)$.

We first give a lemma on the size of the connected components of a random graph which is well-known in the theory of random graphs (See [KAR95]). Since it plays an important role in the rest of discussion on the solvability of the uniform probability model of NK landscapes, we give it a proof for completeness.

Lemma 3.2.5. *Let $G(n, p_e)$ be a random graph with the edge probability p_e, and m be a positive integer. If $\lim_{n\to\infty} p_e n^{\frac{m}{m-1}} = 0$, then with probability asymptotic to 1 as n tends to infinity, every connected component of $G(n, p_e)$ has the size less than m.*

Proof: Let Y_m be the number of trees with size m which are contained in $G(n, p_e)$. Since there are m^{m-2} different trees on m vertices and a tree on m vertices has $m - 1$ edges, we have

$$E[Y_m] = \binom{n}{m} m^{m-2} p_e^{m-1}$$

If $\lim_{n\to\infty} p_e n^{\frac{m}{m-1}} = 0$ and m is fixed, we have $\lim_{n\to\infty} p_e^{m-1} n^m = 0$. Since $\binom{n}{m} \le n^m$, it is easy to see that $\lim_{n\to\infty} E[Y_m] = 0$. The lemma follows from Markov's inequality. □

By combining Algorithms A1 and A2, we obtain the polynomial algorithm A in Table 3.4. We now prove that Algorithm A is correct with probability asymptotic to 1 as n tends to infinity.

Theorem 3.2.6. *For any $p(n)$ such that $\lim_{n\to\infty} p(n)n^{\frac{1}{2^k+1}}$ exists, Algorithm A is polynomial and successfully solves a random instance of $\overline{N}(n, k, p)$ with probability asymptotic to 1 as n tends to infinity.*

```
Algorithm A
Input: An instance of $\overline{N}(n, k, p), f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$;
Output: SOLUBLE/INSOLUBLE(algorithm successes),
        ABORT(algorithm fails)

1. Run Algorithm A1 on $f(x)$;
   If A1 returns INSOLUBLE
      Return INSOLUBLE;

2. Find the connected components of the connection graph of $f$;
   If the maximum size of the components is greater than $2^k + 2$
      Return ABORT;

3. Run step 2 of Algorithm A2 on $f(x)$ and return the results accordingly.
```

Table 3.4: Algorithm A, a combination of Algorithms A1 and A2.

Proof: From Corollary 3.2.2 and Corollary 3.2.4, we know that the time complexity of Algorithm A is $O(n) + O(n^2) + O(n * 2^{2^k+2})$. If $\lim_{n\to\infty} p(n) n^{\frac{1}{2^k+1}} = +\infty$, we know from Corollary 3.2.2 that algorithm A1 will return INSOLUBLE and hence succeeds with probability asymptotic to 1. It follows that

$$\begin{aligned} &\lim_{n\to\infty} Pr\{\text{Algorithm A succeeds}\} \\ &\geq \lim_{n\to\infty} Pr\{\text{Algorithm A1 succeeds}\} \\ &= 1. \end{aligned} \tag{3.2.1}$$

Consider the situation where $\lim_n p(n) n^{\frac{1}{2^k+1}} = c$ is a finite positive constant. Let $\mathcal{M}(n, k, p)$ be the maximum size of the connected components of the connection graph of $\overline{N}(n,k,p)$. Since Algorithm A2 always succeeds, we have

$$\lim_{n\to\infty} Pr\{\text{Algorithm A succeeds}\} \geq \lim_{n\to\infty} Pr\{\mathcal{M}(n,k,p) \leq 2^k + 2\}.$$

We complete the proof by showing that

$$\lim_{n\to\infty} Pr\{\mathcal{M}(n,k) \leq 2^k + 2\} = 1.$$

Consider the connection graph $G = G(V,E)$ of the random NK landscape $\overline{N}(n,k,p)$. G is a random graph and there is an edge between two nodes of G if and only if the two corresponding local fitness functions share a variable and both of the local fitness functions take at least one zero as their fitness value. Therefore, the edge probability of G is

$$p_e(n) \leq \left(1 - \frac{C_{n-2}^k C_{n-2-k}^k}{C_{n-1}^k C_{n-1}^k}\right) * \left(1 - (1-p(n))^{2^{k+1}}\right)^2.$$

Since

$$\begin{aligned}
\frac{C_{n-2}^k C_{n-2-k}^k}{C_{n-1}^k C_{n-1}^k} &= \frac{C_{n-2}^k}{C_{n-1}^k}\frac{C_{n-2-k}^k}{C_{n-1}^k} \\
&= \frac{n-k-1}{n-1}\frac{(n-k-2)(n-k-3)\cdots(n-2k-1)}{(n-1)(n-2)\cdots(n-k)} \\
&\leq \frac{n-k-1}{n-1}\left(\frac{n-k-2}{n-1}\right)^k
\end{aligned}$$

and k is a fixed constant, we have

$$\begin{aligned}
1 - \frac{C_{n-2}^k C_{n-2-k}^k}{C_{n-1}^k C_{n-1}^k} &\leq 1 - \frac{n-k-1}{n-1}\left(\frac{n-k-2}{n-1}\right)^k \\
&= O\left(\frac{1}{n}\right),
\end{aligned}$$

It follows that

$$p_e(n) = O\left(\frac{1}{n}p(n)^2\right).$$

Since $\lim_{n\to\infty} p(n)n^{\frac{1}{2^k+1}} < +\infty$, we have

$$\lim_{n\to\infty} p_e * n^{\frac{2^k+2}{2^k+1}} = 0.$$

It follows from Lemma 3.2.5 that the maximum size of the connected components of the random graph $G = G(n, p_e(n))$ is less than $2^k + 2$ with probability asymptotic to 1. This completes the proof. □

3.3 Thresholds of the Fixed Ratio Model

As has been discussed in the previous section, the uniform probability model $\overline{N}(n, k, p)$ of NK landscapes is asymptotically trivial. This is largely due to the fact that if the parameter p does not decrease very quickly with n, then asymptotically there will be at least one local fitness function that takes the value 0 for all the possible assignments, making the whole decision problem insoluble. In this section, we study the fixed ratio model $N(n, k, z)$. In this model, we require that each local fitness function has fixed number of zero values so that the trivially insoluble situation in the uniform probability model is avoided. We noticed that the same idea has been used in the study of the *flawless CSP* [GMP98].

We will establish several upper bounds on the solvability threshold of the parameter z, and theoretically prove that random instances generated with the parameter z above these upper bounds can be solved with probability asymptotic to 1 by polynomial(even linear) algorithms.

Recall that in the fixed ratio model, we choose the neighborhood structure for each local fitness in the same way as in the uniform probability model $\overline{N}(n, k, p)$. To determine the fitness value for a local fitness function f_i, we randomly without replacement select exactly z tuples $\{s_1, \cdots, s_z\}$ from $\{0, 1\}^{k+1}$, and let $f_i(s_j) = 0$ for each $1 \leq j \leq z$ and $f_i(s) = 1$ for every other $s \in \{0, 1\}^{k+1}$.

It's easy to see that the property "There exists an assignment x such that $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)) = n$" is monotone in the parameter z — the number of tuples at which a local fitness function takes zero. It also can be shown that for two random instances $N(n,k,z_1)$ and $N(n,k,z_2)$ with $z_1 > z_2$, we have

$$Pr\{N(n,k,z_1)\text{is soluble}\} \leq Pr\{N(n,k,z_2)\text{is soluble}\}.$$

According to the threshold conjecture, there exists a z_c such that

$$\lim_{n\to\infty} Pr\{N(n,k,z)\text{is soluble}\} = \begin{cases} 1, & \text{if } z < z_c; \\ 0, & \text{if } z > z_c. \end{cases}$$

The value z_c is called the threshold (or critical value) of the corresponding probability model. An important task in the study of the phase transitions is to locate the threshold for various random combinatorial problems. See [ACH99a, FSS96, KKK98] for further references. Throughout this section, we assume $k = 2$ in the fixed ratio model $N(n,k,z)$.

3.3.1 A Linear Algorithm for the Case $z > 3.0$

Consider a random instance $f = \sum_{i=1}^{n} f_i$ of the fixed ratio model $N(n,2,z)$ with $z = 3.0 + \varepsilon > 3.0$. Without loss of generality, we may write f as

$$f = \sum_{i=1}^{n} f_i = \sum_{i=1}^{\varepsilon n} f_i + \sum_{\varepsilon n+1}^{n} f_i,$$

where f_i has 4 zeroes in its fitness value assignment for $1 \leq i \leq \varepsilon n$ and 3 zeroes for $\varepsilon n + 1 \leq i \leq n$.

Definition 3.3.1. Two local fitness functions f_i and f_j conflict with each other if (1) f_i and f_j have exactly one common variable x and (2)for any assignment $s \in \{0,1\}^n$, we have $f_i(s) * f_j(s) = 0$.

Lemma 3.3.1. *(1) An instance of NK decision problem is insoluble if there exists a pair of conflicting local fitness functions.*

(2) Two local fitness functions conflict with each other only if each of them have at least 4 zeros as their fitness value.

(3) If two local fitness functions conflict with each other and x is the only shared variable, then to make both of the local fitness functions take fitness value 1, x has to be forced to take the value 1 by one local fitness function and 0 by the other local fitness function.

Proof: It follows directly from the definition. □

Theorem 3.3.2. *For the fixed ratio model $N(n, 2, z)$ with $z = 3.0 + \varepsilon$, we have*

$$\lim_{n\to\infty} Pr\{\textit{there is a conflicting pair of local fitness functions in } N(n, 2, z)\} = 1.$$

And thus, a random instance of $N(n, 2, z)$ is insoluble with the probability asymptotic to 1.

Proof: Let

$$f = \sum_{i=1}^{n} f_i = \sum_{i=1}^{\varepsilon n} f_i + \sum_{i=\varepsilon n+1}^{n} f_i,$$

be a random instance of the fixed ratio model, where f_i has 4 zeroes in its fitness value assignment for $1 \leq i \leq \varepsilon n$, and 3 zeroes for $\varepsilon n + 1 \leq i \leq n$. Let I_{ij} be the indicator function of the event that f_i and f_j conflicts with each other, i.e.,

$$I_{ij} = \begin{cases} 1, & \text{if } f_i \text{ and } f_j \text{ conflicts with each other;} \\ 0, & \text{else.} \end{cases}$$

and $S = \sum_{1\leq i,j\leq \varepsilon n} I_{ij}$. We claim that $\lim_{n\to\infty} Pr\{S = 0\} = 0$.

By Chebyschev's inequality, we have

$$\begin{aligned} Pr\{S = 0\} &\leq Pr\{|S - E(S)| \geq E(S)\} \\ &\leq \frac{Var(S)}{(E(S)^2)}. \end{aligned} \tag{3.3.1}$$

Algorithm B1

Input: An instance of $N(n, 2, z)$, $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$;

Output: INSOLUBLE(algorithm succeeds),

ABORT(algorithm fails)

```
int $b[n] = \{-1, \cdots, -1\}$; b[i] = -1(or 0, 1) means the variable $x_i$ has not been
                                  forced(has been forced to 0, has been forced to 1)
For each $1 \leq i \leq n\{$
Let $x_{i_1}, x_{i_2}, x_{i_3}$ be the three variables of $f_i$;
for each $1 \leq m \leq 3$ {
    check to see if $x_{i_m}$ is forced to make $f_i = 1$
    if $x_{im}$ is not forced, continue;
    else if $x_{i_m}$ is forced to $a$ and $b[i_m] = -1$ {
        $b[i_m] = a$; continue;
    }
    else if the forced value of $x_{i_m}$ equals to $b[i_m]\{$
        continue;
    }
    else return INSOLUBLE;
}
return ABORT;
```

Table 3.5: Algorithm B, a linear time algorithm for the fixed ratio model with $z > 3$.

Since for each $1 \leq i \leq \varepsilon n$, f_i has exactly 4 zeros in its fitness value assignment, we have that two local fitness function $f_i, f_j, 1 \leq i, j \leq \varepsilon n$, conflict with each other if and only if they have exactly one common variable x such that one of the following is true: $(1) f_i(s) = 0(\text{or } 1), f_j(s) = 1(\text{or } 0)$ for all the assignments s such that $x = 1$(respectively $x = 0$); and $(2) f_i(s) = 1(\text{or } 0), f_j(s) = 0(\text{or } 1)$ for all the assignments s such that $x = 1$(respectively $x = 0$);

Since the probability that two local fitness functions share a variable is equal to

$$1 - \frac{\binom{n-2}{2}\binom{n-4}{2}}{\binom{n-1}{2}\binom{n-1}{2}},$$

we have

$$Pr\{I_{ij} = 1\} = \left(1 - \frac{\binom{n-2}{2}\binom{n-4}{2}}{\binom{n-1}{2}\binom{n-1}{2}}\right) \cdot 2\left(\frac{1}{\binom{8}{4}}\right)^2 = \Omega(\frac{1}{n}), \quad \varepsilon > 0, \quad 1 \leq i, j \leq \varepsilon n,$$

and hence,

$$E(S) = \sum_{1 \leq i,j \leq \varepsilon n} E(I_{ij}) = \sum_{1 \leq i,j \leq \varepsilon n} Pr\{I_{ij} = 1\} \in \Omega(n).$$

We now consider the variance of S. Since $S = \sum_{1 \leq i,j \leq \varepsilon n} I_{ij}$, we have

$$Var(S) = \frac{\sum_{i,j} Var(I_{ij}) + 2 \sum_{(i,j) \neq (l,m)} [E\{I_{ij} I_{lm}\} - E\{I_{ij}\}E\{I_{lm}\}]}{(E(S))^2}.$$

Let

$$A_1 = \frac{\sum_{i,j} Var(I_{ij})}{(E(S))^2}$$

and

$$A_2 = \frac{2 \sum_{(i,j) \neq (l,m)} [E\{I_{ij} I_{lm}\} - E\{I_{ij}\}E\{I_{lm}\}]}{(E(S))^2}.$$

It is easy to see that $\lim_{n \to \infty} A_1 = 0$. To prove $\lim_{n \to \infty} A_2 = 0$, we consider two cases:

Case 1: $i \neq j \neq m \neq l$. In this case, the two random variables I_{ij} and I_{lm} are actually independent. It follows that $E\{I_{ij} I_{lm}\} - E\{I_{ij}\}Pr\{I_{lm}\} = 0$.

Case 2: $(i,j) \neq (l,m)$ but they have one in common, say $j = l$. In this case, we have

$$\begin{aligned} E\{I_{ij}I_{lm}\} - E\{I_{ij}\}E\{I_{lm}\} &= Pr\{I_{ij}=1\}Pr\{I_{jm}=1|I_{ij}=1\} - \Omega\left(\left(\frac{1}{n}\right)^2\right) \\ &= \Omega\left(\frac{1}{n}\right)Pr\{I_{jm}=1|I_{ij}=1\} - \Omega\left(\left(\frac{1}{n}\right)^2\right) \end{aligned}$$

Given that f_i and f_j conflict with each other, the conditional probability that f_j and f_m conflict with each other is still in $\Omega(\frac{1}{n})$.

Since there are only $C_{\epsilon n}^3$ pairs of I_{ij} and I_{jm} satisfying the condition in Case 2, we know that $\sum_{(i,j)\neq(l,m)} [E\{I_{ij}I_{lm}\} - E\{I_{ij}\}E\{I_{lm}\}]$ is in $\Omega(n)$. And therefore, $\lim_{n\to\infty} A_2 = 0$. It follows that

$$\lim_{n\to\infty} Pr\{S=0\} \leq \lim_{n\to\infty} \frac{Var(S)}{(E(S)^2)} = 0.$$

Since the event $\{S > 0\}$ implies that there exists a conflicting pair of local fitness functions, the theorem follows. □

Based on Lemma 3.3.1 and Theorem 3.3.2, we have a linear algorithm, algorithm B1 in Table 3.5, which solves the NK decision problem $N(n,2,z)$ with probability asymptotic to 1 for any $z > 3$.

3.3.2 2-SAT Sub-problems in $N(n,2,z)$ and a Tighter Upper Bound

In this subsection, we establish a tighter upper bound $z > 2.837$ for the threshold of the fixed ratio model $N(n,2,z)$ by showing that asymptotically, $N(n,2,z)$ contains an unsatisfiable 2-SAT sub-problem with probability 1 for any value of z greater than 2.873. This enables us to have a polynomial algorithm which determines that $N(n,2,z)$ is insoluble with probability asymptotic to 1 for $z > 2.837$.

Definition 3.3.2. [FRG98] (1) Let $U = \{u_i, 0 \leq i \leq 3p\}$ be a set of Boolean variables. A *criss-cross loop* $\mathcal{L}(U, t), t = 3p + 2$, is a set of 2-clauses

$$\bar{u}_0 \vee u_1,\ \bar{u}_1 \vee u_2, \cdots, \bar{u}_{p-1} \vee u_p,\ \bar{u}_p \vee \bar{u}_0,$$

$$u_0 \vee u_{p+1},\ \bar{u}_{p+1} \vee u_{p+2}, \cdots, \bar{u}_{3p-1} \vee u_{3p},\ \bar{u}_{3p} \vee u_0.$$

(2) Let x, y, z be Boolean variables. A 3-module M for the clause $x \vee y$ is a pair of 3-clauses $M = \{x \vee y \vee z, x \vee y \vee \bar{z}\}$.

It is easy to see that a criss-cross loop contains two contradictory cycles and is unsatisfiable. This special structured 2-SAT formula plays an important role in analyzing the phase transition of 2-SAT [BBC99, FRG98].

Definition 3.3.3. A t-3-module $\mathcal{M}$ associated with a criss-cross loop $\mathcal{L}(U, t)$ is the union of the 3-modules for all the 2-clauses in $\mathcal{L}(U, t)$:

$$
\begin{array}{ll}
\mathcal{M} = \{ & M_1 = (\bar{u}_1 \vee u_2 \vee z_1, \bar{u}_1 \vee u_2 \vee \bar{z}_1); \\
& \quad \ldots \\
& M_{p-1} = (\bar{u}_{p-1} \vee u_p \vee z_{p-1}, \bar{u}_{p-1} \vee u_p \vee \bar{z}_{p-1}); \\
& M_p = (\bar{u}_p \vee \bar{u}_0 \vee z_p, \bar{u}_p \vee \bar{u}_0 \vee \bar{z}_p); \\
& M_{p+1} = (\bar{u}_{p+1} \vee u_{p+2} \vee z_{p+1}, \bar{u}_{p+1} \vee u_{p+2} \vee \bar{z}_{p+1}); \\
& \quad \ldots \\
& M_{3p-1} = (\bar{u}_{3p-1} \vee u_{3p} \vee z_{3p-1}, \bar{u}_{3p-1} \vee u_{3p} \vee \bar{z}_{3p-1}); \\
& M_{3p} = (\bar{u}_{3p} \vee u_0 \vee z_{3p}, \bar{u}_{3p} \vee u_0 \vee \bar{z}_{3p}) \\
& M_{3p+1} = (\bar{u}_0 \vee u_1 \vee z_{3p+1}, \bar{u}_0 \vee u_1 \vee \bar{z}_{3p+1}); \\
& M_{3p+2} = (u_0 \vee v_{p+1} \vee z_{3p+2}, u_0 \vee u_{p+1} \vee \bar{z}_{3p+2}); \\
\} &
\end{array}
$$

such that all of the literals of u_i's and z_i's are variable-distinct.

It is easy to see that there are $T = 2t = 2(3p+2)$ clauses and $T-1$ distinct variables in the t-3-module. A t-3-module is *minimally* unsatisfiable, that is, the removal of any clause from it produces a satisfiable set of clauses.

Definition 3.3.4. Given a t-3-module $\mathcal{M}$ and an NK landscape instance $f = \sum_{i=1}^{n} f_i, k = 2$, a sequence of local fitness functions

$$\mathbf{g} = (g_1, \cdots, g_t) \subset (f_1, \cdots, f_n)$$

is said to be a *possible match*(PM) if for each $1 \le m \le t$, the main variable of g_m is one of the three variables that occur in the 3-module M_m. A subsequence $(h_1, \cdots, h_l)$ of a possible match $\mathbf{g}$ is legal if for any $1 \le m < j \le l, \quad h_m \ne h_j$.

Lemma 3.3.3. *Let $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i))$ be an instance of $N(n, 2, z)$ and $\mathcal{M}$ be a t-3-module. Then the number of possible matches for the t-3-module $\mathcal{M}$ is 3^t. Further, the number of legal possible matches is $\Theta\left((\frac{3+\sqrt{5}}{2})^t\right)$.*

Proof: For each $1 \le m \le t$, there are exactly 3 possible choices for g_m: $f_{i_1}(x_{i_1}, \Pi(x_{i_1})), f_{i_2}(x_{i_2}, \Pi(x_{i_2})), f_{i_3}(x_{i_3}, \Pi(x_{i_3}))$, where $x_{i_1}, x_{i_2}, and \quad x_{i_3}$ correspond to the three variables that occur in the 3-module M_m. Therefore, there are 3^t possible matches for the t-3-module.

To prove the second conclusion, we divide the t-3-module into 3 parts $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$, where $\mathcal{M}_1 = (M_m, 1 \le m \le p)$, $\mathcal{M}_2 = (M_m, p+1 \le m \le 3p-1)$, and $\mathcal{M}_3 = (M_{3p}, M_{3p+1}, M_{3p+2})$. Letting L_1, L_2, L_3 are the number of legal possible matches for $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ respectively. Since the literals in $\mathcal{M}_1$ are variable-distinct from the literals in $\mathcal{M}_2$, we have that the number of legal possible matches, L, for the t-3-module $\mathbf{M}$ satisfies

$$L_1 L_2 \le L \le 27 L_1 L_2.$$

We now estimate the order of L_1. To this end, we consider the probability space (Ω, P), where Ω is the set of sequences $(g_1, \cdots, g_p)$ of local fitness functions that possibly match $\mathcal{M}_1$ and P is the uniform probability distribution. Then, the number of legal possible matches is

$$L_1 = |\Omega| \cdot Pr\{\text{a random sample from } \Omega \text{ is legal}\} \tag{3.3.2}$$

Let $\mathbf{g} = (g_1, \cdots, g_p)$ be a random sample from Ω and x_{g_m} denote the main variable of the local fitness function g_m, then we have

$$Pr\{x_{g_m} = |u_m|\} = Pr\{x_{g_m} = |u_{m+1}|\} = Pr\{x_{g_m} = |z_m|\} = \frac{1}{3},$$

where $|u|$ denotes the variable corresponding to the literal u.

Let $B_m, 0 < m \leq p$ be the event that the first m local fitness functions $g_1, \cdots, g_m$ in the possible match $\mathbf{g} = (g_1, \cdots, g_p)$ are mutually distinct. Since in $\mathcal{M}_1$ only consecutive 3-modules share variables, we have

$$B_m = \{(g_1, \cdots, g_m) : \ g_i \neq g_{i+1}, \ 1 \leq i \leq p-1\}.$$

Let $b_m = Pr\{g_m \neq g_{m-1} \mid B_{m-1}\}, m \geq 2$, and $b_1 = 1$. Notice that $B_1 = \Omega$. Then, we have

$$\begin{aligned}
&Pr\{\mathbf{g} = (g_1, \cdots, g_p) \text{is legal}\} = Pr\{B_p\} \\
&= Pr\{g_1 \neq g_2, g_2 \neq g_3, \cdots, g_{t-1} \neq g_t\} \\
&= Pr\{B_1\}Pr\{g_2 \neq g_1 \mid B_1\} \cdot Pr\{g_3 \neq g_2 \mid B_2\} \cdots Pr\{g_p \neq g_{p-1} \mid B_{p-1}\} \\
&= b_1 b_2 \cdots b_p
\end{aligned} \tag{3.3.3}$$

Recalling that x_{g_m} denotes the main variable of the local fitness function g_m, we have

$$\begin{aligned}
b_p &= Pr\{g_{p-1} \neq g_p, x_{g_{p-1}} = |u_p| \mid B_{p-1}\} + Pr\{g_{p-1} \neq g_p, x_{g_{p-1}} \neq |u_p| \mid B_{t-1}\} \\
&= Pr\{g_{p-1} \neq g_p \mid B_{t-1}, x_{g_{p-1}} = |u_p|\} \cdot Pr\{x_{g_{p-1}} = |u_p| \mid B_{p-1}\} + \\
&\qquad Pr\{g_{p-1} \neq g_p \mid B_{p-1}, x_{g_{t-1}} \neq |u_p|\} \cdot Pr\{x_{g_{t-1}} \neq |u_p| \mid B_{t-1}\} \\
&= \frac{2}{3}a_p + (1 - a_p) \\
&= 1 - \frac{1}{3}a_p,
\end{aligned} \tag{3.3.4}$$

where $a_p = Pr\{x_{g_{p-1}} = |u_p| \mid B_{p-1}\}$. For a_p, we have

$$
\begin{aligned}
a_p &= \frac{Pr\{B_{p-1}, x_{g_{p-1}} = |u_p|\}}{Pr\{B_{p-1}\}} \\
&= \frac{1}{Pr\{B_{p-1}\}}(Pr\{B_{p-1}, x_{g_{p-1}} = |u_p|, x_{g_{p-2}} = |u_{p-1}|\} \\
&\qquad + Pr\{B_{p-1}, x_{g_{p-1}} = |u_p|, x_{g_{p-2}} \neq |u_{p-1}|\}) \\
&= \frac{1}{Pr\{B_{p-1}\}}(Pr\{x_{g_{p-1}} = |u_p| \mid B_{p-1}, x_{g_{p-2}} = |u_{p-1}|\} \cdot Pr\{B_{p-1}, x_{g_{p-2}} = |u_{p-1}|\} \\
&\qquad + Pr\{x_{g_{p-1}} = |u_p| \mid B_{p-1}, x_{g_{p-2}} \neq |u_{p-1}|\} \cdot Pr\{B_{p-1}, x_{g_{p-2}} \neq |u_{p-1}|\}) \\
&= \frac{1}{Pr\{B_{p-1}\}}\left(\frac{1}{2}Pr\{B_{p-1}, x_{g_{p-2}} = |u_{p-1}|\} + \frac{1}{3}Pr\{B_{p-1}, x_{g_{p-2}} \neq |u_{p-1}|\}\right)
\end{aligned}
\tag{3.3.5}
$$

The last equation in the above formula is because that given B_{p-1} and $x_{g_{p-2}} = |u_{p-1}|$ (or $x_{g_{p-2}} \neq |u_{p-1}|$), we have two (three, respectively) choices in selecting the local fitness function g_{p-1}. Consider the two terms $Pr\{B_{p-1}, x_{g_{p-2}} = |u_{p-1}|\}$ and $Pr\{B_{p-1}, x_{g_{p-2}} \neq |u_{p-1}|\}$ in (3.3.5), we have

$$
\begin{aligned}
&Pr\{B_{p-1}, x_{g_{p-2}} = |u_{p-1}|\} \\
&= Pr\{g_{p-2} \neq g_{p-1} \mid B_{p-2}, x_{g_{p-2}} = |u_{p-1}|\} \cdot Pr\{B_{p-2}, x_{g_{p-2}} = |u_{p-1}|\} \\
&= \frac{2}{3}\mathrm{Pr}\{x_{g_{p-2}} = |u_{p-1}| \mid B_{p-2}\} \cdot Pr\{B_{p-2}\} \\
&= \frac{2}{3}a_{p-1} \cdot Pr\{B_{p-2}\}
\end{aligned}
\tag{3.3.6}
$$

and

$$
\begin{aligned}
&Pr\{B_{p-1}, x_{g_{p-2}} \neq |u_{p-1}|\} \\
&= Pr\{g_{p-2} \neq g_{p-1} \mid B_{p-2}, x_{g_{p-2}} \neq |u_{p-1}|\} \cdot Pr\{B_{p-2}, x_{g_{p-2}} \neq |u_{p-1}|\} \\
&= Pr\{x_{g_{p-2}} \neq |u_{p-1}| \mid B_{p-2}\} \cdot Pr\{B_{p-2}\} \\
&= (1 - a_{p-1}) \cdot Pr\{B_{p-2}\}
\end{aligned}
\tag{3.3.7}
$$

By plugging (3.3.6) and (3.3.7) into (3.3.5), we get

$$a_p = \frac{Pr\{B_{p-2}\}}{Pr\{B_{p-1}\}}\left(\frac{1}{3}a_{p-1} + \frac{1}{3}(1 - a_{p-1})\right) = \frac{1}{3b_{p-1}}.$$

This, together with (3.3.4), gives us

$$b_p = 1 - \frac{1}{9b_{p-1}}. \tag{3.3.8}$$

It is not difficult to show that the sequence $\{b_p\}$ is decreasing and lower bounded by 0. Letting $\lim_{p} = b$ and taking the limit on both sides, we get

$$b = 1 - \frac{1}{9b}, \tag{3.3.9}$$

and thus, $b = \frac{3\pm\sqrt{5}}{6}$. In our case, $b = \frac{3+\sqrt{5}}{6}$ since $b_1 = 1$. It follows that $b_t \geq b = \frac{3+\sqrt{5}}{6}$. And from (3.3.2), we know that the number of legal possible matches is greater than

$$3^p \left(\frac{3+\sqrt{5}}{6}\right)^p = \left(\frac{3+\sqrt{5}}{2}\right)^p. \tag{3.3.10}$$

To prove that the expected number of legal possible matches L_1 for $\mathcal{M}_1$ is in $\Theta\left(\left(\frac{3+\sqrt{5}}{2}\right)^p\right)$, let $\alpha_p = b_p - \frac{3+\sqrt{5}}{6} = b_p - b$. From (3.3.8) and (3.3.9), we have

$$\alpha_p = b_p - b = \frac{b_{p-1} - b}{9bb_{p-1}} \leq d\alpha_{p-1}, \quad 0 < d < 1.$$

which means that the series $\sum_{m=1}^{p} \alpha_m$ is convergent. It follows that

$$(1 + \frac{\alpha_1}{b}) \cdots (1 + \frac{\alpha_p}{b})$$

converges to a finite positive constant c. Therefore,

$$\begin{aligned} b_1 \cdots b_p &= (b + \alpha_1) \cdots (b + \alpha_p) \\ &= b^p \left(1 + \frac{\alpha_1}{b}\right) \cdots \left(1 + \frac{\alpha_p}{6}\right) \\ &\leq c \left(\frac{3+\sqrt{5}}{6}\right)^p \end{aligned} \tag{3.3.11}$$

for sufficient large p and some constant c.

Similarly, we can show that the number of legal possible matches L_2 for $\mathcal{M}_2$ is in $\Theta\left(\left(\frac{3+\sqrt{5}}{2}\right)^p\right)$. Recalling that the number of legal possible matches L

for the t-3-module satisfies $L_1L_2 \leq L \leq 27L_1L_2$, the second conclusion follows. □

The following Lemma calculates the probability that a matching local fitness function implies the matched 3-module.

Lemma 3.3.4. *Given a 3-module $x \vee y \vee w$, $x \vee y \vee \bar{w}$, and a local fitness function g such that the main variable x_g of g is one of the three Boolean variables $|x|, |y|, |w|$, let $z = 2 + \alpha$, $0 \leq \alpha \leq 1$ be the parameter in the fixed ratio model $\mathcal{N}(n, 2, z)$. Then the probability that g contains the 3-module is*

$$p_0 = \left(\frac{1}{\binom{n-1}{2}}\right)\left(\frac{1}{28}(1-\alpha) + \frac{6}{56}\alpha\right) \tag{3.3.12}$$

Proof: Since x_g is already one of the variables in the 3-module, the probability that the other two variables are also in the 3-module is $\frac{1}{\binom{n-1}{2}}$.

Now, assume that the variables of the local fitness function g are the same as the variables in the 3-module. From the definition of the fixed ratio model, g has two zeros in its fitness value assignment with probability $(1-\alpha)$, and has three zeros in its fitness assignment with probability α. Note that the local fitness function g implies the 3-module $x \vee y \vee w$, $x \vee y \vee \bar{w}$ if and only if

$$g(\bar{x}, \bar{y}, \bar{w}) = 0 \quad \text{and} \quad g(\bar{x}, \bar{y}, w) = 0.$$

From the definition of the fixed ratio model, this happens with the probability

$$\frac{1}{\binom{8}{2}}(1-\alpha) + \frac{6}{\binom{8}{3}}\alpha$$

The Lemma follows. □

With the above preparation, we are ready to prove the upper bound $z > 2.837$ for the threshold of the $\mathcal{N}(n, 2, z)$ model. The idea is to show that if $z > 2.837$, then $\mathcal{N}(z, 2, z)$ contains asymptotically with probability 1, a t-3-module which is an unsatisfiable 2-SAT instance. To make the proof more readable, we present the result in two theorems. In the first theorem, we show that the average number of t-3-modules contained in $\mathcal{N}(n,2,z)$ tends to infinity, while in the second theorem we show $\mathcal{N}(n,2,z)$ contains at least a t-3-module with probability asymptotic to one by using the second moment method.

Theorem 3.3.5. *Let A_t be the number of t-3-modules contained in $\mathcal{N}(n,2,z)$ and $t = \Theta(\ln^2 n)$. Then, if $z = 2+\alpha > 2.837$,*

$$\lim_{n\to\infty} E\{A_t\} = \infty. \tag{3.3.13}$$

Proof. From Lemma 3.3.3, there are more than $(\frac{3+\sqrt{5}}{2})^t$ legal possible matches for a fixed t-3-module. From Lemma 3.3.4, we know that each possible legal match $\mathbf{g} = \{g_1, \cdots, g_t\}$ implies the t-3-module with probability p_0^t. From the proof of Theorem 10.1 in [FRG98], there are

$$2^{t-2} n^{\underline{t-1}} (n-t+1)^t \tag{3.3.14}$$

possible t-3-modules, where $n^{\underline{t-1}} = \frac{n!}{(n-t+1)!}$. Let $r = \left(\frac{1}{28}(1-\alpha) + \frac{6}{56}\alpha\right)$, and write $p_0 = \frac{1}{\binom{n-1}{2}} r$. We have

$$\begin{aligned} E\{A_t\} &= \left(\frac{3+\sqrt{5}}{2} p_0\right)^t \cdot 2^{t-2} n^{\underline{t-1}} (n-t+1)^{\underline{t}} \\ &= \left(\frac{3+\sqrt{5}}{2} r\right)^t \cdot 2^{t-2} n^{\underline{t-1}} (n-t+1)^{\underline{t}} \cdot \frac{1}{\binom{n-1}{2}^t} \\ &= \frac{1}{4(n-t+1)} \left(\frac{3+\sqrt{5}}{2} r\right)^t \cdot \frac{2^t n^{\underline{t}} (n-t+1)^{\underline{t}}}{\binom{n}{2}^t} \left(\frac{\binom{n}{2}}{\binom{n-1}{2}}\right)^t \\ &= \frac{1}{4(n-t+1)} \left(\frac{3+\sqrt{5}}{2} r\right)^t \cdot \frac{4^t n^{\underline{t}} (n-t+1)^{\underline{t}}}{(n(n-1))^t} \left(\frac{n}{n-2}\right)^t \\ &= \frac{1}{4n} (2(3+\sqrt{5})r)^t (1 - O\left(\frac{t^2}{n}\right)), \end{aligned} \tag{3.3.15}$$

where the fourth equation in (3.3.15) is due to the fact that for any positive integer n and q such that $q < \frac{n}{2}$, we have $n^q e^{-q^2/2n} \le n^{\underline{q}} \le n^q$. It follows that

$\lim_{n\to\infty} E\{A_t\} = \infty$ if

$$2(3+\sqrt{5})r > 1. \tag{3.3.16}$$

Solving the inequality (3.3.16) gives us $\alpha > 0.837$, that is, $z = 2+\alpha > 2.837$. This proves Theorem 3.3.5. □

Theorem 3.3.5 shows that the average number of t-3-modules in $N(n, 2, z)$ tends to infinity. Based on the Chebychev's inequality, to prove that $N(n, 2, z)$ contains t-3-modules with probability 1, we need to show that the variance of A_t, the number of contained t-3-modules, is $o(E\{A_t\})$. To do this, we need the following

Lemma 3.3.6. *(Alon and Spencer [ALS92]) Given a random structure(e.g., a random CNF formula), let W be the set of substructures under consideration, $A(w)$ be the set of substructures sharing some clauses with $w \in W$. Let $I_w = 1$ when w is in the random structure and 0 otherwise. If*

(1) elements of W are symmetric;

(2) $\mu = E\{\sum_{w\in W} I_w\} \to \infty$; and

(3) $\sum_{\bar{w}\in A(w)} Pr(\bar{w} \mid w) = o(\mu)$, for each $w \in W$,

then as $n \to \infty$, the probability that the random structure contains a substructure tends to 1.

To use the above Lemma to study the 2-SAT sub-problem in NK landscapes, we view the random structure to be a random instance of $N(n, 2, z)$, and W to be the set of all t-3-modules which is symmetric by their definition.

Theorem 3.3.7. *If $z = 2+\alpha > 2.837$, then $N(n, 2, z)$ is asymptotically insoluble with probability 1.*

Proof: Let A_t be the number of t-3-modules implied by $N(n, 2, z)$ and $t = O(\ln^2 n)$. Theorem 3.3.5 shows that $\lim_{n\to\infty} E\{A_t\} = \infty$. By Lemma 3.3.6, it is enough to show that for each $w \in W$,

$$\sum_{\bar{w} \in A(w)} Pr(\bar{w} \mid w) = o(E\{A_t\}), \tag{3.3.17}$$

where $Pr(\bar{w} \mid w)$ is the conditional probability that $N(n, 2, z)$ implies the t-3-module $\bar{w}$ given that it implies w, and $A(w)$ is the set of all t-3-modules sharing some clauses with w.

Suppose that $\bar{w}$ shares $Q, 1 \leq Q \leq 2t$ clauses with w, and that these Q clauses are distributed among q 3-modules. Further, let q_1 be the number of 3-modules whose two clauses are both shared and $q_2 = q - q_1$ the number of 3-modules that only has one clause shared.

Let T_1 be a 3-module in $\bar{w}$ that shares exactly one clause with a 3-module T_2 in w. We claim that the conditional probability that T_1 is implied by $N(n, 2, z)$ given that w is implied by $N(n, 2, z)$, is

$$\frac{1}{6}\alpha + O(\frac{1}{n}). \tag{3.3.18}$$

Without loss of generality, assume that $T_2 = \{x \vee y \vee u, x \vee y \vee \bar{u}\}$ and $T_1 = \{x \vee y \vee u, \bar{x} \vee y \vee \bar{u}\}$. Since w is implied by $N(n, 2, z)$, there is a local fitness function $g = g(|x|, |y|, |u|)$ that implies T_2. The conditional probability that T_1 is implied, is less than or equal to $P_1 + P_2$ where P_1 is the conditional probability that g also implies the clause $\bar{x} \vee y \vee \bar{u}$ given that g implies T_2, and P_2 is the conditional probability that the clause $\bar{x} \vee y \vee \bar{u}$ is implied by other local fitness functions. By the definition of $N(n, 2, z)$, we have that $P_1 = \frac{1}{6}\alpha$. Since a local fitness function implies $\bar{x} \vee y \vee \bar{u}$ only if it has the same variables with $g = g(|x|, |y|, |u|)$, we have that $P_2 = O(\frac{1}{n})$. The claim is proved. It follows that, for sufficient large n,

$$Pr\{\bar{w} \mid w\} \leq c \left(\frac{3 + \sqrt{5}}{2} p_0\right)^{t-q} \cdot 1^{q_1} \cdot \left(\frac{1}{6}\alpha\right)^{q_2} \tag{3.3.19}$$

where p_0 is defined in Lemma 3.3.4 and c is a fixed constant.

Let $A_{Q,q,q_2}(w)$ be the set of t-3-modules that share Q clauses with w such that these Q clauses are distributed over q different 3-modules. As before, q_1 is the number of 3-modules whose two clauses are both shared and $q_2 = q - q_1$ the number of 3-modules that only has one clause shared. We claim that

$$|A_{Q,q,q_2}(w)| = |A_{2q,q,0}(w)| 6^{q_2}. \tag{3.3.20}$$

where $A_{2q,q,0}(w)$ is the set of t-3-modules that share all the $2q$ clauses in the q 3-modules with w. Let $\overline{M} = \{\overline{M_1}, \cdots, \overline{M_t}\}$ be a t-3-module in which all the clauses $\overline{M_i}, 1 \leq i \leq q$ are shared with w. Let $M = \{M_1, \cdots, M_t\}$ be a t-3-module in which all the clauses in $M_i, 1 \leq i \leq q_1$ are shared and each of the 3-modules $M_i, q_1 + 1 \leq q_1 + q_2$ has only one clause shared. Since for each of the q_2 3-modules, we have 6 ways to choose the non-shared clauses, there are 6^{q_2} such t-3-modules M in $A_{Q,q,q_2}(w)$ that correspond to one t-3-module $\overline{M}$ in $A_{2q,q,0}$. The claims follow. From formula (55) and (56) in [FRG98] and (3.3.17), it follows that

$$|A_{Q,q,q_2}(w)| < \begin{cases} \frac{O(t)}{n^2} 2^{t-q} n^{2(t-q)} 6^{q_2} & , q \leq p+1, \\ \frac{O(1)}{n} 2^{t-q} n^{2(t-q)} 6^{q_2} & ; q > p+1. \end{cases} \tag{3.3.21}$$

Then, we have

$$\begin{aligned} & |A_{Q,q,q_2}(w)| Pr\{\bar{w} \mid w\} \\ & \leq \frac{O(t)}{n^2} 2^{t-q} n^{2(t-q)} 6^{q_2} (\frac{3+\sqrt{5}}{2} p_0)^{t-q} (\frac{1}{6}\alpha)^{q_2} \\ & \leq \frac{O(t)}{n^2} 2^{t-q} n^{2(t-q)} r^{t-q} \frac{1}{\binom{n-1}{2}^{t-q}} \\ & \leq \frac{O(t)}{n} \frac{1}{4n} (4r)^{t-q} \\ & \leq \frac{O(t)}{n} E\{A_t\} (4r)^{-q}, \quad q \leq p+3 \end{aligned} \tag{3.3.22}$$

and

$$|A_{Q,q,q_2}(w)| Pr\{\bar{w} \mid w\} \leq O(1) E\{A_t\} (4r)^{-q}, \quad q > p+3. \tag{3.3.23}$$

Therefore,

$$\sum_{\bar{w}\in A(w)} Pr(\bar{w} \mid w) = \sum_{Q,q,q_2} |A_{Q,q,q_2}(w)| Pr\{\bar{w} \mid w\}$$

$$= \sum_{Q=1}^{t} \sum_{q\le p+3} \sum_{q_2} \frac{O(t)}{n} E\{A_t\}(4r)^{-q} + \sum_{Q-1}^{t} \sum_{q>p+3} \sum_{q_2} O(1) E\{A_t\}(4r)^{-q}. \tag{3.3.24}$$

Since $4r > 1$ for $z > 2.837$, we have

$$\sum_{\bar{w}\in A(w)} Pr(\bar{w} \mid w) \le \frac{O(t^4)}{n} E\{A_t\} + t^3 E\{A_t\}(4r)^{-(p+3)} \tag{3.3.25}$$

$$= o(E\{A_t\}).$$

This completes the proof. □

3.4 The Generalized NK Landscape Model

In the NK landscape model, each local fitness function f_i depends on its main variable x_i and a set of neighboring variables $\Pi(x_i)$, $|\Pi(x_i)| = k$, which are selected randomly. In the generalized NK model([ALT96, WEI96]), each local fitness function f_i is not required to depend on the main variable x_i. Instead, the $k+1$ variables of the local fitness function f_i are all selected randomly. Similar to the definitions of the uniform probability model and fixed ratio model in Section 3.1, we can define the uniform probability model and fixed ratio model for the generalized NK landscape accordingly and write $\overline{N_G}(n,k,p)$ and $N_G(n,k,z)$ for the uniform probability and fixed ratio models respectively.

The results in Section 3.2 and Section 3.3 can be extended to the generalized NK model without much difficulty.

Theorem 3.4.1. *The polynomial Algorithm A given in Section 3.2 successfully solves a random instance of $\overline{N_G}(n,k,p)$ with probability asymptotic to 1 as n tends to infinity for any $p(n)$ such that $\lim_{n\to\infty} p(n) n^{\frac{1}{2k+1}}$ exists.*

Proof: We need to consider two cases: (1) $\lim_{n\to\infty} p(n)n^{\frac{1}{2^{k+1}}} = +\infty$ and (2) $\lim_{n\to\infty} p(n)n^{\frac{1}{2^{k+1}}} \in [0, \infty)$.

For the first case, it is easy to see that Theorem 3.2.1 still holds for $\overline{N_G}(n, k, p)$ since the proof of Theorem 3.2.1 only makes use of the probability distribution with which the fitness values of the local fitness functions are assigned. Then, the proof of Corollary 3.2.2 applies.

For the second case, similar to the proof of Theorem 3.2.6, we need to show that

$$\lim_{n\to\infty} Pr\{M(n, k) \leq 2^k + 2\} = 1,$$

where $M(n, k)$ is the maximum size of the connected components of the connection graph G of $\overline{N_G}(n, k, p)$. To this end, we only need to show that the edge probability $p_e(n)$ of the random graph G satisfies

$$p_e(n) = O\left(\frac{1}{n}p(n)^2\right). \tag{3.4.1}$$

Recalling that in the generalized NK landscape model, the $k + 1$ variables of each local fitness function are all selected randomly, we have by the definition of the connection graph that

$$\begin{aligned}
p_e(n) &\leq \left(1 - \frac{C_n^{k+1} C_{n-k-1}^{k+1}}{C_n^{k+1} C_n^{k+1}}\right) * \left(1 - (1 - p(n))^{2^{k+1}}\right)^2 \\
&= \left(1 - \frac{(n-k-1)(n-k-2)\cdots(n-2k-1)}{n(n-1)\cdots(n-k)}\right) * \left(1 - (1 - p(n))^{2^{k+1}}\right)^2 \\
&\leq \left(1 - \left(\frac{n-k-1}{n}\right)^k\right)\left(1 - (1 - p(n))^{2^{k+1}}\right)^2 \\
&= O\left(\frac{1}{n}p_e(n)^2\right).
\end{aligned} \tag{3.4.2}$$

This completes the proof. □

For the fixed ratio model $N_G(n, 2, z)$ of the generalized NK landscape model, we have the following upper bound on the insoluble threshold:

Theorem 3.4.2. *If $z = 2+\alpha > 2.667$, then $N(n,2,z)$ is in soluble asymptotically with probability 1.*

Proof: Let A_t be the number of t-3-modules contained in $N(n,2,z)$ and $t = O(\ln^2 n)$. For a fixed t-3-models and a given subsequence of local fitness functions $g = \{g_1, \cdots, g_t\} \in \{f_1, \cdots, f_n\}$, the probability that $g = \{g_1, \cdots, g_t\}$ implies the t-3-module is

$$\left(\frac{1}{\binom{n}{3}}\left(\frac{1}{28}(1-\alpha)+\frac{6}{56}\alpha\right)\right)^t, \tag{3.4.3}$$

where $\frac{1}{\binom{n}{3}}$ is the probability that a local fitness function has three specific variables, and $r = \left(\frac{1}{28}(1-\alpha)+\frac{6}{56}\alpha\right)$ is the probability that a local fitness function having the same variables with a 3-module implies the 3-module. Write $p_0 = \frac{1}{\binom{n}{3}}r$. Since there $n^{\underline{t}}$ ways to choose the subsequence $g = \{g_1, \cdots, g_t\}$, and there are $2^{t-2}n^{\underline{t-1}}(n-t+1)^t$ possible t-3-modules ([FRG98]), we have

$$\begin{aligned}
E\{A_t\} &= p_0^t n^{\underline{t}} 2^{t-2} n^{\underline{t-1}} (n-t+1)^{\underline{t}} \\
&= \frac{1}{4(n-t+1)} r^t 2^t n^{\underline{t}} n^{\underline{t}} (n-t+1)^t \cdot \frac{1}{\binom{n}{3}^t} \\
&= \frac{1}{4(n-t+1)} r^t \frac{(2*3!)^t n^{\underline{t}} n^{\underline{t}} (n-t+1)^t}{(n(n-1)(n-2))^t} \\
&= \frac{1}{4n}(12r)^t(1-O\left(\frac{t^2}{n}\right)).
\end{aligned} \tag{3.4.4}$$

It follows that $\lim\limits_{n\to\infty} E\{A_t\} = \infty$ if

$$12r > 1. \tag{3.4.5}$$

Solving the above inequality gives us $\alpha > 0.667$, that is, $z = 2+\alpha > 2.667$. The rest of the proof is to show that the second condition in Lemma 3.3.6 holds and is almost the same as the proof of Theorem 3.3.7. □

3.5 Experiments

Our study of the threshold phenomena in NK landscapes started with experimental investigation. Many of the theoretical results in the previous section are also motivated by the observations made in our experiments. In this section, we describe the approach and methods we used in the experimental study, and report the results and observations we have made.

3.5.1 Strategies and Algorithms

In our experiments, an instance of the decision problem of NK landscape is converted to an equivalent 3-SAT problem, and then the 3-SAT problem is solved using Roberto's relsat—an enhanced version of the famous Davis-Putnam algorithm for SAT problems implemented in C^{++}. The source code of relsat can be found at http://www.cs.ubc.ca/ hoos/SATLIB/solvers.html, and algorithmic details of the implementation is discussed in [BSC97].

Let $f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)), \quad x = (x_1, \cdots, x_n) \in \{0, 1\}^n$ be an instance of the decision problem of NK landscape. From the discussion in Section 3.3.1, each local fitness function f_i is equivalent to a k-SAT problem C_i and the number of clauses in C_i equals to the number of assignments at which f_i takes zero as its fitness value. The NK landscape decision problem is thus equivalent to the following k-SAT problem:

$$\varphi = C_1 \bigwedge C_2 \cdots \bigwedge C_n. \tag{3.5.1}$$

Once we get the equivalent SAT problem φ, Roberto's relsat can be readily used to solve the problem. In the experiments, we generated random instances of the NK landscape decision problem from the random model $N(n, 2, z)$. As a result, the equivalent SAT problem for each random NK landscape instance is a 3-SAT problem with n variables and (on average) zn clauses. By definition,

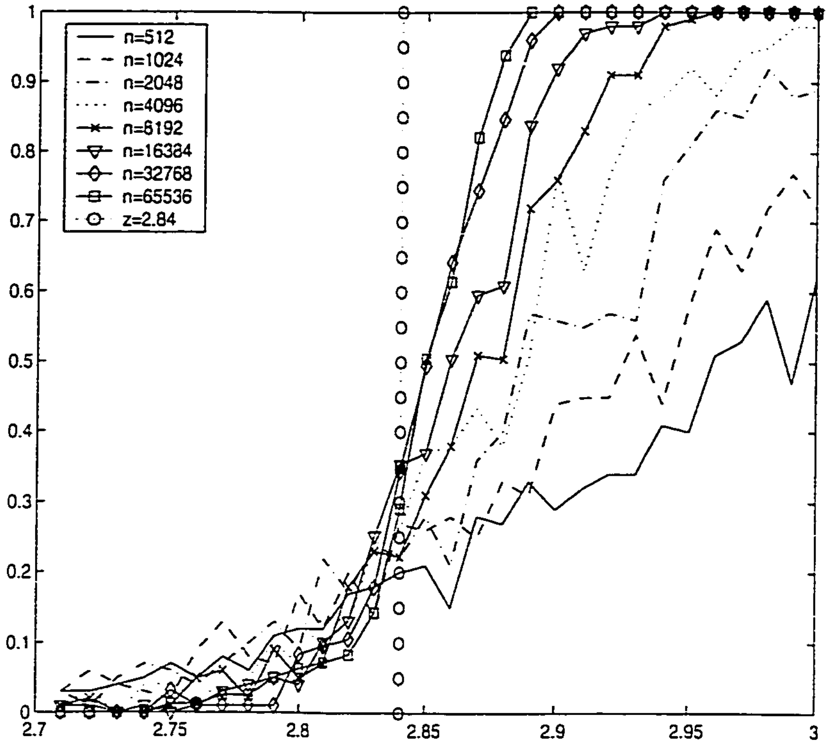

Figure 3.1: Fractions of insoluble instances(Y-axis) as a function of z (X-axis).

the parameter z is between 0 and 8. For $z \leq 1$, the 3-SAT can be solved easily by setting the literals that correspond to the main variables of the local fitness function to true. As z increases, we get more and more clauses and the 3-SAT problem becomes more and more constrained. The aims of the experiments are three-fold:(1)Investigating if there exists a threshold phenomenon in the random NK landscape model; (2) Locating the threshold of the parameter z; and (3)Determining if there are any hard instances around the threshold.

3.5.2 Experiments on the Original Fixed Ratio Model

In this part of the experiments, we generate 100 random instances of $N(n, 2, z)$ for each of the parameters $n = 2^9 \cdots 2^{16}$ and $z = 2.71 + offset, 0 \leq offset \leq$

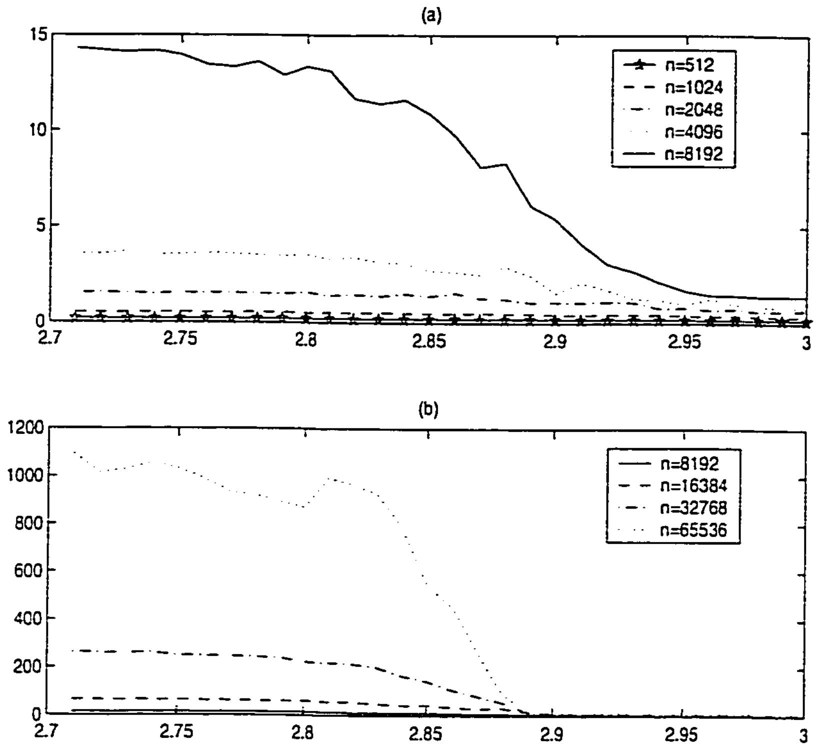

Figure 3.2: Average search cost(Y-axis, in seconds) as a function of z (X-axis).

0.29. These instances are then converted to 3-SAT instances and solved by *relsat*. Figure 3.1 shows the fraction of insoluble instances as a function of the parameter z. It can be seen that there exists a threshold phenomenon and the threshold is around 2.83. This shows that our upper bound $z = 2.837$ is very tight.

In Figure 3.2, we draw the average search cost (in seconds) as a function of the parameter z. As we can see, the average search cost drops quickly at the threshold. By examining the data, we find that it takes much less time to prove the insolubility of an instance. In fact, the data shows that more than 99 percent of the insoluble instances are solved quickly in the preprocessing stage of relsat. This explains the dramatic decreasing of the average search cost because the fraction of insoluble instances increases quickly around the

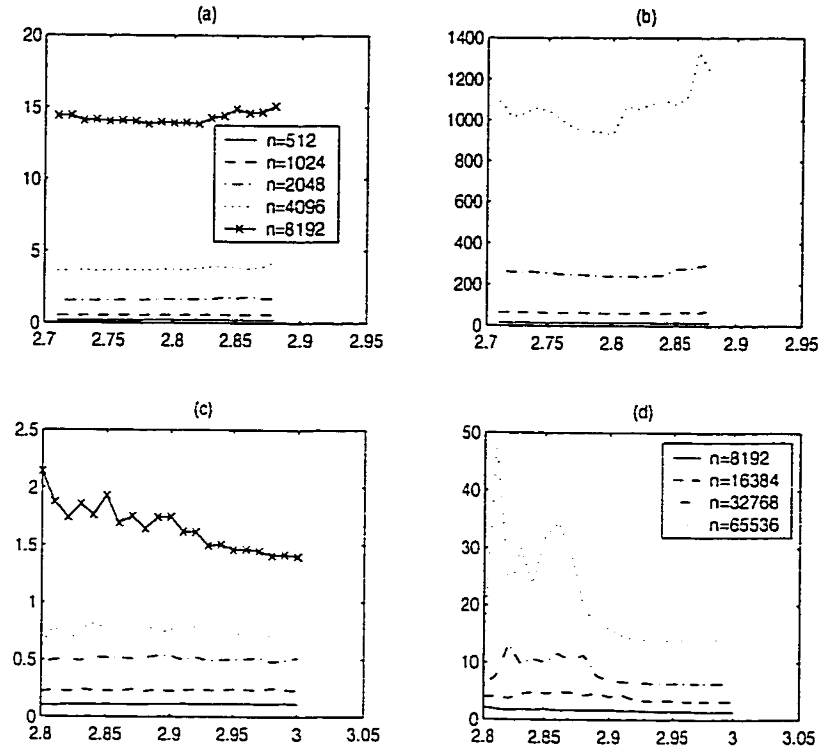

Figure 3.3: Search cost (Y-axis, in seconds) as a function of z (X-axis) for soluble instances (sub-figures (a) and (b)) and insoluble instances(sub-figures (c) and (d)).

threshold. It also suggests that there must be some"small" structures that make the instances insoluble. In Figure 3.3, we separate the search cost into two parts: the average search cost for soluble instances and the average search cost for insoluble instances. As the figure illustrates, it takes much more time to find a solution to the soluble instances than to prove the insoluble instances. Also shown in Figure 3.3 is the fact that in both soluble and insoluble cases, the average search cost is almost constant as a function of the parameter z. To see how the search cost scales with respect to the problem size, we plot in Figure 3.4 the square root of the average search cost as a function of n. The figure suggests that the average search is in $O(n^2)$ for any parameter z.

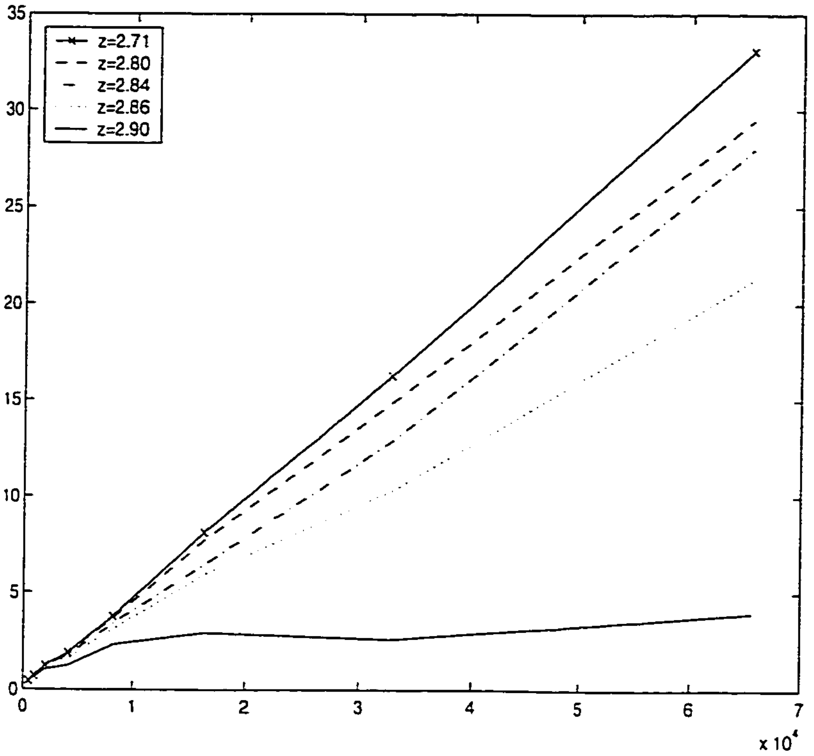

Figure 3.4: Square root of the average search cost (Y-axis, in seconds) as a function of n (X-axis).

3.5.3 Experiments on the 2-SAT sub-Problem

This is the part of the experiments that motivated our theoretical analyses in Section 3.3.2. The idea can be explained as follows. Let

$$f(x) = \sum_{i=1}^{n} f_i(x_i, \Pi(x_i)), \quad x = (x_1, \cdots, x_n) \in \{0,1\}^n$$

be an instance of the decision problem of NK landscape and

$$\varphi = C_1 \bigwedge C_2 \cdots \bigwedge C_n. \tag{3.5.2}$$

the equivalent 3-SAT problem where C_i is the set of 3-clauses equivalent to the local fitness function f_i. For each i, there is a set of 2-clauses D_i(possibly empty) implied by C_i. For example, if C_i has three 3-clauses

$$((x, y, z), (x, \bar{y}, z), (x, y, \bar{z})),$$

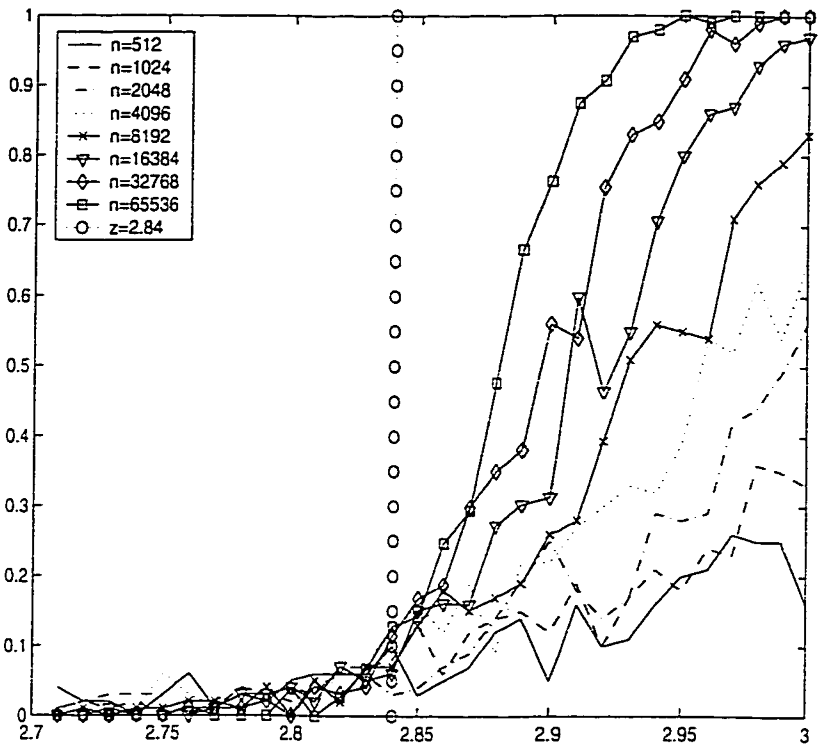

Figure 3.5: Fractions of insoluble instances(Y-axis) as a function of z (X-axis) for 2-SAT sub-problems.

then the set of 2-clauses D_i would be $((x, z), (x, y))$. The conjunction of D_i, denoted by $\bar{\varphi}$, is a 2-SAT problem. It is obvious that the original 3-SAT problem φ is satisfiable only if the 2-SAT sum-problem $\bar{\varphi}$ is satisfiable. In the experiments, we generate instances of the NK landscape $N(n, 2, z)$, convert them to the equivalent 3-SAT problems, and extract the 2-SAT sub-problems. These 2-SAT problems are then solved by the relsat solver. If the 2-SAT problem is unsatisfiable, then the original NK landscape instance is also insoluble.

As in section 3.5.2, we generate 100 instances of $N(n, 2, z)$ at each of the parameters $n = 2^9 - 2^{16}$ and $z = 2.71 + offset, 0 \leq offset \leq 0.29$. The results are shown in Figures 3.5-3.8, in parallel to the Figures 3.1-3.4 of the results on the original 3-SAT problems in section 3.5.2. We see that the patterns of insoluble fractions and search cost are similar to those we found in the orig-

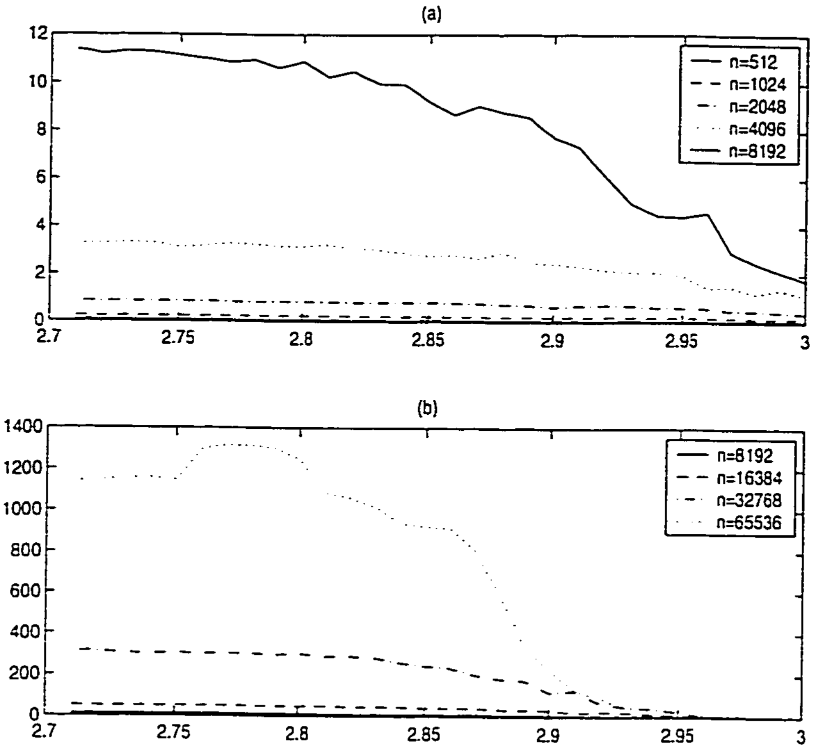

Figure 3.6: Average search cost(Y-axis, in seconds) as a function of z (X-axis) for 2-SAT sub-problems.

inal 3-SAT problems. There is a soluble-insoluble phase transition occurring around 2.83, but the fraction of unsatisfiable instances is lower than the fraction in the original 3-SAT problems. This may be possibly explained in two ways: (1)In the original 3-SAT problem, other than the 2-SAT sub-problems, there may be some other sub-problems that can be used to determine the unsatisfiability of the instances: or (2) The gap in the fraction of unsatisfiable instances is due to the effect of the finite problem size. We are currently not sure if the second explanation makes sense because we have tried the problem size up to $2^{17} = 131072$.

Comparing Figure 3.3 and Figure 3.7, it is interesting to notice that for the relsat, the average search cost of satisfiable instances for the 2-SAT sub-problems remains the same as that for the original 3-SAT problems, while the

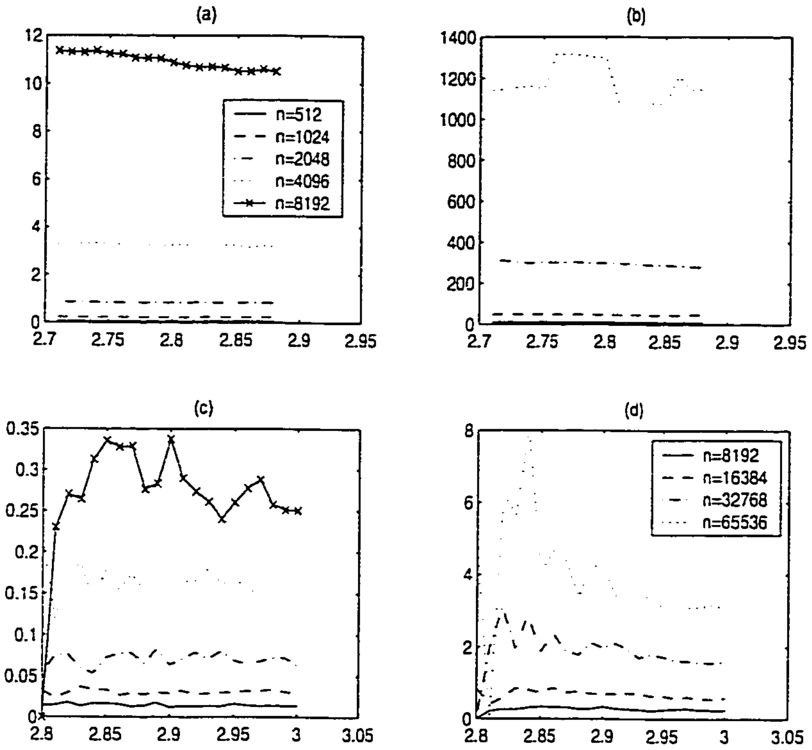

Figure 3.7: Search cost (Y-axis, in seconds) as a function of z (X-axis) for soluble instances (sub-figures (a) and (b)) and insoluble instances(sub-figures (c) and (d)) for 2-SAT sub-problems.

average search cost of the unsatisfiable instances for the 2-SAT sub-problems is much less than that for the original 3-SAT problems. This tells us that the difficulty of solving a soluble instance of NK landscape is almost the same as that of solving a 2-SAT problem which is polynomially solvable, and hence is easy. Therefore, on average the NK landscape $N(n, 2, z)$ is also easy at parameters below the threshold where almost all of the instances are soluble.

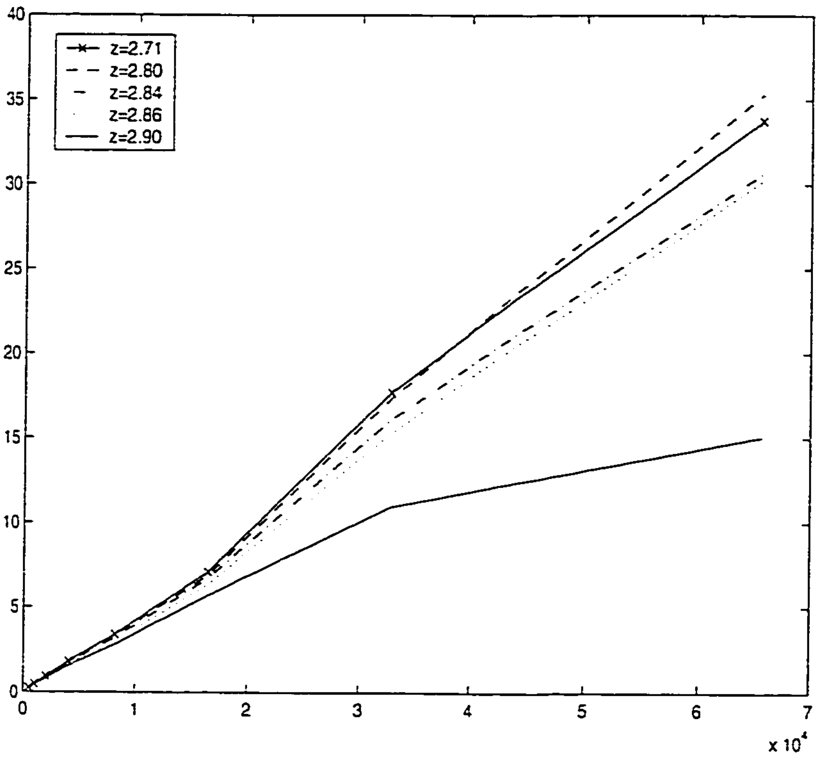

Figure 3.8: Square root of the average search cost (Y-axis, in seconds) as a function of n (X-axis) for 2-SAT sub-problems.

Chapter 4

Conclusion

The NK landscape model, proposed by Kauffman [KAU89], has been widely used as a prototype and benchmark in the study of genetic algorithms. In the literature, it has been discussed from the perspectives of statistics and computational complexity. In this thesis, we have studied the decision version of the NK landscape model in terms of the threshold phenomena and phase transition. We have theoretically and empirically investigated the problem under two random models: the uniform probability model and the fixed ratio model.

Our analyses show that the uniform model is trivially sovable as the problem size tends to infinity. For the fixed ratio model, we have derived two upper bounds for the threshold of the solubility phase transition, and proved that the problem with the control parameter above the upper bounds can be solved polynomially with probability asymptotic to 1 due to the existence of easy subproblems such as 2SAT. A series of experiments has also been conducted to investigate the hardness of the problem with the control parameters around and below the threshold. From the experiments, we have observed that the problem is also easy around and below the threshold of the phase transition.

There are several problems that are worth further exploration.

(1). Establishing lower bounds on the threshold of phase transition of the fixed ratio model. As in the study of the SAT phase transition [FRA84, FRG98, FRP83, FRS95a, FSS96], this usually requires us to come up with some polynomial algorithms which show that the problem is asymptotically satisfiable if the control parameter is below the lower bounds. Currently, we only have a trivial lower bound of $z = 1$.

(2). Explaining the gap between the theoretical upper bound of the threshold and the empirical thresholds obtained by solving the original NK landscape decision problem and the 2-SAT sub-problem (see Sections 3.3.2, 3.5.2, and 3.5.3). Our theoretical analysis in Section 3.3.2 gives the upper bound $z = 2.837$ by using only the set of 2-clauses generated by resolving each local fitness function separately. Our empirical results, however, indicate that in the region [2.84, 2.9], there are certain fraction of instances that are insoluble, but their insolubility cannot be determined by the 2-SAT sub-problems generated according to our theoretical analysis. We have done the experiments for the problem size up to $n = 65536$, and believe that it is large enough for the frequencies to converge to the probabilities in the ordinary sense. We still cannot come up with a satisfactory explanation for this phenomenon.

(3). Generalizing the analysis to SAT and CSP problems. Both the uniform probability model and the fixed ratio model can be converted to equivalent SAT problems. This is the approach that we have used in the empirical study. It should be noted that by converting the random NK landscape model to SAT, we get a new random SAT model which is different from those widely used in the study of SAT phase transition. In fact, the new random SAT model has a special structure in which clauses are divided into several closely related clusters. Further exploration of this special random SAT model may give us some interesting results about the nature of the SAT problem. In the meantime, maybe this clustered random model is more suitable for modeling the SAT and CSP problems encountered in practice.

Bibliography

[ACH99a] D.Achlioptas, "Setting 2 Variables at a Time Yields a New Lower Bound for Random 3-SAT", *MicroSoft Tech. Report MSR-TR-99-96*, 1999.

[ACH99b] D.Achlioptas, *Threshold Phenomena in Random Graph Colouring and Satisfiability*, 1999.

[AGK00] D.Achlioptas, C.Gomes, H.Kautz, and B.Selman, "Generating Satisfiable Problem Instances", To appear in AAAI00, Austin, Texas, 2000.

[AKK97] D.Achlioptas, L.M.Kirousis, E.Kranakis, D.Krizanc, M.Molloy, and Y.C.Stamation, "Random Constraint Satisfaction: A more Accurate Picture", In *Proceedings of CP97*, pp.107-120, Springer, 1997.

[ALS92] N.Alon and J.H.Spencer, *The Probabilistic Method*, New York: Wiley, 1992.

[ALT96] L.Altenberg, "NK Fitness Landscapes", in T.Back, et al.(Eds.), *The Handbook of Evolutionary Computation, Section B.2.7.2*, 1997, Oxford University Press.

[BBC99] B.Bollabás, C.Borgs, J.T.Chayes, J.H.Kim, and D.B.Wilson, "The Scaling Window of the 2-SAT Transition", *MicroSoft Tech. Report MSR-TR-99-41*, 1999.

[BFU93] A.Z.Broder, A.M.Frieze, and E. Upfal, "On the Satisfiability and Maximum Satisfiability of Random 3-CNF Formulas", in *4th Annual ACM-SIAM Symposium on Discrete Algorithms*(Austin, TX, 1993), pp.322-330, New York: ACM Press, 1993.

[BHH97] T.Bäck, U.Hammel, and H-P.Schwefel, "Evolutionary Computation:Comments on the History and Current States", *IEEE Transactions on Evolutionary Computation 1(1)*, pp.3-17, 1997.

[BOT86] B.Bollobás and A.Thomason, " Threshold Functions", *Combinatorica 7*, pp.35-38, 1986.

[BSC97] R.J.Bayardo and R.C.Schrag, "Using CSP Look-Back Techniques to Solve Real-World SAT Instances", in *Proc. of the 14th National Conf. on Artificial Intelligence* , 203-208, 1997.

[CHV91] V.Chvátal, " Almost all graphs with 1.44n edges are 3-colorable", *Random Structures and Algorithms 2(1)*, pp.11-28, 1991.

[CHR92] V.Chvátal and B.Reed, " Mick Gets some (the odds are on his side", in *Proc. of 33rd Symposium on the Foundations of Computer Science*, pp.620-627, IEEE Press, 1992.

[CKT91] P.Cheeseman, B.Kanefsky, and W.Taylor, " Where the really hard problems are", in *Proceedings of the 12th IJCAI*, pp.331-337, 1991.

[CMI97] S.A.Cook and D.G.Mitchell, "Finding Hard Instances of the Satisfiability Problem: A Survey", *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1997.

[CUL96] J.Culberson and J.Lichtner, "On Searching A-ary Hypercubes and Related Graphs", in *Foundations of Genetic Algorithms 4*, Richard K. Belew and Michael D. Vose (eds.), pp.263-290, 1996.

[CUL99a] J.Culberson and I.P.Gent, " Well Out of Reach: Why Hard Problems Are Hard?", *Technical Report APES-13-1999*, APES Group, 1999. Http:// apes.cs.strath.ac.uk/apesreports.html

[CUL99b] J.Culberson and I.P.Gent, "Frozen Development in Graph Coloring", *Technical Report APES-19*, APES Group, 1999. Http://apes.cs.strath.ac.uk/apesreports.html

[DBM00] O.Dubois, Y.Boufkhad, and J.Mandler, " Typical Random 3-SAT Formula and the Satisfiability Threshold", to appear in SODA 2000.

[EIB96] A.E.Eiben and C.A.Schippers, " Multi-Parent's Niche: n-ary Crossovers on NK Landscape", In *Parallel Problem Solving from Nature-PPSN IV*, M.M.Voigt, et al (Eds.), pp.319-328, Berlin: Springer, 1996.

[ERD59] P.Erdös, "Graph Theory and Probability", *Canadian J. of Math. 11*, pp.34-38, 1959.

[ERR60] P.Erdös and A.Renyi, "On the Evolution of Random Graphs", *Publ. Math. Inst. Hungar. Acad. Sci. 5*, pp.17-61,1960.

[FRA84] J.Franco, "Probabilistic Analysis of the Pure Literal Heuristic for the Satisfiability Problem", *Ann. Oper. Res. 1*, pp.273-289, 1984.

[FRE96] J.W.Freeman, " Hard Random 3-SAT Problems and the Davis-Putman Procedure', *Artificial Intelligence 81*, pp.183-198, 1996.

[FRG98] J.Franco and A.V.Gelder, " A Perspective on Certain Polynomial Time Solvable Classes of Satisfiability," 1998. To appear in *Discrete Applied Mathematics.*

[FRI99] E.Friedgut, "Necessary and Sufficient Conditions for Sharp Thresholds of Graph Properties, and the k-SAT problem", *J. Amer. Math. Soc. 12*, pp.1017-1054,1999.

[FRM97] A.Frieze and C.McDiarmid, " Algorithmic Theory of Random Graphs", *Random Structures and Algorithms 10*, pp.5-42, 1997.

[FRP83] J.Franco and M.Paul, "Probabilistic Analysis of the Davis-Putnam Procedure for Solving Satisfiability", *Discrete Applied Mathematics 5*, pp.77-87, 1983.

[FRS95a] J.Franco and R.Swaminathsn, "Average Case Results for Satisfiability Algorithms Under the Random Clause Width Model", in *15th International Symposium on Mathematical Programming*, University of Michigan, Ann Arbor, 1995.

[FRS95] J.Franco and R.Swaminathsn,"Toward a Good Algorithm for Determining Unsatisfiability of Propositional Formulas", *Tech. Report*, 1995. http://www.ececs.uc.edu/ franco/Reports/hitset.ps

[FSS96] A.M.Frieze and S.Suen, " Analysis of two Simple Heuristics on a Random Instance of k-SAT," *J. of Algorithm 20(2)*, pp.312-355, 1996.

[GEW96] I.P.Gent and T.Walsh, " The Satisfiability Gap", *Artificial Intelligence 81*, pp.59-80, 1996.

[GMP98] I.P.Gent, E.MacIntyre, P.Prosser, B.M.Smith, T.Walsh, "Random Constraint Satisfaction: Flaws and Structure", APES Research Group, University of Strathclyde Report APES-08-1998, 1998. http://www.cs.strath.ac.uk/ apes/reports/apes-08-1998.ps.gz

[GOE92] A.Goerdt, "A Threshold for Satisfiability", in *Proc. of 17th International Symposium on Foundations of Computer Science*, I.M.Havel and V.Koubek (eds.), pp. 264-274, Berlin: Springer, 1992.

[GOL79] A.Goldberg, "On the Complexity of the Satisfiability Problem", *Courant Computer Science Report No.16*, New York University, 1979.

[GUJ96] J.Gu,P.W.Purdom, J.Franco, and B.Wah, " Algorithms for the Satisfiability (SAT) Problem: A Survey", *DIMACS Series on Discrete Mathematics and Theoretical Computer Science 35*, pp.19-151, American Mathematical Society, 1997.

[HOL92] J.Holland, *Adaptation in Natural and Artificial Systems*, Cambridge: MIT Press, 1992.

[HOR97] W.Hordijk, "A Measure of Landscapes", *Evolutionary Computation 4(4)*, pp.335-360, 1997.

[IST98] G.Istrate, " The Phase Transition in Random Horn Satisfiability and its Algorithmic Implications", submitted *Random Structure and Algorithms*. http://cnls.lanl.gov/ gistrate/

[JON95] T.C.Jones, *Evolutionary Algorithms, Fitness Landscapes and Search*, PhD Thesis, University of New Mexico, Albuquerque, NM.

[JSV99] S.Janson, Y.C.Stamation, and M.Vamvakari, "Bounding the Unsatisfiability Threshold of Random 3-SAT", submitted to *Random Structure and Algorithms*, 1999.

[KAR95] M.Karonski, " Random Graphs", in *Handbook of Combinatorics, vol 1*, R.L. Graham, M.Grotschel and L.Lovasz (eds), Cambridge, MA: MIT press, 1995.

[KAU89] S.A.Kauffman, "Adaptation on Rugged Landscape", in D.Stein(Ed.), *Lectures in Science of Complexity*, vol.1, pp.527-618, 1989, Addison-Wesley Longman.

[KAU93] S.A.Kauffman, *The Origins of Order*, Oxford University Press, New York, 1993.

[KGV83] S.Kirkpatrick, C.D.Gelatt, M.P.Vecchi, "Optimization by Simulated Annealing", *Science*, vol.220, pp.671-680, 1983.

[KKK98] L.M.Kirousis, P.Kranakis, D.Krizanc, and Y.Stamation, "Approximating the unsatisfiability threshold of random formulas ", *Random Structures and Algorithms 12(3)*, pp.253-269, 1998.

[KOS83] J.Komlos and E.Szemerédi, "Limit Distributions for the existence of Hamilton Circuits in a Random Graph", *Discrete Mathematics 43*, pp.55-63, 1983.

[KRP95] A.Kamath, R.Motwani, K.Palem, and P.Spirakis, "Tail Bounds for Occupancy and the Satisfiability Threshold Conjecture", *Random Structure and Algorithms 7(1)*, pp.59-80, 1995.

[KSE94] S.Kirkpatrick and B.Selman, "Critical Behavior in the Satisfiability of Random Boolean Expression", *Science 264*, pp.1297-1301, 1994.

[LOV68] L.Lovasz, "On Chromatic Number of Finite Set-systems," *Acta Math. Acad. Sci. Hung. 19*, pp.59-67, 1968.

[LUC91] T.Luczak, "Size and connectivity of the k-core of a random graph", *Discrete Mathematics 91(1)*, pp.61-68, 1991.

[MAV95] A.EL Maftouhi and W.Fernandez de la Vega, " On Random 3-SAT", *Combin. Probab. Compt. 4(3)*, pp.189-195, 1995.

[MDS91] B.Manderick, M.deWeger, and P.Spiessens, "The Genetic Algorithm and the Structure of the Fitness Landscape", in R.K. Belew and L.B. Bookers (Eds.), *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991, pp.143-150, Sana Mateo, CA: Morgan Kauffman.

[MLE96] D.G.Mitchell and H.J. Levesque, "Some Pitfalls for experiments with random SAT", *Artificial Intelligence 81*, pp.111-125, 1996

[MSL92] D.G.Mitchell, B.Selman, and H.Levesque, "Hard and Easy Distributions of SAT problems", in *Proceedings of the 10th Natl. Conf on Artificial Intelligence*, pp. 459-465, Menlo Park, CA: AAAI Press, 1992.

[MFH92] M.Mitchell, S.Forresr, and J.H.Holland, " The Royal Road for Genetic algorithms: Fitness Landscape and GA Performance", in F.J.Varela et al.(Eds.), *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, 1992, pp.245-254, Cambridge, MA: MIT Press.

[MOL92] M.Molloy, " The Chromatic Number of Sparse Random Graphs", M.Math Thesis, University of Waterloo, 1992.

[MOL98] M.Molloy, "The Probabilistic Methods", in *Probabilistic Methods for Algorithmic Discrete Mathematics*, M.Habib, C.McDiarmid, et al.(eds), New York: Springer, 1998.

[MOZ96] R.Monasson, R.Zecchina, "Entropy of the Random K-satisfiability Model", *Phys. Rev. Lett. 76(21)*, pp.3881-3885, 1996.

[MOZ97] R.Monasson, R.Zecchina, "Statistical Mechanics of the Random K-satisfiability Model", *Phys. Rev. E(3) 56(2)*, pp.1357-1370, 1997.

[MZK99a] R.Monasson, R.Zecchina, S.Kirkpatrick, M.Selman, and L.Troyansky, "2+p-SAT: Relation of Typical-case Complexity to the Nature of the Phase Transition", *Random Structure and Algorithms 15(3-4)*, pp.414-435, 1999.

[MZK99b] R.Monasson, R.Zecchina, S.Kirkpatrick, M.Selman, and L.Troyansky," Determining Computational Complexity From Characteristic Phase Transition", *Nature 400(8)*, pp.133-137, 1999.

[POT98] M.Potter, "NK landscape Model: A Tutorial", http://www.cs.gmu.edu/ mpotter/nk-generator/.

[PSW96] B.Pittel, J.H.Spencer, and N.C.Wormald, " Sudden emergence of a giant k-core in a random graph", *J. of Combin. Theory, Ser. B 67(1)*, pp.111-151, 1996.

[SML92] B.Selman, H.J.Levesque, and D.Mitchell, " A new method for solving hard satisfiability problems", in *Proc. AAAI-92*, pp.440-446, San Jose, CA, 1992.

[SMH96] B.Selman, D.G.Mitchell, and H.J.Levesque, "Generating hard satisfiability problems", *Artificial Intelligence 81*, pp.17-29, 1996.

[THO95] R.K.Thompson, *Fitness Landscapes Investigated*, Master's Thesis, University of Montana, Mansfield Library, Missoula, MT, 59812, 1995

[VCU98] B.Vandegriend and J.Culberson, " The G(n,m) Phase Transition is not Hard for the Hamiltonian Cycle", *Journal of Artificial Intelligence Research 9*, pp.219-245, 1998.

[WEI96] E.D.Weinberger, " NP Completeness of Kauffman's NK Model: a tunable rugged fitness landscape", *Working Papers 96-02-003*, Santa Fe Institute, 1996, Santa Fe, FM.

[WEI91] E.D.Weinberger, "Local Properties of Kauffman's NK Model", *Physical Review A 44(10)*, pp.6399-6413, 1991.

[WEI90] E.D.Weinberger, "Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference", *Biological Cybernectics 63*, pp.325-336, 1990.

[WHI91] L.D.Whitley, "Fundamental Principles of Deception in Genetic Search", In *Foundations of Genetic Algorithms,*G.J.E. Rawlins (ED.), pp.221-241, Morgan Kaufmann: San Mateo,1991.

[WRI32] S.Wright, " The Roles of Mutation, Inbreeding, CrossBreeding and Selection in Evolution", in *Proceedings of the Sixth International Conference on Genetics*, D.Jones(Ed.), Vol. 1, pp.356-366, 1932.

[WTZ99] A.H.Wright, R.K.Thompson, J.Zhang, "The Computational Complexity of N-K Fitness Functions", *Technical Report*, Department of Computer Science, University of Montana, 1999.