INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Assessing the performance of the simulated annealing algorithm using information theory

Fleischer, Mark Alan, Ph.D.
Case Western Reserve University, 1994

Copyright ©1993 by Fleischer, Mark Alan. All rights reserved.
ASSESSING THE PERFORMANCE
OF THE
SIMULATED ANNEALING ALGORITHM
USING INFORMATION THEORY

by

Mark Fleischer

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Dissertation Advisor: Sheldon H. Jacobson

Department of Operations Research
CASE WESTERN RESERVE UNIVERSITY

January, 1994
Copyright © (1993) by Mark Alan Fleischer
CASE WESTERN RESERVE UNIVERSITY

GRADUATE STUDIES

We hereby approve the thesis of

MARK A. FLEISCHER

candidate for the PH.D.
degree.*

(signed)

(chafr)

(Jain H.)

(Prof. Papachristou)

(date) 7/29/93

*We also certify that written approval has been obtained for any proprietary material contained therein.
I grant to Case Western Reserve University the right to use this work, irrespective of any copyright, for the University's own purposes without cost to the University or to its students, agents and employees. I further agree that the University may reproduce and provide single copies of the work, in any format other than in or from microforms, to the public for the cost of reproduction.

[Signature]
ASSESSING THE PERFORMANCE
OF THE
SIMULATED ANNEALING ALGORITHM
USING INFORMATION THEORY

Abstract

by

Mark Fleischer

The simulated annealing (SA) algorithm has traditionally been viewed as a simulation of a thermodynamic system that can be used to solve combinatorial optimization problems. In this dissertation, SA is studied from the perspective of information theory and viewed as an inhomogeneous Markov information source. The concepts of the entropy of ergodic information sources and the Asymptotic Equipartition Property (AEP) from information theory are then utilized to explain the finite-time behavior of SA.

This dissertation extends the AEP to the case of strongly ergodic inhomogeneous Markov chains and then specializes it to incorporate the concept of convergence in distribution. The theory is then applied to Markov chains of the type encountered in SA. This makes it possible to define typical sequences of states in an annealing experiment and the relationship between them and the entropy measure of an SA experiment. The theory developed shows that the finite-time behavior of SA is directly related to this entropy measure. This entropy effect can therefore be used as a guide in developing methods to improve the finite-time performance of SA. Four methodologies are employed to illustrate this entropy effect which involve modification of the neighborhood structure and polynomial transformations between NP-hard problems.
DEDICATION

To the memory of my beloved father Harold C. Fleischer,
and to my mother Renee S. Fleischer
ACKNOWLEDGEMENTS

My first acknowledgement goes to my parents, sisters and brothers-in-law. Without their love and support, none of this would have been possible. I want to thank my dissertation advisor, Dr. Sheldon H. Jacobson, for showing me the ropes, for helping me accomplish my goals, for being a friend and a mentor. There were countless times I relied on his experience and expertise. His laser-like mind was naturally a big help, but it was so many of his other qualities that helped me not only survive the process of learning to do research, but succeed at it.

I want to thank Professor Daniel Solow who also served on my dissertation committee and at one time was a co-advisor. He practices what he preaches and is a dedicated and inspiring educator. I thank him for the opportunity to work on his upcoming management science textbook and for showing me that a working relationship can be one of mutual respect and friendship. I want to thank Professors John Wilson and Christos Papachristou for also serving on my dissertation committee. All through this process they gave me guidance and encouragement when I needed it.

I want to thank the entire Department of Operations Research and its chairman, Dr. Hamilton Emmons, for their financial support and for sponsoring support from BP Oil Company and Prentice-Hall, Inc. I also want to thank Dean Edward Verhosek for his acts of kindness, understanding, encouragement and help. He is a real gentleman. Many times in the last umpteen years I stood by Elaine Iannacelli's office door to garner a bit of her charm, friendship, and encouragement. She was always there to lend an ear and help out in the many tasks that needed to be done on the road to completing this dissertation. Often she went out of her way to make things a bit easier for me. I will always be fond of her. Tedda Nathan was also someone I relied on for support and friendship. Her humor and good will made the good days better and the tough days all the more bearable.

Now that I have had a chance to take stock of this entire affair, it seems almost impossible that it was only by virtue of this graduate program I have been in for so long that I made such warm
and enduring friendships! These were some of the greatest dividends I have received from my graduate school career! How could I not have known these people before? Dr. Mendel Singer showed me so many survival techniques, offered his friendship and love, and many times saved me from myself. He is an enlightened one and a life-long friend. My friend Dr. Viswanathan (Vish) with whom I shared my office and a number of experiences at school and at BP Oil Co. and other places made the program all the more rewarding. His sense of humor was a joy to behold. Dr. (soon-to-be) Khalid Bouzina is a friend with whom I shared much comaraderie. His outlook and joy of life made the difficult times almost fun to experience. Dr. Mian Aka and Dr. (soon-to-be) Xiaomei Xu, Dr. Bardia Kamrad and Dr. Ali Benmerzouga and many others all put up with my sense of humor and bounced it back to me—I will miss their company.

I want to thank my lifelong friends Drs. Alan Weiss and Lester Longley for always being there and for making me feel I was part of the clan even before my induction. Their help, friendship and guidance made this process possible. Dr. Jane Buder-Shapiro gave me encouragement, friendship and love and made the ultimate goal seem real and achievable to me. Last, but not least, I want to thank my sweet Melissa (Melissa Kay). Her abiding love, friendship, and support helped me to achieve my goals, and sustained me during the most difficult times. She put up with all my mood swings during the writing of this dissertation, gave me her patient love and charm and has taught me the most wonderful and beautiful things about life! I am forever grateful to her.
# TABLE OF CONTENTS

Introduction ........................................................................................................... 1

Chapter 1: The History and Theory of Simulated Annealing ........................................... 4
1.1. Simulated Annealing: Basic Theory ..................................................................... 4
1.2. Historical Background ......................................................................................... 5
1.3. Markov Chains and Simulated Annealing ............................................................. 7

Chapter 2: Elements of Information Theory ................................................................ 8
2.1. Entropy as a Measure of Uncertainty .................................................................... 8
2.2. The Entropy of Homogeneous Markov Chains .................................................... 10
2.3. The Asymptotic Equipartition Property ............................................................... 13
2.4. The Relationship Between Entropy and the Set of Typical Sequences ................. 16

Chapter 3: Information Theory and Simulated Annealing ......................................... 18
3.1. Extensions of the AEP to the Inhomogeneous Case ............................................. 18
3.2. The Entropy of Annealing Experiments ............................................................... 33
   3.2.1. The Effect of Neighborhood Structure on Entropy .................................... 34
   3.2.2. Polynomial Transformations ....................................................................... 40
3.3. Summary ........................................................................................................... 41

Chapter 4: The Performance of Simulated Annealing ................................................. 43
4.1. Measuring Performance ....................................................................................... 43
4.2. The Convergence of SA ...................................................................................... 44
   4.2.1. SA and the AEP in the Limit ....................................................................... 45
   4.2.2. The Convergence Rate of SA ................................................................... 46
4.3. The Final State—Considerations ......................................................................... 50
   4.3.1. The Distribution of State Pairs ................................................................... 50
4.4. Entropy and Convergence in the Limit ............................................................... 54
4.5. Scaling Properties of SA ................................................................................... 62
   4.5.1. The Magnitude of the Error ..................................................................... 66

Chapter 5: Experimental Methodology ..................................................................... 67
5.1. Generic Annealing Experiments ........................................................................ 67
   5.1.1. The Generic Configuration Space ............................................................... 68
      5.1.1.1. Creating Generic Objective Function Values ......................................... 69
      5.1.1.2. Creating the Neighborhood Structure ................................................... 69
5.2. Experiments on Combinatorial Optimization Problems ...................................... 70
   5.2.1. NP-completeness and NP-hard Optimization Problems .............................. 70
   5.2.2. Objective Functions and Neighborhood Size ............................................. 71
      5.2.2.1. Independent Set ................................................................................... 72
      5.2.2.2. Clique ................................................................................................. 74
      5.2.2.3. Vertex Cover .................................................................................... 74
      5.2.2.4. 3-SAT ............................................................................................... 75
   5.2.3. Implementation Schemes ............................................................................ 75
   5.2.4. Effect of Transformations on the Configuration Space ............................... 77
      5.2.4.1. Transformations between Independent Set, Clique and Vertex Cover .... 77
      5.2.4.2. Transformation from 3-SAT to Clique .................................................... 79
5.3. The Cooling Schedule ....................................................................................... 80
5.4. Calculation of the Entropy Measure ........................................................................ 82
  5.4.1. Calculating $\Delta f$ ......................................................................................... 82
  5.4.2. Calculation of the State Probabilities ............................................................... 85
5.5. Calculation of the Expected Value, Variance, and Norm ....................................... 86
  5.5.1. The Expected Value and Variance ................................................................. 86
  5.5.2. The Norm ..................................................................................................... 87

Chapter 6: Experimental Results .................................................................................... 88
6.1. Generic Experiments ............................................................................................. 88
  6.1.1. Solution Space A .......................................................................................... 89
    6.1.1.1. Experiment A1 ....................................................................................... 90
    6.1.1.2. Experiment A2 ....................................................................................... 91
    6.1.1.3. Experiment A3 ....................................................................................... 93
  6.1.2. Solution Space B .......................................................................................... 95
    6.1.2.1. Experiment B1 ....................................................................................... 96
    6.1.2.2. Experiment B2 ....................................................................................... 98
    6.1.2.3. Experiment B3 ....................................................................................... 100
6.2. Combinatorial Experiments ................................................................................ 102
  6.2.1. Transformation Experiment A ....................................................................... 102
  6.2.2. Transformation Experiment B ....................................................................... 105
6.3. Conclusion ........................................................................................................... 107

Chapter 7: Simulated Annealing: Present and Future .................................................. 109
7.1. Summary of This Research ................................................................................... 109
  7.1.1. Modeling SA Using Information Theory ....................................................... 109
    7.1.1.1. Extensions of The AEP ............................................................................ 110
    7.1.1.2. Using the Final State as an Estimator of The Optimal Solution .......... 110
    7.1.1.3. Development of Scaling Properties of SA ............................................. 111
    7.1.1.4. Theoretical Association of Entropy and the Finite-Time Performance of SA ....................................................................................................... 111
  7.1.2. Experimental Validation of the Theory ......................................................... 112
    7.1.2.1. Generic Experiments ............................................................................. 112
    7.1.2.2. Combinatorial Experiments ................................................................ 113
  7.1.3. Applications .................................................................................................. 113
  7.1.4. Conclusion .................................................................................................... 114
7.2. Directions for Future Research ........................................................................... 115
  7.2.1. Estimating the Entropy .................................................................................. 115
  7.2.2. The AEP ..................................................................................................... 116
  7.2.3. Counting the Combinations of State Pairs ............................................... 116
  7.2.4. Bounding the Error in Using the Scaling Property ................................... 117
  7.2.5. Relating the Eigenvalues and Entropy ....................................................... 117
  7.2.6. Noisy Information Sources ....................................................................... 118
  7.2.7. Parallel Implementations ............................................................................. 118
  7.2.8. Other Transformation Experiments ......................................................... 119

Appendix A .................................................................................................................. 120

Appendix B .................................................................................................................. 145

References .................................................................................................................... 148
TABLE OF FIGURES

Chapter 3
Figure 1: Examples of Optimal and Suboptimal States in Two Annealing Sequences .......... 23
Figure 2: Frequency of Optimal States in Each Subsequence ........................................... 24
Figure 3: Value of $p$ in $p$ .................................................................................................... 38

Chapter 4
Figure 1: Generic Configuration Space .................................................................................. 48
Figure 2: Number Ranges of State Pairs ............................................................................... 58
Figure 3: State of the System at Low Temperatures ................................................................. 59
Figure 4: Four Classes of State Pairs ..................................................................................... 61
Figure 5: State of the System at a Higher Temperature ............................................................ 63
Figure 6: Illustration of a Super-Node .................................................................................... 63

Chapter 5
Figure 1: 3-SAT to Clique Transformation .......................................................................... 79

Chapter 6
Figure 1: Experiment A1 ......................................................................................................... 91
Figure 2: Experiment A2 ......................................................................................................... 93
Figure 3: Experiment A3 ......................................................................................................... 95
Figure 4: Experiment B1 ......................................................................................................... 97
Figure 5: Experiment B2 ......................................................................................................... 99
Figure 6: Experiment B3 ......................................................................................................... 101
Figure 7: Transformation Experiment A .................................................................................. 104
Figure 8: Entropy Versus Penalty Parameter ......................................................................... 106

Chapter 7
Figure 1: Function Used in the Inhomogeneous AEP .............................................................. 116
LIST OF TABLES

Chapter 3

Table 1: Relationship Between Row Entropy and Neighborhood Size ........................................36
Table 2: Effect on Entropy Contribution Caused by the Addition of a New Neighbor ...............40

Chapter 4

Table 1: Example of Pair Distribution ......................................................................................52
Table 2: Example of Pair Symbols ............................................................................................52
Table 3: Correspondence Between Sequences and Transition Symbols ..................................53

Chapter 5

Table 1: Neighbors to Solution 00100100 ..............................................................................84

Chapter 6

Table 1: Experiment A1 ............................................................................................................90
Table 2: Experiment A2 ............................................................................................................92
Table 3: Experiment A3 ............................................................................................................94
Table 4: Experiment B1 ............................................................................................................96
Table 5: Experiment B2 ............................................................................................................98
Table 6: Experiment B3 ..........................................................................................................100
Table 7: Computer Output for 3-SAT A ..................................................................................103
Table 8: Computer Output For Clique A ................................................................................103
Table 9: Computer Output For 3-SAT B ................................................................................105
Table 10: Computer Output For Clique B ..............................................................................105
INTRODUCTION

This dissertation explores how information theory can be used to improve our understanding of the finite-time performance of the simulated annealing (SA) algorithm. Since its inception just over ten years ago, the SA algorithm has been seen as a valuable tool for obtaining solutions to combinatorial optimization problems. In that time, the SA algorithm has been analyzed by many scientists seeking to improve the understanding of how the algorithm behaves on real world problems and also to discover methods to improve its performance in solving these problems. In this dissertation, a new perspective is used to add to our theoretical understanding of SA. This perspective comes from viewing SA as an information source. By drawing on some of the ideas in information theory, modifying and extending them where necessary, results are obtained that provide new insights into the finite-time behavior of SA.

Although the asymptotic behavior of SA has been extensively studied, it is the finite-time behavior that interests practitioners. Implementations on real world problems necessarily involve finite-time runs of the SA algorithm. Many researchers have attempted to develop systematic methods to improve the finite-time performance of SA by altering what is known as the configuration space, the set of solutions together with their associated neighbors (Tovey, [1988]). Most of these efforts have involved modifying the neighborhood size used in SA. This has lead to differing and, at times, seemingly contradictory results. Waterman and Goldstein [1988] apply SA to several instances of the Traveling Salesman Problem (TSP) using different neighborhood sizes. They report that the finite-time performance of SA degrades if the neighborhood size is too large or too small, the implication being that some optimal neighborhood size exists. Cheh et al. [1991] show that for some problems, including the TSP, performance improved for smaller neighborhood sizes. Yao [1991], on the other hand, reports that larger neighborhood sizes improve the performance of SA. The objective here is to provide some underlying theory to explain these
disparate experimental results (without contradicting them) while at the same time providing guidelines for improving the finite-time performance of SA by using the perspective of information theory.

The research presented relies on information theoretic concepts such as the entropy of ergodic information sources and the Asymptotic Equipartition Property (AEP) which are utilized to address the finite-time behavior of SA. This research shows that the finite-time behavior of SA is related to an entropy measure associated with the problem instance in question. In general, the higher this entropy measure, the greater the likelihood that an SA implementation will generate globally optimal solutions near the end of the simulation.

Information theory provides new tools that suggest how improvements in the finite-time performance of SA can be obtained. In the near term, this analysis addresses questions investigated by other researchers regarding the neighborhood size used in SA implementations. Longer term consequences are intimated by our use of concepts from information theory and the theory of computational complexity and polynomial reductions of problem instances. Experiments show that increases in the entropy measure associated with a problem instance lead to improvements in SA's finite-time performance. This opens up a host of possibilities for future research into the relative merits of using stochastic approaches for solving NP-hard problems and how such problems can be modified in order to obtain better solutions using such stochastic approaches.

From a practical standpoint, this research provides new experimental issues to explore such as under what circumstances is the entropy measure associated with a problem is high or low. This in turn can lead to new understanding of the type or even class of combinatorial problems for which SA will perform well. It also provides a theoretical foundation for the more anecdotal findings of when the SA algorithm performs well.

The dissertation is organized into seven chapters. Chapter 1 provides a brief overview of the SA algorithm. This covers the so-called Metropolis Acceptance Criterion, cooling schedules, and
the notion of convergence in distribution. Chapter 2 provides background on information theory. Specifically, it describes the entropy measure of uncertainty and how this uncertainty can be defined for both homogeneous and inhomogeneous Markov chains. Chapter 3 discusses the AEP. The AEP is used to explain the relationship between the entropy of homogeneous Markov chains and the number of chains of finite length having frequencies of states close to their expected values. Concepts from the AEP are then extended to develop an analogous theory for strongly ergodic inhomogeneous Markov chains of the type encountered in SA.

In Chapter 4, issues concerning measures of performance of SA are outlined. Theoretical relationships between entropy and estimators of the globally optimal solution are then developed for a special case of the SA algorithm. Scaling properties of SA are then developed and utilized to extend these results to a more general case. Chapter 5 describes the experimental methodology and implementation issues revolving around SA. Chapter 6 presents the results of experiments that demonstrate the relationships described in Chapter 4. Chapter 7 provides a summary and conclusion of the research presented here, as well as directions for future research.

Each chapter's equations, theorems and other items are numbered from 1. References to them in other chapters will include the chapter number. For example, a reference in Chapter 5 to the second equation of Chapter 2 will be indicated by (2.2). Otherwise, it will be referred to simply as (2). The proofs of theorems and lemmas developed by the author are presented in Appendix A unless its appearance in the text improves its clarity. For proofs of theorems developed by others, a clear reference is made. Miscellaneous materials such as computer code and minor proofs are provided in Appendix B.
CHAPTER 1

THE HISTORY AND THEORY OF SIMULATED ANNEALING

1.1. SIMULATED ANNEALING: BASIC THEORY

SA essentially combines aspects of local search with that of Monte Carlo experimentation by stochastically exploring the set of solutions, referred to as the configuration space, under the influence of a control parameter, referred to as temperature. This search method is based on the so-called Metropolis Acceptance Criterion (Metropolis et al. [1953]) which models how a thermodynamic system moves from its current state to a candidate state in which the energy content is being minimized with the probability expression

$$\Pr \{ \text{Accept candidate } j \text{ as next state} \} = \begin{cases} e^{-\frac{\Delta f_{ji}}{t_k}} & \Delta f_{ji} > 0 \\ 1 & \Delta f_{ji} \leq 0 \end{cases} \quad (1)$$

where \( i \) and \( j \) denote the current and candidate states of the system, respectively, \( t_k \) represents the temperature at time index \( k \), and \( \Delta f_{ji} = f_j - f_i \) where \( f_j \) and \( f_i \) denote the energy (objective function) values associated with states \( j \) and \( i \), respectively. The candidate state \( j \) is chosen at random from among the set of neighbors of state \( i \), defined by \( N(i) \), and becomes the current state (i.e., current time index) is updated from \( k \) to \( k+1 \) based on the probability given in (1). Note that (1) can be rewritten as

$$\Pr \{ \text{Accept candidate } j \text{ as next state} \} = e^{-\frac{\Delta f_{ji}^+}{t_k}} \quad \text{where } \Delta f_{ji}^+ = \begin{cases} \Delta f_{ji} & \Delta f_{ji} > 0 \\ 0 & \text{otherwise} \end{cases}$$

This acceptance probability is the basic element of the search mechanism in the SA algorithm. If the probability of generating a candidate state \( j \) from the neighbors of state \( i \) is \( 1/|N(i)| \) where \( |N(i)| \) is the cardinality of \( N(i) \), then a transition matrix can be defined with transition probabilities.
\( \tilde{p}_n^{[k]} = p_n(t_k) = \begin{cases} \frac{1}{N(i)} e^{-\beta_n(t_k)} & j \neq i, j \in N(i) \\ 0 & j \neq i, j \notin N(i) \\ 1 - \sum_{\substack{j=1 \\substack{\neq i}}}^{N} p_n(t_k) & j = i \end{cases} \) (2)

which defines an inhomogeneous Markov chain. A property of these probabilities is that as the temperature monotonically decreases to zero, the diagonal elements of the associated transition matrix are nondecreasing to one, while the nondiagonal elements are nonincreasing to zero.

The key characteristic of this search mechanism is that it provides a means of escaping from local optima by allowing for uphill moves, i.e., moves which may worsen the objective function value. As the SA algorithm searches the configuration space at descending values of \( t_k \), the state distribution associated with the inhomogeneous Markov chain converges to a distribution in which all the probability is equally concentrated on the set of globally optimal solutions.

1.2. HISTORICAL BACKGROUND

Since its development by Kirkpatrick, et al. [1983], SA methods for solving combinatorial optimization problems have received a great deal of attention in the literature. The initial research focus on SA dealt with the convergence properties of the algorithm. In particular, the rate at which \( t_k \) approached zero was the primary focus (see for example Mitra et al. [1988], Hajek [1988] and Anily and Federgruen [1987]). Other work has generally concentrated on empirical results in the application of SA to specific problems (see Johnson et al. [1989, 1990a,b], Bonomi and Lutton [1984] and Kirkpatrick et al. [1983]). Very little work has focused on studying why SA works well on some problems and not so well on others. Differences in performance appear to be related to the topology of the configuration space—what distinguishes problems from SA's perspective, and what seems to be the important factor in explaining why SA works well on some problems and not on others.
Some of this early work touched on relating the topology of the configuration space to convergence of SA. Mitra et al. [1988] showed that SA can be modelled as an inhomogeneous Markov that converges in probability given a cooling schedule of the form

$$t_k = \frac{\gamma}{\ln(k_0 + k + 1)}$$  \hspace{1cm} (3)

where $k_0$ is a constant, $k$ is the time index, and $\gamma$ is a function of the configuration space. Thus, the SA algorithm converges in probability and therefore, converges in distribution to the set of globally optimal solutions as the temperature tends to zero. In addition, the stationary distribution also converges in distribution (see Aarts and Korst [1989 pp.18]). More explicitly,

$$\lim_{t \to 0} \pi_i(t) = \frac{1}{s^*} \hspace{0.5cm} \forall \pi_i > 0$$  \hspace{1cm} (4)

where $\pi_i(t)$ corresponds to the $i$th state of the stationary distribution at temperature $t$, and $s^*$ corresponds to the cardinality of the set of globally optimal solutions. Hajek [1988] later determined necessary and sufficient conditions for SA to converge to the set of optimal states by showing a relationship involving the maximum depth over all local minima in the configuration space. He showed that convergence in probability is guaranteed if and only if

$$\sum_{x \in X} e^{-d^*/t_k} = +\infty$$  \hspace{1cm} (5)

where $d^*$ is a constant denoting the maximum depth over all local minima, not including the globally optimal solutions (see Hajek [1988] for a complete exposition). Note that $\gamma$ in Mitra et al. [1988] and $d^*$ in Hajek [1988] are crude measures of the topology of the configuration space. Nonetheless, these results indicate that convergence in probability is related to the topology of the configuration space. Empirical results for finite-time runs, however, illustrate differences in performance for different problems. See for example Johnson et al. [1989], [1990b] or Aarts and Korst [1989 p.91] for an exposition on performance experiences. The goal of this dissertation therefore, is to develop some theoretical foundation for explaining these differences in performance.
1.3. MARKOV CHAINS AND SIMULATED ANNEALING

The following notation will be used to define inhomogenous Markov chains.

Definitions:

\[ P(m) \] is the transition matrix of the Markov chain at time index \( m \).

\[ v(m + 1) = v(m)P(m) \], the state distribution vector at time index \( m+1 \), where \( v(0) \) is the state distribution vector at time index 0.

\[ P(m, m+n) = \prod_{i=0}^{n-1} P(m+i) \], the Markov transition matrix after \( n \) transitions where the first transition matrix is defined at time index \( m \), where \( P(m, m) = I \) (i.e., with zero transitions, the probability of moving to a different state is zero for all states).

\[ v(m+n) = v(m)P(m, m+n) \], the state distribution vector after \( n \) transitions where the first transition matrix is defined at time index \( m \).

\[ p_{ij}^{(m,n)} = [P(m,n)]_{ij} \] and equals the \( ij \) element of the matrix \( P(m,n) \).

Using this notation, a strongly ergodic inhomogeneous Markov chain is defined as one satisfying

\[ \lim_{n \to \infty} \sup_{v(0)} \|v(m,n) - q\| = 0 \] (6)

for all \( m \), where \( q \) is some fixed vector (see Mitra et al., and Isaacson et al. [1973 pp.139]. In SA, as the algorithm proceeds with the temperature modified in each iteration by the cooling schedule defined in (3), the state distribution vector approaches some fixed vector where all the probability is distributed only among the globally optimal solutions. These equations and relationships constitute the basic elements of the SA algorithm.
CHAPTER 2

ELEMENTS OF INFORMATION THEORY

2.1. Entropy as a Measure of Uncertainty

This chapter explores the entropy concept in information theory. Entropy is a scalar quantity which provides a method of measuring the uncertainty associated with an ensemble of events, each having a probability of occurrence. Since its first use as a measure of information by Shannon [1948], it has been extensively studied by scientists in many different fields. It has also been discussed in the context of SA (Aarts and Korst [1989] pp.20).

Since SA has its roots in simulating a thermodynamic system, SA and entropy are intimately related. The basic element of SA, the Metropolis acceptance criterion (see (1.1)) was defined to maximize the entropy of the set of possible states at a given value of t (Bonomi and Lutton [1984]). Its connection to information theory however, has not been developed. Since the SA algorithm can be modeled as an inhomogeneous Markov chain, the entropy of this Markov chain may provide a natural measure of the configuration space and suggests that entropy may be useful in assessing the finite-time performance of SA.

Definition:

The entropy $H(A)$ of a finite scheme $A$ of $n$ independent events $A_1, A_2, \ldots, A_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, is

$$H(A) = H(p_1, p_2, \ldots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log p_i$$

where $\sum_{i=1}^{n} p_i = 1$ and $p_i > 0$ for all $i$, and $\lambda$ is a scaling constant. Natural logarithms will be used from now on since the transition probabilities in SA involve the base of the natural logarithms.
The entropy measure is uniquely defined by certain properties, as stated in the following theorem.

**Theorem 1:**

If a measure of uncertainty associated with a finite scheme of \( n \) independent events \( A_1, A_2, \ldots, A_n \) with probabilities \( p_1, p_2, \ldots, p_n \), respectively, satisfies the following properties:

1) the measure should be maximized when all probabilities in the ensemble are equal; *i.e.*, 
   \[
   \max_{p_1, p_2, \ldots, p_n} H(p_1, p_2, \ldots, p_n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).
   \]

2) the measure for two independent ensembles \( A \) and \( B \) jointly occurring, indicated by the ensemble \( A \cup B \), should be additive; *i.e.*, 
   \[
   H(A \cup B) = H(A) + H(B).
   \]

3) the measure for two ensembles where the probabilities of one are conditioned on the other should be additive in a meaningful way; *i.e.*, 
   \[
   H(A \cup B) = H(A) + \sum_A p_A H(B|A),
   \]
   where \( H(B|A) \) is the entropy of \( B \) given event \( A \).

4) the addition of impossible events to an ensemble (i.e., events with probability zero), should not change the value of the uncertainty; *i.e.*, 
   \[
   H(p_1, p_2, \ldots, p_{n-1}, 0) = H(p_1, p_2, \ldots, p_{n-1}),
   \]
   then the entropy measure is the only measure of uncertainty that satisfies all these properties.

**Proof:** See Khinchin [1957 pp.9] or Ash [1965 pp.5].

The following corollary to this theorem will be needed later.
Corollary (Gibbs Inequality):

Let \( p \) and \( q \) be two probability ensembles, i.e., \( \sum_{i=1}^{n} p_i = 1 \) and \( \sum_{i=1}^{n} q_i = 1 \) with \( p_i > 0, q_i > 0 \) for all \( i, i = 1, 2, ..., n \). Then, \(-\sum_{i=1}^{n} p_i \ln p_i \leq -\sum_{i=1}^{n} p_i \ln q_i\) with equality if and only if \( p_i = q_i \) for all \( i \).

Proof: See Goldie and Pinch [1991 p.15].

2.2. THE ENTROPY OF HOMOGENEOUS MARKOV CHAINS

A measure associated with a Markov chain as a whole, is the average entropy of each state of the transition matrix associated with the Markov chain. Such a measure "characterizes the chain as a whole; it is clear that it is uniquely determined by giving the state probabilities ... and the transition probabilities ...." Khinchin [1957 p.14]. The use of entropy in this way with respect to SA has not been utilized. To date, the entropy measure has been used to describe the uncertainty associated with the stationary distribution in SA at fixed temperatures. This, however, is nothing more than a restatement of the fact that the stationary distribution converges in probability to the optimal state distribution; i.e., where all globally optimal states have equal probability and non-optimal states have zero probability (see (1.4) and Aarts and Korst [1989] pp.20). The present use of entropy describes the configuration space at each temperature in the algorithm. Other aspects of information theory are used to demonstrate certain relationships between an entropy measure of the configuration space and the behavior of an SA experiment.

Since each row of the transition matrix associated with an SA experiment is an ensemble of events, i.e., possible transitions to other states in the configuration space, an entropy measure associated with a transition matrix can be defined. First, the entropy of a homogeneous transition matrix, one that remains constant over time, is defined.
Definitions:

\[ s = \text{the number of distinct states of the Markov chain where } \{S\} \text{ is the set of all states.} \]

Therefore, \( s = |S| \). In the SA framework, it corresponds to the number of states in the annealing experiment (number of points in the configuration space) where each such state has an associated objective function value.

\[ h_i = -\lambda \sum_{j=1}^{s} p_{ij} \ln p_{ij} \]. This corresponds to the entropy of each row \( i \) (state \( i \)) of the transition matrix with \( h = (h_1, h_2, \ldots, h_s) \) and corresponds to the vector of row entropies.

\( \pi = (\pi_1, \pi_2, \ldots, \pi_s) \) corresponds to the vector of stationary probabilities.

\[ H_{mc} = -\lambda \sum_{i=1}^{s} \sum_{j=1}^{s} \pi_i p_{ij} \ln p_{ij} \text{ where } \sum_{j=1}^{s} p_{ij} = 1, p_{ij} > 0 \text{ for all } i, j \text{ and } p_{ij} \text{ corresponds to the probability of moving from solution } i \text{ to solution } j \text{ in one transition.} \]

\( H_{mc} \) is referred to as the entropy of the Markov chain defined by transition probabilities \( \{p_{ij}\} \). Without loss of generality, set \( \lambda = 1 \). For a complete discussion, see Khinchin [1957 p.14]. Therefore, \( H_{mc} = \pi \cdot h \), the dot product of the vectors \( \pi \) and \( h \).

\( \Sigma_n = \sigma_0, \sigma_1, ..., \sigma_{n-1} \), a fixed sequence of states of length \( n \) with each such state denoted by \( \sigma_k \) generated by a Markov chain. The Markov chain may be either homogeneous or inhomogeneous, depending on the context.

\( Z_n = z_0, z_1, ..., z_{n-1} \), a random sequence of states of length \( n \) with each such state denoted by the random variable \( z_k \), generated by a Markov chain. The Markov chain may be either homogeneous or inhomogeneous, and will be made clear from the context.

\[ \Pr\{\Sigma_n\} = \Pr\{Z_n = \Sigma_n\} = \Pr\{z_0 = \sigma_0, z_1 = \sigma_1, ..., z_{n-1} = \sigma_{n-1}\} = v_{\sigma_0} P_{\sigma_0 \sigma_1} P_{\sigma_1 \sigma_2} \cdots P_{\sigma_{n-2} \sigma_{n-1}}, \]

the probability measure associated with some fixed sequence \( \Sigma_n \). Each such sequence therefore has a uniquely defined probability measure given a fixed value \( v_{\sigma_0} \) for the probabilities of the initial state. When the measure is associated with an
inhomogeneous Markov chain, the probability measure will also be a function of the
time indices. This will be clear from the context. See Chapter 3 for the definition of
the probability measure associated with an inhomogeneous Markov chain.

\[ \Pr \{ Z_n \} = \nu_0 p_{00} p_{10} \ldots p_{n-2, n-1}, \] 
the probability measure associated with random sequence \( Z_n \).

This probability measure is a function of the \( p_{i,i+1} \) which is a function of the random
variables \( z_k \), hence, is also a random variable.

The row entropy, \( h_i \) is by definition the entropy associated with state \( i \), and "can be
regarded as a measure of the amount of information obtained when the Markov chain moves one step
ahead." Therefore, the average of this quantity over all initial states, i.e., the quantity \( H_{mc} \) is
therefore to be regarded as a measure of the average amount of information obtained when the given
Markov chain moves one step ahead." Khinchin [1957 p.14]. Since the entropy measure for
conditional events is additive (see 3) in Theorem (1)), the entropy of a sequence of transitions in a
Markov chain should be the sum of the entropies of each transition, weighted appropriately by the
state distribution vector, i.e., the vector of probabilities of being in each state of the system. This
concept is captured in the following theorem.

**Theorem 2:**

If \( H^{(r)} \) is the total entropy of \( r \) consecutive transitions (\( r \)-step entropy) and \( \pi^{(k)} = \pi^{(k-1)} P \) is
the state distribution vector after the \( k-1 \)th transition, then

\[
H_{mc}^{(r)} = -\lambda \sum_i \sum_j \pi_i^{(0)} p_{ij} \ln p_{ij} - \lambda \sum_i \sum_j \pi_i^{(1)} p_{ij} \ln p_{ij} - \ldots - \lambda \sum_i \sum_j \pi_i^{(r-1)} p_{ij} \ln p_{ij}
\]

\[
= \pi^{(0)} \cdot h + \pi^{(1)} \cdot h + \ldots + \pi^{(r-1)} \cdot h \tag{1}
\]

**Proof:** This follows directly from 3) in Theorem 1. For a complete discussion, see Khinchin [1957
pp.9].
Corollary:

If $\pi^{(n)} = \pi$ for all $n$, as when the chain is in steady-state, then the $r$-step entropy is

$$H_{mc}^{(r)} = rH_{mc}$$

(2)

Proof: This follows from an inductive proof, as presented in Khinchin [1957 pp.15]. □

2.3. THE ASYMPTOTIC EQUIPARTITION PROPERTY

This section describes the Asymptotic Equipartition Property (AEP) of homogeneous Markov chains. It has also been referred to as the $E$-property (Khinchin [1957]) and as the Shannon-McMillan Theorem (Ash [1965]) and is one of the central elements in information theory. The AEP relates the probability of occurrence of a typical sequence of states, defined below, to the entropy of the Markov chain. In the next section, the AEP is extended to the inhomogeneous case and used to prove the main result.

The reason the AEP is useful in assessing the finite-time performance of SA experiments is because it provides insight into the likelihood that sequences of states generated by information sources, such as annealing sequences, have certain statistical properties. For instance, as an SA experiment proceeds, it generates a sequence of states in such a way that the frequency and, hence likelihood, that globally optimal states are visited increases. Eventually, the annealing experiment terminates having visited a finite number of states and stopping in a particular state. One expects that if the experiment is allowed to run for a sufficiently long period of time, the likelihood that the final state is optimal also increases. This reflects the fact that, using a proper cooling schedule (see Mitra et al. [1988]), annealing experiments correspond to strongly ergodic Markov chains. Thus, most of the time, SA experiments demonstrate convergence in distribution. The case where an annealing experiment visits states in an unusual manner, such as when it never visits optimal states, therefore has a low probability of occurrence. The AEP, a concept associated only with homogeneous Markov information sources, addresses the issue of measuring the probability and number of highly likely sequences of finite-length generated by a Markov information source. If a
similar property as the AEP could be applied to strongly ergodic inhomogeneous Markov information sources, then such a property would show that increasing the probability and number of highly likely sequences also increases the probability and number of sequences which exhibit convergence in distribution. This could then provide predictive value that an annealing experiment ends in a globally optimal state.

The AEP states that sequences of \( n \) states in a homogeneous Markov chain can be partitioned into two mutually exclusive and exhaustive sets: \( G_1 \), with a high collective probability, and \( G_2 \), with a low collective probability. The size of each set and the probability that a random sequence is in one of these sets is affected by the length of these sequences and the entropy of the associated homogeneous Markov chain.

Based on Khinchin [1957 p. 14], given a random sequence of states \( Z_n \) of length \( n \), governed by a homogeneous Markov chain on a total of \( s \) distinct states, then from the definitions in Section 2.2, the probability that a random sequence \( Z_n \) equals a particular sequence \( \Sigma_n \) is

\[
Pr\{\Sigma_n\} = \prod_{i=1}^{s} \prod_{j=1}^{s} v_{\sigma_0} \rho_{ij}^{m_{ij}}
\]

(3)

where \( v_{\sigma_0} \) is the probability of being in the initial state \( \sigma_0 \) and \( m_{ij} \) equals the actual number of transitions from \( i \) to \( j \) in a given sequence of length \( n \). If the transition matrix is irreducible and aperiodic, then it is ergodic (see Khinchin [1957 pp.16]). Such Markov chains are referred to as regular Markov information sources in the context of information theory (see Ash, pp.185). Thus, after a sufficiently large number of transitions, the relative frequencies of the states in the sequence will closely match that of the stationary probabilities \( \pi_i \). If \( m_i \) is the number of occurrences of state \( i \) in a random sequence of states of length \( n \), then from the weak law of large numbers, for any \( \delta > 0 \), and for all states \( i \),

\[
\lim_{n \to \infty} \Pr \left( \left| \frac{m_i}{n} - \pi_i \right| \leq \delta \right) = 1
\]

(4)
(see Bernoulli's Theorem in Gnedenko [1976 pp. 199]), or by the definition of a limit, for any $\varepsilon > 0$ and $\delta > 0$ there exists an $n^* > 0$ such that for all $n \geq n^*$,

$$\Pr\left\{\left|\frac{m_i}{n} - \pi_i\right| \leq \delta \right\} > 1 - \varepsilon.$$  \hspace{2cm} (5)

Using this as a guidepost, a criterion for partitioning all possible sequences of states into the two mutually exclusive sets $\Omega_1$ and $\Omega_2$ can be established. A sequence $\Sigma_n$ is defined to be typical or standard, and belongs to $\Omega_1$, if and only if the relative frequencies of all states closely matches that of the steady-state distribution. Thus, by definition, for given values of $n$ and $\delta > 0$,

$$\Sigma_n \in \Omega_1 \Leftrightarrow \forall i, \quad \left|\frac{m_i}{n} - \pi_i\right| \leq \delta.$$ \hspace{2cm} (6)

where $m_i$ is the number of occurrences of state $i$ in sequence $\Sigma_n$. From (5), the probability that a random sequence $Z_n$ satisfies this criterion for membership in $\Omega_1$ is

$$\Pr\{Z_n \in \Omega_1\} = \Pr\left\{\left|\frac{m_i}{n} - \pi_i\right| \leq \delta \right\} > 1 - \varepsilon$$ \hspace{2cm} (7)

for all states $i$, and $n$ sufficiently large, given any values of $\varepsilon > 0$ and $\delta > 0$. Using the definition of $m_{ij}$ for sequence $\Sigma_n$, this membership criterion can be restated in terms of the transition probabilities of a Markov chain

$$\Sigma_n \in \Omega_1 \Leftrightarrow \forall ij, \quad \left|\frac{m_{ij}}{n} - \pi_i\pi_{ij}\right| \leq \delta.$$ \hspace{2cm} (8)

Although the justification for this seems intuitively obvious, it is by no means trivial. Nonetheless, for regular Markov information sources, the relative frequencies of any specified string, or subsequence of states, such as $ij$ pairs, approaches its expected value with probability one i.e., by the strong law of large numbers (Goldie et al. [1991 pp. 68]). Therefore, given an arbitrary initial starting vector, for strings $\{Z_{m+1} = j, Z_m = i\}$, $\Pr\{Z_{m+1} = j, Z_m = i\} \rightarrow \pi_i\pi_{ij}$ as $m \rightarrow \infty$. For a rigorous proof, see Goldie et al. [1991 pp. 53].
2.4. **The Relationship Between Entropy and the Set of Typical Sequences.**

The following theorem shows how the AEiP relates the entropy of homogeneous Markov chains to the probability of $\Omega_1$ sequences.

**Theorem 3:**

For all sequences fix and for all $\eta > 0$ and $n$ sufficiently large,

$$\frac{-\ln \Pr\{\Sigma_n\} - H_{mc}}{n} \leq \eta,$$

(9)

where $\eta = \frac{-\ln v_{\eta_0}}{n} - \delta \sum_{i=1}^t \sum_{j=1}^t \ln p_{ij}$ can be made arbitrarily small for large $n$ and small $\delta$.

**Proof:** See Khinchin [1957 pp. 17].

Note that $\eta$ is independent of the frequencies of states in the sequence. Expression (9) can be rewritten as

$$nH_{mc} - n\eta \leq \ln \left( \frac{1}{\Pr\{\Sigma_n\}} \right) \leq nH_{mc} + n\eta$$

(10)

or,

$$e^{-n(H_{mc} + \eta)} \leq \Pr\{\Sigma_n\} \leq e^{-n(H_{mc} - \eta)}.$$ 

(11)

Consequently, since $\eta$ is small for $n$ large and $\delta > 0$ small such that $n\delta$ is small,

$$\Pr\{\Sigma_n\} = e^{-nH_{mc}}.$$ 

(12)

Therefore, the probability of a random sequence $Z_n$ being in $\Omega_1$ is approximately equal to $e^{-nH_{mc}}$. Thus, the probability of a particular $\Omega_1$ sequence occurring is related to the length of the sequence and the entropy of the Markov chain. From (7), the probability of a random sequence $Z_n$ of length $n$ being an element of $\Omega_1$ can be made arbitrarily close to one for small $\varepsilon > 0$. Since $\Omega_1 \cup \Omega_2$ is the entire set of length $n$ sequences all (i.e., $|\Omega_1 \cup \Omega_2| = s^n$), the probability of a random sequence being an element of $\Omega_2$ can therefore be made arbitrarily close to zero. Thus, $\Pr\{Z_n \in \Omega_1\} < \varepsilon$. But
because of the bounds in (11), for $n$ sufficiently large and $\delta > 0$ sufficiently small, $\Omega_1$ sequences have approximately the same probability. Since the probability of a random sequence $Z_n$ not being in $\Omega_1$ is small, there are approximately $e^{nH_{\infty}}$ sequences in $\Omega_1$.

Since the AEP relates the number of typical sequences to the entropy of the homogeneous Markov chain in question, it would be desirable if the above analysis could be modified to relate some entropy expression to an inhomogeneous Markov chain in which the relative frequencies of states are close to their expected values. The problem in the inhomogeneous case, however, is that the expected values of the occurrence of states change over time. In SA, for example, an inhomogeneous Markov chain converges in distribution, i.e., the expected numbers of optimal states increases in the latter stages of an annealing experiment. Thus, in order to modify the AEP in an appropriate way, the ramifications of the homogeneity of the Markov process must be understood.
CHAPTER 3

INFORMATION THEORY AND SIMULATED ANNEALING

3.1. EXTENSIONS OF THE AEP TO THE INHOMOGENEOUS CASE

In this chapter, a relationship is established between the entropy of strongly ergodic inhomogeneous Markov chains and the size of set $\Omega_1$. The theory developed is then used to establish a similar relationship for strongly ergodic inhomogeneous Markov chains of the type used to model SA with the goal of illuminating the finite-time behavior of annealing experiments. The following definitions are needed to extend the AEP to the inhomogeneous case.

Definitions:

$$h_i^{[j]} = -\lambda \sum_{j=1}^{i} P_{ij}^{[j]} \ln p_{ij}^{[j]}.$$  Thus, $h_i^{[j]}$ equals the entropy of row (state) $i$ of the transition matrix at time index $k$.

$h(k) = (h_1^{[k]}, h_2^{[k]}, \ldots, h_r^{[k]})$, the vector of entropies of the rows in the transition matrix at time index $k$.

$v(k) = (v_1^{[k]}, v_2^{[k]}, \ldots, v_s^{[k]})$, the state distribution vector at time index $k$.

$$H(k) = -\lambda \sum_{i=1}^{r} v_i^{[k]} \sum_{j=1}^{i} p_{ij}^{[k]} \ln p_{ij}^{[k]} = v(k) \cdot h(k),$$  the entropy of an inhomogeneous Markov chain at time index $k$, where $p_{ij}^{[k]}$ is defined by (1.2).

$$H^*(m, n-2) = \sum_{k=m}^{n-2} H(k) = \sum_{k=m}^{n-2} v(k) \cdot h(k),$$  is the sum of the entropies of each transition from time index $m$ to time index $n-2$.

$m_{ij}(m, n-1)$ = the number of $ij$ transitions occurring in a sequence between time index $m$ and time index $n-1$.

$$\tilde{p}_{ij}(m, n-2) = \left[p_{ij}^{[k_1]} \times p_{ij}^{[k_2]} \times \cdots \times p_{ij}^{[k_r]} \right]$$  where $r = m_{ij}(m, n-1)$ and $k_y$ is the time index of the $y^{th}$ transition. Thus, $k_1 \geq m$, $k_r \leq n - 2$. This is the geometric mean of all $ij$ transition
probabilities of the $ij$ transitions actually occurring in a sequence between time indices $m$ and $n - 1$.

$$\bar{E}\{m_{ij}(m,n-2)\} = \sum_{k=m}^{n-2} v_{ij}^{(k)} p_{ij}^{(k)}$$

where $\bar{E}$ is an operator on the $i$th element of $v(k)$ and the $ij$th element of the transition matrix $P(k)$. This is the expected number of transitions from state $i$ to state $j$ between time indices $m$ and $n-2$, where $n-2$ is the time index of the last transition in a sequence.

$\Sigma(m,n-1) = (\sigma_m, \sigma_{m+1}, \ldots, \sigma_{n-1})$, a fixed sequence of states generated by a Markov chain from time index $m$ to time index $n-1$, where $n-1$ is the time index corresponding to the last state in a sequence.

$Z(m,n-1) = (z_m, z_{m+1}, \ldots, z_{n-1})$, a random sequence of states generated by a Markov chain from time index $m$ to time index $n$. Note that $Z_n \equiv Z(0,n-1)$.

$Pr\{Z_n\} = v_0^{[0]} p_{\sigma_0 \sigma_1}^{[0]} p_{\sigma_1 \sigma_2}^{[1]} \cdots p_{\sigma_{n-2} \sigma_{n-1}}^{[n-2]}$, the probability measure associated with a random sequence of length $n$ generated by an inhomogeneous Markov chain. As in the homogeneous case, it is also a random quantity.

$Pr\{\Sigma_n\} = Pr\{Z_n = \Sigma_n I\} = v_0^{[0]} p_{\sigma_0 \sigma_1}^{[0]} p_{\sigma_1 \sigma_2}^{[1]} \cdots p_{\sigma_{n-2} \sigma_{n-1}}^{[n-2]}$, the probability measure associated with a fixed sequence of length $n$ generated by an inhomogeneous Markov chain. Note the relationship of the time index for the last transition and time index associated with the last state.

Note that $\bar{E}\{m_{ij}(m,m+r-1)\} = v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m+r-1]} p_{ij}^{[m+r-1]}$ in the inhomogeneous case corresponds to the term $n \pi_i p_{ij}$ in the homogeneous case. It is therefore a useful measure of the expected number of $ij$ transitions for any Markov chain.

The literature in information theory seems to deal exclusively with homogeneous Markov chains. One possible explanation for this is that the value of $H_{mc}$ can be easily calculated and hence, has some utility. The theorem that follows, however, shows that the concept of the AEP can be
extended to the case of inhomogeneous Markov chains which are strongly ergodic when the
definition of the entropy measure is based on the total uncertainty associated with the
inhomogeneous Markov chain. That such definitions of entropy may not be easily calculated is not
of immediate concern since the goal here is to demonstrate how the entropy of inhomogeneous
Markov chains directly affects the finite-time performance of SA. The following theorem establishes
the AEP for strongly ergodic inhomogeneous Markov chains.

**Theorem 1:**

Let $Z_n$ be some random sequence of $n$ states generated by a strongly ergodic inhomogeneous
Markov chain. Then

$$\lim_{n \to \infty} E \left[ -\ln \Pr \left( Z_n \right) \right] = \lim_{n \to \infty} \frac{H^* (0, n-2)}{n} \leq -\sum_{j=1}^{i} q_j \ln q_j = \ln |s^*|$$  \hspace{1cm} (1)

where the $q_j$ are the limiting values for the $v_i$.

**Proof:** See Appendix A. ■

The following corollaries to this theorem will be useful later on.

**Corollary 1 to Theorem 1:**

Let $Z(m, n-1)$, be a random sequence of states starting at time index $m$ to time index $n-1$
generated by a strongly ergodic inhomogeneous Markov chain (this therefore corresponds to
the last $n - m$ states in the random sequence $Z_n$). Then, for all $m$

$$\lim_{n \to \infty} E \left[ -\ln \Pr \left( Z(m, n-1) \right) \right] = \lim_{n \to \infty} \frac{H^* (m, n-2)}{n - m} \leq -\sum_{j=1}^{i} q_j \ln q_j = \ln |s^*|$$  \hspace{1cm} (2)

**Proof:** See Appendix A. ■

Note that (2) reduces to (1) by setting $m$ to zero. In this sense, (1) is a special case of (2).

**Corollary:**

Given any $\epsilon > 0$ and $\delta > 0$, and for all $m \geq 0$, then for $n$ sufficiently large,

$$\Pr \left[ \left| -\ln \Pr \left( Z(m, n-1) \right) - \frac{H^* (m, n-2)}{n - m} \right| \leq \delta \right] > 1 - \epsilon$$  \hspace{1cm} (3)

**Proof:** See Appendix A. ■
Note the similarity of the bracketed quantity in (3) to the statement in (2.7) for the homogeneous case. The following definition of typical sequences therefore follows naturally from these theorems.

**Definition:**

The set $\Omega_1$ is defined to be the set of all sequences $\Sigma_n$ such that for some $\delta > 0$, the following condition (based on Theorem 1) holds

$$ \left| - \ln \Pr\{\Sigma(m,n-1)\} - \frac{H^*(m,n-2)}{n-m} \right| \leq \delta $$

(4)

where $\Sigma(m,n-1)$ corresponds to the last $n-m$ states of the $\Sigma_n$.

Note that $H^*(m,n-2)/(n-m)$ is the average entropy per transition for the last $n-m-1$ transitions, whereas in the homogeneous case, $H_{nc}$ is also the average entropy per transition but is independent of the time index.

Based on the same reasoning used in the homogeneous case, the following statements can be made for all such sequences of states $\Sigma_n \in \Omega_1$:

$$ e^{-H^*(m,n-2)\cdot(n-m)} \leq \Pr\{\Sigma(m,n-1)\} \leq e^{-H^*(m,n-2)\cdot(n-m)\delta} $$

and $\Pr\{\Sigma(m,n-1)\} = e^{-H^*(m,n-2)}$ and the size of $\Omega_1$ is approximately $e^{H^*(m,n-2)}$, and grows exponentially with $H^*(m,n-2)$. The differences between the homogeneous case and the inhomogeneous case can be highlighted, however, if the probability measure of a sequence $\Sigma_n$ of length $n$ generated by an inhomogeneous Markov chain is written as

$$ \Pr\{\Sigma_n\} = v_{d_0}^{(n)} \prod_{i=1}^{s} \prod_{j=1}^{s} \hat{p}_{ij}(0,n-2)^{m_{ij}(0,n-2)} $$

(5)

where the $\hat{p}_{ij}(0,n-2)$ takes into account the dependence of $p_{ij}^{(k)}$ on $k$.

The foregoing results and statements, however, lead to difficulties insofar as showing a relationship between entropy and the finite-time performance of SA. This is because they concern the sequence as a whole and thus fail to capture the notion of convergence in distribution. For one
thing, it is difficult to show how the AEP for strongly ergodic inhomogeneous Markov chains relates to the frequency of states. Recall in (2.8), the criterion for membership in \( \Omega_i \) in the homogeneous case was based on how close \( m_{ij}/n \) is to \( \pi_i \pi_{ij} \). Also, the definition of \( m_{ij}^{(0,n-2)} \) equals the number of \( ij \) transitions for the entire sequence and therefore ignores the effect on the sequence probability caused by the ordering of the states. It is therefore conceivable that the probability measure of some sequence satisfies the bounds in (4) yet does not demonstrate convergence in distribution, which strong ergodicity implies. In SA, however, there is convergence in distribution, and therefore the likelihood is low that the latter states of an annealing sequence are suboptimal. The criterion used in defining a typical annealing sequence should somehow exclude such unlikely or atypical sequences. Consequently, the ordering of the states, and not just their frequency, should affect whether an annealing sequence is defined as typical.

It is curious that the above results for strongly ergodic inhomogeneous Markov chains have not been previously developed. This may be due to the fact that there was no additional and practical benefits from doing so. In the homogeneous case, the practicality of the AEP has been amply demonstrated in coding theory as it illuminates some of the theoretical considerations involved in the efficient transmission of information (see any of the references involving information or communication theory). The AEP for strongly ergodic Markov chains shown above has the same practicality as in the homogeneous case. Sequences of symbols, as a whole, generated by strongly ergodic inhomogeneous Markov information source can thus be efficiently transmitted, as the above results imply, for exactly the same reasons as in the homogeneous case. The reason for this is that the set of finite sequences of symbols (states) generated either by a regular Markov information source or a strongly ergodic inhomogeneous Markov information source can be partitioned into a highly likely group and an unlikely group. To efficiently transmit information, one encodes the highly likely sequences with as few symbols as possible and the unlikely sequences with arbitrary encoding. From this perspective, the mathematical differences between the homogeneous and
inhomogeneous cases are irrelevant. No further benefits derive from demonstrating an AEP for inhomogeneous Markov chains. In addition, the fact that the above results also ignore the chief feature of strongly ergodic inhomogeneous Markov chains, *i.e.*, convergence in distribution, is perhaps why these results have not hitherto been developed — a lack of motivation. Nonetheless, the theory developed below shows how the AEP for strongly ergodic inhomogeneous Markov information sources, developed above, illuminates some theoretical considerations in the finite-time performance of the SA algorithm. To show this, some method of incorporating the concept of convergence in distribution into the AEP for strongly ergodic inhomogeneous Markov information sources must be devised. This can be done by dividing each sequence generated by a strongly ergodic inhomogeneous Markov chain into a series of subsequences. This idea is illustrated in Figures 1 and 2.

\textbf{Examples of Optimal and Suboptimal States in Two Annealing Sequences}

\textbf{Time} \rightarrow

\begin{align*}
A & \quad \text{TIIIIIOIOIOIOOOIOOOIOOOOIOOOO} \\
B & \quad \text{OOOOIOOOOIOOOOIOOOIOOOIIOIIII}
\end{align*}

\textbf{Figure 1}

Let the sequences $A$ and $B$ indicate sequences of states generated by an SA experiment, where the symbol O corresponds to optimal states and the symbol I corresponds to suboptimal states. Sequence $A$ could therefore be considered as an example of a typical annealing sequence since the frequency of the states O increases over the course of the experiment. Sequence $B$ has exactly the same distribution of states (the sequence of states has merely been reversed), yet this sequence should not be considered as a candidate for membership in the class of typical annealing sequences (even though in the homogeneous case, if sequence $A$ is typical, then sequence $B$ would also be typical) because the frequency of optimal states \textit{decreases} over the course of the annealing sequence.
Sequence $A$ reflects the notion of convergence in distribution while sequence $B$ does not. The only difference in the two sequences is the order in which the states appear. Thus, for strongly ergodic inhomogeneous Markov chains, not only is the frequency of states important in defining typical sequences, but also, where in the sequence the states occur.

To develop a reasonable criterion for defining typical annealing sequences, each such sequence can be divided into a series of subsequences. By doing this, the notion of convergence in distribution can be captured while ignoring the ordering of states within the subsequences. To illustrate, in Figure 2 note that sequence $A'$ is identical to sequence $A$ but has been divided into four successive subsequences.

<table>
<thead>
<tr>
<th>Frequency of Optimal States in Each Subsequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time -&gt;</td>
</tr>
<tr>
<td>$A'$ 1111100 1001010 0010010 00010000</td>
</tr>
<tr>
<td>2 5 6 7</td>
</tr>
<tr>
<td>$A''$ 0101110 0100101 01000100 00001000</td>
</tr>
<tr>
<td>2 5 6 7</td>
</tr>
<tr>
<td>$B'$ 00001000 01000100 01001001 0101111</td>
</tr>
<tr>
<td>7 6 5 2</td>
</tr>
<tr>
<td>$B''$ 00001000 00100010 10010010 11111000</td>
</tr>
<tr>
<td>7 6 5 2</td>
</tr>
</tbody>
</table>

Figure 2

The number below each subsequence indicates the number of optimal states in that subsequence. Notice that in sequence $A'$ these numbers increase from one subsequence to the next, hence this sequence could be considered as a typical annealing sequence. Now consider sequence $A''$, a sequence with an identical distribution of states as in sequences $A$ and $A'$. Sequence $A''$ is the same sequence as $A$ and $A'$ but has the ordering of states in each of its subsequences reversed. Notice that the frequency of optimal states in each subsequence remains the same and is therefore increasing.
from one subsequence to the next. Thus, sequence $A''$ could also be a member of the set of typical annealing sequences.

In contrast, consider sequences $B'$ and $B''$. Sequence $B'$ is identical to sequence $B$ in Figure 1 but has been divided into a series of subsequences. Since $B$ does not exhibit convergence in distribution and therefore should not be a member of the group of typical sequences, neither should sequence $B'$. The decreasing numbers of optimal states in the subsequences in $B'$ bear this out. Likewise, reversing the order of states in each subsequence in sequence $B'$, a sequence which cannot be a member of the typical group, as is done in sequence $B''$, does not change the fact that the sequence does not reflect convergence in distribution and, likewise, cannot be a member of the typical group.

Therefore, when the sequence as a whole is considered, the ordering of states, in addition to the frequency distribution, matters as sequences $A$ and $B$ illustrate. By dividing sequences up into subsequences, the ordering of the states within subsequences can be made irrelevant, as it is in the homogeneous case, insofar as whether the sequence should be considered as a typical annealing sequence. Only the frequency of states in each subsequence need be considered for the membership criterion as sequences $A', A'', B'$ and $B''$ illustrate. It is still necessary however, to consider the ordering of states when calculating probabilities, since in an inhomogeneous Markov chain, the transition probabilities change with each transition, hence at each time index.

Using the above ideas, it is now possible to characterize the criterion for membership in the set of typical sequences which exhibit convergence in distribution for strongly ergodic inhomogeneous Markov chains in terms of the frequencies of states. This set will be denoted by $\Gamma_1$ to distinguish it from $\Omega_1$ (it will be shown later on that these two sets are not identical). Using an analogy from the homogeneous case, the membership criterion for $\Gamma_1$ will be based on a modified version of (2.8) that reflects the inhomogeneity of the process. Since (5) is written in terms of the numbers of $ij$ pairs between time indices $k$ and $k+r-1$, i.e., $m_{ij}(k, k+r-1)$, the membership criterion
should require that the number of $ij$ pairs is close to the expected value of $m_{ij}(k, k+r-1)$. Noting that 
\[ \bar{E}\{m_{ij}(k, k+r-1)\} = v_i^{[k]} p_{ij}^{[k]} + v_i^{[k+1]} p_{ij}^{[k+1]} + \ldots + v_i^{[k+r-1]} p_{ij}^{[k+r-1]} \], the criterion for inclusion of sequences in $\Gamma_1$ can be defined in terms of the number of $ij$ pairs in a subsequence of transitions.

**Definition:**

The set $\Gamma_1$ is defined as the set of all sequences $\Sigma_n$, where for some fixed $\delta > 0$, each subsequence $\tau$ of length $r_\tau$ of sequence $\Sigma_n$, starting at time index $k$, $1 \leq k < n-1$, and $\forall i, j, 1 \leq i \leq s, 1 \leq j \leq s$, the following inequality holds:

\[
\left| \frac{m_{ij}(k, k+r_\tau-1)}{r_\tau} - \bar{E}\{m_{ij}(k, k+r_\tau-1)\} \right| \leq \delta. \tag{6}
\]

This means that the set $\Gamma_1$ consists of sequences in which each of its subsequences $\tau$ of length $r_\tau$ has state frequencies consistent with (6), that is, has a distribution of $ij$ pairs within $r_\tau \delta$ of their expected values. In SA, the requirement that all subsequences be consistent with (6) can be relaxed to requiring that only the last subsequence be consistent with (6). This is because the finite-time performance of SA will be based only on the distribution of the final state in an annealing sequence. To keep the analysis tractable, subsequences must be nonoverlapping. Thus, each sequence is the concatenation of all of its subsequences. Since the probability measure is central to this analysis, to simplify notation, a subsequence can be encoded as a sequence of transition probabilities. No loss of information occurs from this method of encoding the states of a subsequence as it is always possible to reconstruct the sequence of states by noting the sequence of transition probabilities (this is analogous to the definition of a Markov information source where every arc emanating from a state corresponds to a symbol in an alphabet distinct from all other symbols associated with the other arcs emanating from that state. Such an information source is said to be unifilar. In SA, the configuration space can be characterized as a unifilar information source by properly assigning
symbols to each arc. For a complete description, see Ash [1965 pp.187] or Shannon [1948, 1963 pp.45]). Thus, writing \( \Pr\{\Sigma_n\} \) in terms of the transition probabilities of each subsequence.

\[
\Pr\{\Sigma_n\} = \frac{v_{\sigma_0}^{[0]} p_{\sigma_0 \sigma_1} p_{\sigma_1 \sigma_2} \ldots p_{\sigma_{n-1} \sigma_n} \times p_{\sigma_{n+1} \sigma_{n+1}} \times \ldots \times p_{\sigma_{R_n-1} \sigma_{R_n-1}} x \ldots \times \frac{p_{\sigma_{R_n-1} \sigma_{R_n-1}}}{p_{\sigma_{R_n-1} \sigma_{R_n-1}}}}{\text{subsequence 1 of length } n} \times p_{\sigma_{n+1} \sigma_{n+1}} \times \ldots \times \frac{p_{\sigma_{R_n-1} \sigma_{R_n-1}}}{p_{\sigma_{R_n-1} \sigma_{R_n-1}}}
\]

(7)

where \( v_{\sigma_0}^{[0]} \) is the probability of the initial state \( \sigma_0 \) and \( R_n = \sum_{k=1}^{n} k \). The following definition is needed.

Definition:

\[
\Pr\{T_\tau\} = \frac{p_{\sigma_{t-1} \sigma_{t-1}} \times \ldots \times p_{\sigma_{R_\tau-1} \sigma_{R_\tau-1}}}{\text{subsequence of transitions of a given sequence } \Sigma_m \text{ where } R_\tau \text{ is defined above.}}
\]

Using this definition, the probability measure associated with the sequence \( \Sigma_n \) can be written in terms of the probability measures of each subsequence of transitions. Thus,

\[
\Pr\{\Sigma_n\} = v_{\sigma_0}^{[0]} \prod_{\tau} \Pr\{T_\tau\}
\]

(8)

Without loss of generality, consider the probability measure of one subsequence of transitions, \( \Pr\{T_\tau\} \), in a sequence \( \Sigma_n \in \Gamma_1 \) where \( R_{n-1} = k \). Thus, subsequence \( \tau \) has transitions from time index \( k \) to time index \( k + r_\tau - 1 \). The probability measure of a subsequence of transitions can therefore be related to the number of \( ij \) transitions by the following equation.

\[
\Pr\{T_\tau\} = \prod_{i=1}^{k} \prod_{j=1}^{r_\tau} p_{ij} (k, k + r_\tau - 1) m_{ij}^{(k + r_\tau - 1)}
\]

(9)

Since \( \Sigma_n \in \Gamma_1 \) from (6),

\[
m_{ij}^{(k, k + r_\tau - 1)} = \hat{E}\{m_{ij} (k, k + r_\tau - 1)\} + r_\tau \delta_{ij}^{(k, k + r_\tau - 1)}
\]

\[
= v_i^{[k]} \sum_{j=1}^{l} p_{ij}^{[k]} + \ldots + v_i^{[k + r_\tau - 1]} p_{ij}^{[k + r_\tau - 1]} + r_\tau \delta_{ij}^{(k, k + r_\tau - 1)}
\]

(10)

where \( |\delta_{ij}^{(m, m + r_\tau - 1)}| \leq 1 \). Substituting (10) into (9), taking the natural log of both sides, and multiplying by \(-1 \) gives
\[- \ln \Pr \{ T_\tau \} = - \sum_{i=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{S}} \left( v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m\tau, r-1]} p_{ij}^{[m\tau, r-1]} + r \theta_{ij}^{\tau} + \delta_{ij} \right) \ln \hat{p}_{ij}^{\tau} (m, m + r, -1) \]  

(11)

Several points should be noted to identify the similarities and differences between the homogeneous case and the inhomogeneous case. First, the expression $v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m\tau, r-1]} p_{ij}^{[m\tau, r-1]}$ in (11) corresponds to $\pi_i p_{ij}$ in the homogeneous case. In fact, they are the same when $v_i = \pi_i$ for all $i$ and $p_{ij}^{[k]} = p_{ij}$ for all $k$. Second, in the homogeneous case, the terms $p_{ij} \ln p_{ij}$ can be summed over all $i, j$ to obtain a scalar $H_{mc}$, which is the same for all sequences. Differences in the expected frequency of states in the sequence and the actual number of transitions are accounted for by $\theta_{ij}$. In the inhomogeneous case, the terms $\left( v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m\tau, r-1]} p_{ij}^{[m\tau, r-1]} \right) \ln \hat{p}_{ij}^{\tau} (m, m + r, -1)$ must be summed over all $i, j$, but $\hat{p}_{ij}$ depends on the position in the sequence of the $ij$ transitions. Thus, $\hat{p}_{ij}$ is not a fixed value and instead depends on where in the sequence the $ij$ transitions occur, hence, depends on the ordering of the states in the sequence. Nor is $\hat{p}_{ij}$ a probability measure since $\sum_j \hat{p}_{ij} \neq 1$. Therefore the $\hat{p}_{ij}$ are not elements of a stochastic matrix.

To summarize these differences, in the homogenous case, the value of $m_{ij}$ affects the probability measure of a sequence while the ordering of states does not. In the inhomogeneous case, both $m_{ij}$ and the ordering of states in the sequence affect the probability measure of that sequence. Although the nonstationarity of an inhomogeneous Markov chain makes the relationship between the probability measure of a sequence and an entropy measure more complex, such a relationship exists. In order to establish this relationship, the following results are needed in order to relate the $\hat{p}_{ij}$ to appropriate probability measures.

**Lemma 1:**

Given a Markov chain with transition probabilities in (1.2) and a cooling schedule in (1.3) and time indices $k$ and $k+r$, then for all $ij$ transitions, there exists a $t_y \in [t_{k+r}, t_k]$ such that $\hat{p}_{ij} (k, k+r) = p_{ij} (t_y)$.

**Proof:** See Appendix A. ■
Lemma 1 states that as a consequence of the monotonicity of the transition probabilities in SA, the value of the geometric mean of the transition probabilities between two temperature values is equal to the value of the transition probability at some particular temperature value that lies between the extreme temperature values.

**Lemma 2:**
Under the assumptions stated in Lemma 1, for any fixed subsequence \( \tau \) of transition probabilities of length \( r_T \), there exists a time index \( m^* \) such that for all \( m > m^* \), there exist values \( t_i > 0 \), such that

\[
\sum_{i=1}^{s} \sum_{j=1}^{s} m_{ij}(m,m+r_T-1) \ln p_{ij}(t_j) = \sum_{i=1}^{s} \sum_{j=1}^{s} m_{ij}(m,m+r_T-1) \ln p_{ij}(t_i).
\]

**Proof:** See Appendix A. 

Lemma 2 states that for each state \( i \), the summation of the logarithms of the transition probabilities to neighboring states \( j \) equals the sum of the logarithms of the transition probabilities to neighboring states \( j \) at a single temperature value \( t_i \).

From Lemmas 1 and 2, \( p_{ij}(t_i) \) can be substituted for \( \hat{p}_{ij}(m,m+r_T-1) \) in (11) to give

\[
-\ln \Pr(T_r) = -\sum_{i=1}^{s} \sum_{j=1}^{s} \left[ v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m+r_T-1]} p_{ij}^{[m+r_T-1]} + r_T \delta_{ij}(m,m+r_T-1) \ln p_{ij}(t_i) \right] \ln p_{ij}(t_j)
\]

(12)

\[
= -\sum_{i=1}^{s} \sum_{j=1}^{s} \left[ v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m+r_T-1]} p_{ij}^{[m+r_T-1]} \ln p_{ij}(t_i) + \right.
\]

\[
-\sum_{i=1}^{s} \sum_{j=1}^{s} r_T \delta_{ij}(m,m+r_T-1) \ln p_{ij}(t_i)
\]

(13)

(14)

The following definitions are needed to show how the entropy measure relates to (12).

**Definitions:**

\[
g_i^{[k]} = -\lambda \sum_{j=1}^{s} p_{ij}^{[k]} \ln p_{ij}(t_i), \text{ where } \lambda \text{ is some positive constant.}
\]

\[
g(k) = \{g_1^{[k]}, g_2^{[k]}, \ldots, g_s^{[k]}\}.
\]
\[ G(k) = -\lambda \sum_{i=1}^{s} \sum_{j=1}^{r} v_{ij} \ln p_{ij}(t_i) + v(k) \cdot g(k), \] the dot product of the vector \( g(k) \) and the state distribution vector \( v(k) \).

\[ G^*(m,m+r-1) \equiv \sum_{i=m}^{m+r-1} g_i \cdot v(i) = \sum_{i=m}^{m+r-1} G(i) = \sum_{i=m}^{m+r-1} G(k). \] Note the similarity of \( G(k) \) and \( G^*(m,m+r-1) \) to \( H(k) \) and \( H^*(m,m+r-1) \), respectively.

These definitions lead to the following results for strongly ergodic inhomogeneous Markov chains.

**Theorem 2:**

Given a strongly ergodic inhomogeneous Markov chain in which the diagonal and non-diagonal elements of the associated transition matrix are monotone with respect to time index \( k \) (see (1.2) and (1.3)), for each subsequence \( \tau \) of length \( r_\tau \) sufficiently large and time index \( k \) sufficiently large which satisfies (6), then for all \( \eta_\tau > 0 \) there exists a \( G^*(m,m+r_\tau-1) > 0 \) such that

\[
\left| -\ln \Pr\{T_\tau\} - \frac{G^*(m,m+r_\tau-1)}{r_\tau} \right| \leq \eta_\tau \quad (15)
\]

**Proof:** See Appendix A. \( \blacksquare \)

Since \( G^*(m,m+r_\tau-1) > 0 \) refers to the \( \tau^{th} \) subsequence of length \( r_\tau \) to simplify the notation, define \( G^*_\tau = G^*(m,m+r_\tau-1) \). In a similar manner, \( \theta_j(\tau) = \theta_j(m,m+r_\tau-1) \). Considering the sequence as a whole and recalling that each of its subsequences is disjoint, the probability of the sequence is equal to the product of the probabilities of each of its subsequences (see (8)). This leads to the following corollary to Theorem 2:

**Corollary:**

If for each subsequence in \( \Sigma_n \), the bound (6) is satisfied, then there exists an \( \eta_\tau > 0 \), an \( N > 0 \) and \( \sum_\tau G^*_\tau > 0 \) such that

\[
\left| -\ln \Pr\{\Sigma_n\} - \frac{\sum_\tau G^*_\tau}{N} \right| \leq \eta_\tau \quad (16)
\]

**Proof:** See Appendix A. \( \blacksquare \)
Note the similarity of (15) and (16) to (2.7). In (2.7), the homogeneous case contains the quantity $H_{mc}$ which is constant for all sequences, rather than $\sum G_\tau^*/N$, as in (16). In the inhomogeneous case, $G_\tau^*$ depends on the number and ordering of states in the subsequence, since the order affects the geometric mean of the transition probabilities and, from Lemmas 1 and 2, the value of $p_\ell(t_i)$ which affects the values of $\xi(k)$ and $G(k)$, hence $G_\tau^*$.

Now consider the $\tau^{th}$ subsequence of two $\Gamma_1$ sequences of length $r_\tau$ with identical frequencies of states but with a different ordering of these states. Because of the inhomogeneity of the Markov process, the probability values of these subsequences will be different. This means that the two values of $G_\tau^*$ must also be different for the inequality in (15) to hold for both subsequences. The most significant observation to be made, however, is that the values of $G_\tau^*$ for all $\tau^{th}$ subsequences in $\Gamma_1$ sequences have the same lower bound, which is independent of all sequences and depends only on the transition probabilities (1.2), the cooling schedule (1.3) and the initial state distribution vector. This result is stated in the following theorem.

**Theorem 3:**

Given a strongly ergodic inhomogeneous Markov chain defined by the transition matrix in (1.2) and a cooling schedule given in (1.3), for which Theorem 2 holds, then for all possible sequences of finite length and for any interval of time indices $m$ and $m + r_\tau - 1$, $G^*(m,m+r_\tau-1) \geq H^*(m,m+r_\tau-1)$.

**Proof:** See Appendix A. ■

The differences between homogeneous Markov information sources and strongly ergodic inhomogeneous Markov information sources are now apparent. In the homogeneous case, when $m_{ij}$ is close to $n p_{ij} - \ln \Pr(S_n)/n$ is close to $H_{mc}$ and vice versa. This symmetry is not present in the inhomogeneous case. Thus, if $-\ln \Pr(S_n)/n$ is close to some value $G^*(m,m+r_\tau-1)$, it implies that $m_{ij}(m,m+r_\tau-1)$ is close to $E[m_{ij}(m,m+r_\tau-1)]$ only when the values of the $G^*(m,m+r_\tau-1)$ are suitably bounded. This asymmetry is due to the inhomogeneity of the Markov chain.
The differences between the homogeneous and inhomogeneous cases therefore lead to two definitions of typical sequences, both of which are necessary in relating entropy measures to the performance of SA. The first set, defined as set $\Omega_1$ and based on Theorem 1 and the concept of convergence in probability, allows us to make a statement regarding the total probability of typical sequences and relate it to an entropy measure (see (4)). The second set, defined as set $\Gamma_1$, is based on the frequency of $ij$ transitions (see (6)). Theorems in Chapter 4 show how this set allows us to make a statement about the distribution of the final state in an annealing experiment. By using the two sets together, it is possible to make a probabilistic statement about the final state in an annealing experiment. This is possible because the two sets become the same set asymptotically due, in part, to the fact that the values of $G^*(m, m+r_e-1)$ approach $H^*(m, m+r_e-1)$ from above. This fact is stated in the following corollary.

**Corollary:**
Given the inhomogeneous Markov chain described above, for a given $r_e > 0$,

$$\lim_{m \to \infty} \frac{G^*(m, m+r_e-1)}{r_e} = \lim_{m \to \infty} \frac{H^*(m, m+r_e-1)}{r_e}.$$

**Proof:** See Appendix A. ■

If the number of typical annealing sequences (i.e., sequences which exhibit convergence in distribution) increases, then it is increasingly likely that a randomly generated sequence will be in the typical group and therefore have a higher number of optimal or near optimal states near the end of the annealing sequence. The question then arises on how to increase the number of typical annealing sequences. The following theorem shows that the size of the group which exhibits convergence in distribution grows exponentially with the value of $H^*(m,n-2)$. 
Theorem 4:
For any $\delta' > 0$, there exist time indices $m$ and $n$ such that the size of a subset of the group which exhibits convergence in distribution, defined as $\Gamma'_1$, where $\Gamma'_1 \subset \Gamma_1$, can be estimated by

$$|\Gamma'_1| \geq e^{H^*(m,n-2)+\delta'}.$$ 

Proof: See Appendix A. ■

The next section discusses two possible approaches which can lead to increases in $H^*(m,n-2)$, thereby improving the finite-time performance of SA.

3.2. The Entropy of Annealing Experiments

Now that a theoretical foundation between the entropy of Markov chains and SA has been established, the question arises as to how this entropy measure may affect the performance of SA. Several possibilities exist for changing the value of this entropy measure and, hence, possibly changing the performance of SA. The first method involves modifying the neighborhood structure of the configuration space. This has been explored by several researchers in trying to assess the impact of the neighborhood size on the performance of SA and have yielded mixed results (see for example Goldstein and Waterman [1988], Yao [1991] and Cheh et al. [1991]). Modifying the neighborhood structure changes the value of $H^*(m,n-2)$ and therefore changes the finite-time performance of SA, but as indicated below, increasing the neighborhood size does not guarantee increases in the value of $H^*(m,n-2)$. Hence, the effect of neighborhood changes on the value of $H^*(m,n-2)$ can explain these mixed results. A second method, which has not been previously considered, attempts to increase $H^*(m,n-2)$ by transforming the original combinatorial problem, thus changing the size of the configuration space and, hence, the dimension of the transition matrices.
3.2.1. The Effect of Neighborhood Structure on Entropy

Before the effect of changes in entropy on SA’s performance are explored, some discussion of how the neighborhood size affects the entropy measure and the number of typical sequences is given. The neighborhood size directly affects the Markov transition matrices and therefore directly affects the entropy measure since for any particular solution that SA happens to be in (the current state), the more neighbors there are, intuitively, the more uncertainty exists as to what state SA will be in during the next iteration. But another implication of changes in neighborhood size is that certain sequences of states become possible when before they may have been impossible, that is, a sequence may have a probability measure of 0. For example, if states $S_i$ and $S_j$ are not neighbors, then any sequence of states with a pair $S_iS_j$ has probability measure 0. If by adding to the neighborhood size $p_{ij}$ becomes greater than 0, then more possible sequences of states are possible. Thus, the neighborhood structure directly affects the entropy of the associated Markov chain, the number of possible sequences and, by implication, the number of typical sequences. These two concepts are expressed in the following theorems.

**Theorem 5:**
Given a transition matrix defined by (1.2) in which each state has the same number of neighbors, $N$, then for all time indices $k$, $H(k) \leq \ln(N+1)$.

**Proof:** See Appendix A. ■

**Theorem 6:**
Let the number of possible sequences of length $n$, that is, sequences where $Pr[\Sigma_n] > 0$ be denoted by $N(n)$. Then, given a transition matrix defined by (1.2) in which each state has the same number of neighbors, $N$, there exists a positive number $K$ such that $N(n) \leq K(N+1)^n$.

**Proof:** See Appendix A. ■

Note that since the number of typical sequences must be less than or equal to the number of possible sequences, from Theorem 4, the following inequalities hold:

$$e^{H(\{s_{\Sigma_N}\}+S_0)} \leq \sum_i |\Gamma_i| \leq K(N+1)^n = e^{n\ln(N+1)+\ln K}.$$
Theorem 5 indicates that the logarithm of the neighborhood size is an upper bound on the entropy of the inhomogeneous Markov chain at all time indices. Theorem 6 indicates that the number of possible sequences grows exponentially in terms of the neighborhood size. Thus, the neighborhood size directly affects the entropy of sequences as well as the number of possible and typical sequences there are. How the neighborhood size affects the entropy is somewhat problematic. Since the upper bound in Theorem 5 is a constant, the interplay of the factors that affect the entropy measure are ignored. These factors, however, play an essential role in determining when and by how much the neighborhood size affects the entropy measure. The following example and analysis, therefore, explains this interplay of the important factors affecting the entropy measure when the neighborhood is increased and illustrates how increases in entropy can be achieved.

One way to increase the value of $H^*_{m,n-2}$ is suggested by its definition. Since $H^*_{m,n-2}$ is based on $w(k)$ and $H(k)$, its value can be increased by increasing the value of $h_i^{[k]}$ for those states $i$ which become more heavily weighted by the $v_i^{[k]}$, the components of $v(k)$, as the time index $k$ increases from $m$ to $n-2$. Since annealing sequences converge in distribution, and if $n$ is sufficiently large, then the sum of the values of $v_i^{[k]}$ associated with the globally optimal states converge to unity as the time index increases and the temperature monotonically decreases to zero. By the definition of $H^*_{m,n-2}$, if $h_i^{[k]}$ associated with the optimal states is increased, then $H^*_{m,n-2}$ must also increase.

It is not practical, however, to increase $h_i^{[k]}$ for just the globally optimal states since the globally optimal states are not known in advance in any real-world implementation of SA. If they were, the application of the SA algorithm would be unnecessary. Consequently, in order to increase the value of $h_i^{[k]}$ for the optimal states, it is necessary to increase this value for all states. Since $h_i^{[k]}$ measures the uncertainty associated with moving from state $i$ to its successor state, increasing the number of possible successor states, i.e., increasing the number of neighbors of state $i$, may increase
the entropy $h_i^{[x]}$ associated with each state $i$. Thus, increasing the neighborhood size of all states may increase $H^*(m,n-2)$ and hence, improve the finite-time performance of SA.

It is important to realize however that increasing the number of neighbors of any particular state $i$ does not guarantee increases in $h_i^{[x]}$. This is illustrated in the following table, which shows the values of the $\Delta f^*$.s of the neighbors to state $i$ with neighborhood sizes $N$ equal to 2, 4, and 6 along with the corresponding entropy values $h_i^{[x]}$ at temperature value $T_k = 1$.

<table>
<thead>
<tr>
<th>$\Delta f^*$.s</th>
<th>$N$</th>
<th>$h_i^{[x]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1, 1}</td>
<td>2</td>
<td>0.9128</td>
</tr>
<tr>
<td>{1, 1, 2, 2}</td>
<td>4</td>
<td>0.8849</td>
</tr>
<tr>
<td>{1, 1, 0}</td>
<td>4</td>
<td>1.4961</td>
</tr>
<tr>
<td>{1, 1, 2, 1.5, 1.5}</td>
<td>6</td>
<td>0.9110</td>
</tr>
</tbody>
</table>

Table 1

Table 1 shows that if the neighborhood size increases, it does not necessarily follow that $h_i^{[x]}$ also increases. It appears that this entropy value depends on the interplay of a number of factors, such as the particular values of the $\Delta f^*$.s of the new neighbors as well as the number of new neighbors.

In general, the relationship between neighborhood size and entropy is complicated and depends on how the various solutions are connected. Nonetheless, it is still possible to characterize this relationship in a general way. Consider a solution in an SA experiment with $N$ neighbors. This corresponds to a row in the Markov transition matrix associated with SA in which $N+1$ elements are greater than 0, corresponding to the $N+1$ transition probabilities composed of the $N$ neighbors plus the diagonal element of the transition matrix. When an additional neighbor is added, the transition probabilities of both the original neighbors and the diagonal element are affected. Thus, the effect on the entropy $h_i$ of adding a new neighbor to solution $i$ can be separated into three components: the
first, comprised of the original neighbors \(i\) to solution \(i\), defined as \(N(i)\). The contribution to the entropy measure before the addition of the new neighbor is therefore

\[
- \sum_{j \in N(i)} p_{ij}^{[1]} \ln p_{ij}^{[1]} = \sum_{j \in N(i) \setminus \{i\}} \frac{1}{N} e^{-N e^{-i \frac{N_j}{i} + \ln N}}
\]

where the transition probabilities are defined in (1.2). From this expression, it is easy to see that the effect of adding a new neighbor to the existing neighborhood \(N(i)\) merely increases the number \(N\) to \(N+1\). Therefore, the contribution of the original neighborhood to the entropy measure after the addition of the new neighbor is:

**Part 1:**

\[
- \sum_{j \in N(i) \setminus \{i\}} p_{ij}^{[1]} \ln p_{ij}^{[1]} = \sum_{j \in N(i) \setminus \{i\}} \frac{1}{N+1} e^{-N e^{-i \frac{N_j}{i} + \ln (N+1)}}
\]

Recall from (1.2) that the maximum value of non-diagonal elements is \(1/N\), where this value is achieved when the neighboring objective function values are less than or equal to that of the current solution. Thus, the contribution from part 1 almost always decreases with increasing \(N\), since \(1/N \to 0\) as \(N \to \infty\) and \(N/N \to 0\) as \(N \to \infty\), except in the case of \(N\) increasing from 2 to 3 when solution \(i\) is a local maximum for \(N = 2\) and 3. Furthermore, the rate of this decrease is greatest for smaller values for \(N\).

The second part is comprised of the diagonal element, which is also defined in (1.2) and will be denoted here as \(d\). From (1.2) the value of the diagonal element will change with the addition of a new neighbor. Let \(d_{N(i)}\) be the value of the diagonal element of the transition matrix based on the original neighborhood, that is, before the addition of the new neighbor. Then the value of the diagonal element after a new solution has been included in the neighborhood, that is, after the addition of the new neighbor, is given by

\[
d = d_{N(i)} - p_{ij}^{[1]} - \delta_{N(i)},
\]

where \(p_{ij}^{[1]}\) is the transition probability associated with the new neighbor \(j^*\) and

\[
\delta_{N(i)} = \left(\frac{1}{N_{\text{len}}} - \frac{1}{N}\right) \sum_{j \in N(i)} e^{-N e^{-i \frac{N_j}{i} + \ln N}}.
\]
the change in the transition probabilities of the original neighbors caused by the increase in the neighborhood size. Consequently, the contribution to the entropy measure of the diagonal element after the addition of the new neighbor is:

$$\text{Part 2: } - d \ln d = \left( d_{N(i)} - p_{q^*}^{[k]} - \delta_{N(i)} \right) \ln \left( d_{N(i)} - p_{q^*}^{[k]} - \delta_{N(i)} \right).$$

Finally, the contribution to the entropy measure of the new neighbor is:

$$\text{Part 3: } - p_{j^*}^{[k]} \ln p_{j^*}^{[k]} = \frac{1}{N+1} e^{-\frac{N}{i}} \left[ \frac{N}{i} + \ln(N+1) \right].$$

Parts 1, 2, and 3 are clearly interdependent—the larger $p_{j^*}^{[k]}$ is, the smaller $d$ is, and vice versa. Their effects on the entropy measure are therefore more complicated than in Part 1. Nonetheless, it is possible to characterize these effects by examining how the function $f(p) = -p \ln p$ behaves for $0 \leq p \leq 1$. This function is illustrated in the graph below:

**Value of $p \ln p$**

![Graph showing the function $p \ln p$]

Notice that this function is concave and achieves its maximum value at $p = 1/e = 0.3678$. Consequently, if $d_{N(i)}$ is large, that is, close to 1, and the addition of solution $j^*$ to the neighborhood reduces the value of the diagonal element by a small amount to some value greater than $1/e$, then the diagonal element will increase its contribution to the entropy measure.
Since the new neighbor is a non-diagonal element, its upper bound is $1/N$ and achieves this value when its objective function value is less than or equal to that of solution $i$. Thus, $d_{N(i)}$ will tend to be large when solution $i$ is a local minimum because the non-diagonal element will be smaller than $1/N$. In addition, the value of $d_{N(i)}$ will tend to be large when the original neighborhood $N(i)$ is small, since the number of non-diagonal elements will be small.

In the case of the other extremes, such as when $N(i)$ is large and solution $i$ is not a local minimum, the diagonal element of the transition matrix corresponding to solution $i$ will be small and the addition of a new neighbor will further reduce the value of $d$. In this case, the value of $d_{N(i)}$ will tend to be small, less than 0.36, and further decreases will decrease the non-diagonal element's contribution to the entropy measure.

Finally, the contribution of part 3 will always add to the entropy measure of solution $i$, since it is a completely new component to the neighborhood. Whether this contribution is large or small will depend essentially on two factors: the neighborhood size $N$, and the value of $\Delta f_{N(i)}^+$. If the neighborhood size is large, then the maximum value of $p_{y^k}^{[1]}$ is relatively small and, consequently, solution $j^*$ will contribute only a small amount to the total entropy measure. Likewise, if $\Delta f_{N(i)}^+$ is large, such as when a steep uphill move from solution $i$ to $j^*$ exists, then $p_{y^k}^{[1]}$ will again be small and the contribution of $j^*$ to the entropy will be small (notice in part 3 that the exponential term with $\Delta f_{N(i)}^+$ dominates the effect of $\Delta f_{N(i)}^+/|t|$). On the other hand, if $N$ is small, say 2 or 3, and the value of $\Delta f_{N(i)}^+$ is small, such as when there is a fairly shallow uphill move from solution $i$ to $j^*$, then the contribution of part 3 will tend to be greater.

Thus, these three parts each contribute varying amounts to the total entropy of solution $i$ and can be characterized by dependence on the neighborhood size, and the values of the $\Delta f_{N(i)}^+$'s. The following table summarizes these characterizations.
Effect on Entropy Contribution
Caused by the Addition of a New Neighbor

<table>
<thead>
<tr>
<th>Part</th>
<th>Increase $h_i$?</th>
<th>Decrease $h_i$?</th>
<th>Increase $h_{i\cdots}$</th>
<th>Decrease $h_{i\cdots}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>special case</td>
<td>almost always</td>
<td>if $N(i) = 2.$</td>
<td>if $N$ becomes large.</td>
</tr>
<tr>
<td>2</td>
<td>possible</td>
<td>possible</td>
<td>if $N$ is small and $i$ is a local min.</td>
<td>if $N$ is large.</td>
</tr>
<tr>
<td>3</td>
<td>always</td>
<td>never</td>
<td>is large if $N$ is small and $\Delta f_{i}^{+}$ is small</td>
<td>small</td>
</tr>
</tbody>
</table>

Table 2

Based on this table, the greatest increases in entropy should occur when the neighborhood size is increased from a small value, say 2 to 4, as opposed to an increase from 8 to 10. This is all the more likely since the state vector element corresponding to the global optimum (a local minimum) increases with time and thus gives a greater weight to the diagonal element, part 2, when $N$ is small. Whether the addition to the entropy caused by part 3 outweighs any decreases caused by changes in parts 1 and 2 will depend on the particulars of the neighborhood. In view of the fact that increases in entropy seem to favor small values of $N$, as $N$ gets larger, the likelihood of increases in entropy should become less predictable. As will be shown in Chapter 6, these characterizations seem to hold in general.

3.2.2. Polynomial Transformations

Another method which can change the configuration space, and hence the transition matrix, is by a transformation of the problem. If possible, this is done in such a way that the original problem and the transformed problem have the same optimal solution. Thus, running SA on either problem can produce similar results. PLS transformations are a means of relating two combinatorial optimization problems such that the local optima in both problems are matched (Johnson et al. [1988]). This allows for the possibility of solving an optimization problem by solving the optimization problem it transforms to.
Transformations, however, tend to increase the size of a problem and, hence, degrade the ability of practitioners to solve them. This is because heuristic approaches to problem solving usually suffer when the solution space increases. Thus, transformation algorithms are not generally implemented for any practical purpose. Rather, they are used to prove the class membership of a particular type of optimization problem and therefore serve a theoretical purpose. In the context of information theory and the foregoing theorems, however, transforming a problem instance of one type of problem into a problem instance of another type of problem, will allow us to test these theories because the transformed problem will likely have a different value of $H^*(m,n-2)$. This is because for some transformations, the number of states as well as the number of neighbors of each state changes.

3.3. **Summary**

The concept of entropy and the AEP in information theory has been utilized to show how the finite-time performance of SA can be related to the entropy of inhomogeneous Markov chains of the type used to model SA. The finite-time performance of SA can be measured by the probabilities of occurrence of annealing sequences of fixed length which the Markov chain generates. Using the AEP, it has been shown that the number of typical annealing sequences, that is, sequences which exhibit convergence in distribution, is related to these probabilities of occurrence and an entropy-like measure, $G^*(m,n-2)$, that depends on the particular sequence or subsequence in question. Furthermore, all values of $G^*(m,n-2)$ from time index $m$ to time index $n-2$ have the same lower bound of $H^*(m,n-2)$ that depends only on the transition matrices, the cooling schedule, and the initial state distribution vector of the inhomogeneous Markov chain. It has been shown that increasing the value of $H^*(m,n-2)$ by modifying the transition matrices of the Markov chain can increase the probability that the inhomogeneous Markov chain randomly generates a typical annealing sequence and, hence, the probability of generating sequences with globally optimal states as their final state.
Two methods which can alter the value of $H^*(m,n-2)$ have been described. The first, modifying the neighborhood structure, shows that this method has the potential to increase the value of $H^*(m,n-2)$ and, hence, improve the finite-time performance of SA, though is not guaranteed to do so. This is consistent with the experimental results in Goldstein and Waterman [1988]. They demonstrate that too small or too large a neighborhood size can degrade performance of SA, while certain neighborhood sizes can improve the finite-time performance of SA. The foregoing analysis indicates that the reason for this is that some neighborhood sizes will, depending on the problem instance, be associated with lower values of $H^*(m,n-2)$ while other neighborhood sizes will be associated with larger values of $H^*(m,n-2)$. The second method, based on the theory of PLS transformations (polynomial transformations of NP-Hard problems), provides another avenue for modifying the value of $H^*(m,n-2)$ and therefore modifying the finite-time performance of SA.
CHAPTER 4

THE PERFORMANCE OF SIMULATED ANNEALING

4.1. MEASURING PERFORMANCE

In demonstrating a relationship between the performance of SA and the entropy measure described in Chapter 3, the question arises as to what is an appropriate performance measure for SA. Such a performance measure should have both practical and theoretical value in relation to the entropy measure. As stated earlier, the purpose here is not to show the best method of implementing SA, but rather to identify a fundamental relationship between the finite-time performance of SA and this entropy measure. Once this is done, this knowledge can then be used to further improve on implementation strategies. In considering such performance measures, it is useful to examine what implementation strategies have been previously used in evaluating the performance of SA.

A typical method of implementing SA is to keep track of the best solution obtained during the simulation. This means that during the sequence of iterations in running an SA experiment, the best solution is updated whenever the current solution improves upon the best solution obtained up to that point. Since this strategy requires very little additional memory or CPU time, it is one of the most practical. Such a strategy can easily be implemented and superimposed upon any other reasonable implementation of SA. The standard implementation of SA, however, does not incorporate this point.

Another method of implementing SA uses taboo search (Glover [1989]). In this method, neighbors of the current solution are chosen only if they have not been previously chosen. Again, this method can be superimposed on the standard SA implementation. Since both these methods can be superimposed on the standard implementation, the question arises as to how good the performance of the standard implementation is. The theory behind the standard implementation states that the SA algorithm converges in probability to the globally optimal states as the time index \( n \to \infty \) given the
cooling schedule described in Mitra et al. [1986]. Thus, the question of how fast this convergence occurs would be a useful measure of performance. Alternatively, it seems reasonable that for any finite-time run of SA, there should be a high probability that the final state is a globally optimal one when the time index \( n \) is sufficiently large. In Chapter 3, we allude to the fact that when there is a high entropy measure, the probability that the final state is a globally optimal one should be high. A reasonable approach toward demonstrating this would therefore be to run the SA algorithm many times, using the same initial conditions and cooling schedule, and compute the average value of the final state. If it can be shown that the expected value of the last state is close to the globally optimal solution when the entropy measure is high, then we can have a high confidence that, under such circumstances, any single run of SA should produce a final state that is close to the globally optimal state, provided the variance of the final state (random variable) is low.

Before using such estimators of the global optimum based on finite-time SA experiments, it is prudent to examine relationships in the asymptotic cases. Relationships between performance and entropy are shown in the asymptotic or limiting cases of SA and are described below. By way of review, the next subsection highlights the connection between typical sequences and entropy.

### 4.2. The Convergence of SA

A relationship has been established between the number of typical sequences, their total probability of occurrence, and the entropy measure defined in Chapter 3. In essence, the larger the entropy measure, the more typical sequences there are for any fixed length set of sequences and the larger their total probability. The question of whether more typical sequences improves performance, that is, whether there is a large number of optimal states in these sequences has not been addressed. Certainly, if typical sequences are made to somehow contain more optimal states, then increasing the number and total probability of these typical sequences, such as by increasing the entropy measure, would be desirable. In Chapter 3 methods to increase this entropy measure were discussed, such as changing the neighborhood structure of the configuration space. But in changing
the neighborhood structure, the transition matrix also changes. While this change may increase or decrease the entropy measure, it also has some effect on how fast the inhomogeneous Markov chain converges. Before attempting to analyze the effect on convergence of changing the entropy measure, the issues concerning convergence in general must be reviewed. It is noteworthy, however, that the notion of convergence of SA can be expressed in an information theoretic context. The following shows how, in the limiting case, the AEP expresses the convergence of SA.

4.2.1. **SA and the AEP in the Limit**

Recall from Theorem 3.4 in Chapter 3, that the number of typical sequences is $|\Gamma_1| \geq e^{H_n \cdot m \cdot z - z}$. Also note that from Theorem 3.1 the entropy measure for a strongly ergodic Markov chain approaches some constant. In SA, this constant is equal to the logarithm of the number of optimal solutions. Consequently, the ratio of the number of typical sequences to the entire sequence space approaches zero, that is

$$\lim_{n \to \infty} \frac{|\Gamma_1|}{s^*} = 0.$$ 

Yet from the proof of Theorem 3.4, the total probability of the typical sequences is arbitrarily close to 1 (see Appendix A). But those typical sequences have an increasing number of optimal states. This is evident from that fact that the state distribution elements corresponding to globally optimal solutions $i^*$ approach the value

$$\lim_{n \to \infty} v_{i^*} = q_{i^*} = \frac{1}{s^*},$$

where $s^*$ is the number of globally optimal states. Thus,

$$\lim_{n \to \infty} \frac{\hat{E}(m_{1,i}, (m,n-2))}{n-m-2} = \lim_{n \to \infty} \frac{v_{i^*}^m p_{i^*}^{m-1} + v_{i^*}^{m-1} p_{i^*}^{m-2} + \ldots + v_{i^*}^{n-2} p_{i^*}^{n-2}}{n-m-2} = 1$$

Therefore, the fraction of optimal states in the typical sequences approaches unity. In summary then, the relative proportion of the typical sequences to the entire sequence space decreases to 0, but the
total probability of those typical sequences increases to 1 as does the proportion of optimal states in those few typical sequences. In other words, all the probability in the sequence space becomes concentrated on a limited number of sequences, each of which reflects the convergence in probability of the SA algorithm. In this way, the AEP for strongly ergodic Markov chains is consistent with the convergence of SA.

4.2.2. **The Convergence Rate of SA**

The analysis in the preceding section illustrates how the SA algorithm converges in probability, from the perspective of the AEP in which a sequence space is partitioned into typical and atypical sequences. Unfortunately, this analysis is valid only for the limiting cases where the SA algorithm continues to search the configuration space forever. Insofar as the finite-time behavior of SA is concerned, this analysis suffers from one serious flaw mentioned earlier, that the changes to the transition matrices that affect the entropy measure can also affect how fast SA converges in probability. If these two effects were independent, then it would be possible to extend the above analysis to SA experiments of finite length. For the most part, changes to the transition matrices that improve performance also increase the entropy, yet high values of entropy would, intuitively, seem to be the antithesis of convergence—increasing uncertainty should slow down convergence. Yet, no theoretical formulation that connects the entropy of Markov chains to their convergence rate (for example, a connection between the entropy of Markov chains and the eigenvalues of the transition matrices) has been developed even for the homogeneous case! In the following sections, an approach using scaling properties of SA is employed to fashion some connection between convergence of SA and entropy. Before this is done, however, some background into how fast SA converges is in order. This will help to illuminate those factors affecting convergence that also affect the entropy measure.

*Mitra et al.* [1986] provides the following estimate on the rate of convergence for SA:
\[ \|v(kr) - e^*\| \in O\left( \frac{1}{k^{\min(a,b)}} \right) \]  

where \( v(kr) \) is the state distribution vector at time index \( kr \), \( e^* \) is the state vector in which all the probability is concentrated on only the optimal solutions, and \( a \) and \( b \) are related to the cooling schedule and the configuration space of the particular problem on which SA is being applied. It is these latter factors that are of interest to us. As Mitra indicates, the factors \( a \) and \( b \) depend, in part, on the "connectivity properties of the graph underlying the Markov chain..." [Mitra pp.769]. These will be examined in some detail below.

The factor \( a \) described by Mitra is a function of several properties of the configuration space. Written in terms of the notation used here,

\[ a = \left( \frac{k}{L} \right)^{\gamma} \]  

where \( r \) is the radius of the associated graph of the configuration space, \( L \) is the maximum difference in the objective function values of any two neighboring states in the configuration space and \( \gamma \) is the constant associated with the cooling schedule described in Chapter 1 (Mitra pp. 762. Note that the factor \( 1/N \) is the minimum generation function referred to in Mitra and is used here because the generation function is the same for all of the solutions in configurations spaces where the number of neighbors are the same for all solutions.). The value for \( r \) is defined as

\[ r = \min_{1 \leq i \leq S_m} \max_{j \in S} d(i,j) \]  

where \( S_m \) is the set of solutions that are all local maxima, i.e., \( S_m = \{ i \in S | f_j \leq f_i \ \forall j \in N(i) \} \) and \( d(i,j) \) is the minimum number of edges in a path from \( i \) to \( j \) that minimizes the maximum path length. Thus, \( r \) is the minimum number of edges that must be traversed in order to get from one to solution to any other solution. The factor \( L \) is defined by

\[ L = \max_{i \in S} \max_{j \in N(i)} |f_j - f_i| \]  

The factor \( b \) is defined more simply by \( b = \delta/\gamma \) where \( \delta \) is the difference in objective function values between the least-cost and the next-to-least-cost solution. From these definitions it is easy to see that the convergence rate
depends on a variety of values, all dependent on the topology of the configuration space. These factors can be examined qualitatively to see how increases in $a$ and $b$ affect the entropy measure.

The factor $r$ is clearly affected by the neighborhood structure. In general, the more neighbors there are, the smaller the radius of the graph. This relation can be made explicit for the generic configuration spaces, introduced in Chapter 6. In such configuration spaces, the configuration space is defined by a graph in which all the nodes correspond to solutions and have the same number of neighbors as in Figure 1 which depicts eight solutions each with a neighborhood size of two.

\[ r = \left\lfloor \frac{|S| - 1}{N} \right\rfloor \]

**Figure 1**

For this particular graph, the radius $r = 4$ because traversing 4 arcs is the minimum number of arcs required to get from one node to any other node (e.g., the number of arcs between diametrically opposite nodes). If the neighborhood were increased from 2 to 4, then the radius is decreased to 2. In general, the radius of these types of graphs can be related to neighborhood size by the following:
where $|S|$ is the size of the configuration space. Thus, with an increase in the neighborhood size, there is a decrease in the radius and, therefore, an increase in $a$ which can lead to faster convergence (see (1)). The factor $a$ is also larger when $L$ is smaller which tends to happen when the configuration space is flatter, i.e., objective function values for neighboring solutions tend to be close to each other. This also corresponds to higher entropy (since there is more uncertainty when the configuration space is relatively smooth). Since we expect the entropy to increase with increases in neighborhood size (although this is not guaranteed), and decreases in $L$, increases in entropy should also correspond to faster convergence rates insofar as the factor $a$ is concerned. On the other hand, a larger value for $a$ increases the chance that $b$ will control the convergence rate. Thus, a large value for $a$ may or may not lead to faster convergence.

The factor $b$, being proportional to the difference in objective function values between the least-cost and the next-to-least-cost solution, the smaller this value is, the more likely it controls the convergence rate (see (1)). Again, smaller values should correspond to higher entropy values. This is because the transition probability is higher and is more likely to contribute a larger value to the entropy measure (see Section 3.2.1 in which the next-to-least-cost solution is the added solution caused by an increase in neighborhood size, that is, corresponds to Part 3).

In general, the larger the factors $a$ and $b$ are, the higher the convergence rate. But these factors are all affected by the topology of the configuration space and their interplay is complicated. While higher values suggest faster convergence, it is the minimum of $a$ and $b$ that determines the convergence rate. This adds a significant complication which seems to preclude a straightforward connection between entropy and the convergence rate for SA. In the next sections, an approach is described that yields a more direct connection between convergence rate of SA and entropy. Before this is done however, certain properties of the final state of typical sequences must be identified.
4.3. **The Final State—Considerations**

In this section, a performance measure is considered in more detail. The main reason for defining the typical group, $\Gamma_1$, is to enable us to make statements about the final state. These aspects will be described in the next subsection. Before that is done, some basic issues concerning the final state must be addressed.

In using the final state of an SA experiment as a measure of performance, we are, in effect, using the final state as an estimator of the globally optimal solution. Consequently, issues concerning the bias and the variance of this estimator come into play. Specifically, we are concerned with how the entropy measure may affect the bias and the variance of this estimator. In the experiments described in Chapter 6, the configuration spaces have a globally optimal value of 0.

To compute the mean squared error of $X$, the final state (random variable) in the SA, let $\theta$ equal the parameter that is being estimated. By definition, the mean squared error in terms of the bias and variance of the estimator is:

$$\text{MSE}(\theta) = E[(X - \theta)^2]$$

$$= \text{Bias}^2 + \text{Var}(X)$$

Since the global optimum that is being estimated, $\theta$, is known to be 0, then $\text{MSE}(\theta = 0) = E(X^2)$ and the Bias = $E(X)$, where $X$ is the objective function value (random variable) of the final state in the SA experiment. The methods used in calculating these statistics will be described in Chapter 5. These performance measures will then be used in the experiments described in Chapter 6.

4.3.1. **The Distribution of State Pairs**

In this subsection, properties of the final state of typical sequences are described. These properties are the direct result of the definition of the typical group in (3.6). This definition states that a typical sequence is one where the distribution of pairs of states is close to the expected number
of such pairs as determined by the Markov chain. This definition will be used to enable us to
establish the distribution of the final states among all the typical sequences.

Notice that the criterion for the typical group does not require any specific ordering of the
pairs of states. It seems intuitively obvious that if typical sequences have a certain distribution of
pairs, as the criterion above requires, then the last pairs of all such sequences should have the same
distribution. This phenomenon is easily seen if we just consider single states instead of pairs of
states. One simply examines all the permutations of sequences with a given combination of states.

To see this, consider the following example. Suppose that the sequence of states
\( A B B A A A \) is typical. If we just consider single states, there are \( 6/(4!2!) = 15 \) different
sequences, the multinomial coefficient for this particular combination of states. Since A's constitute
2/3 of the states in this combination, 2/3, or 10 of the 15 sequences should have an A as the final
state. This is easily verified by noting that there are only \( (6-1)![(4-1)!2!] = 5/(3!2!) = 10 \)
permutations possible when one of the A's is restricted to the last state. A similar conclusion is
possible for state B.

Unfortunately, permutations on single states alter the frequency of the pairs of states. Thus,
some of the permutations of the typical sequences may actually be atypical (i.e., sequences in \( \Gamma_2 \)).
For example, the sequence \( A B A B A A \) has the same combination of single states as the sequence
above, but does not have the same number of pairs of states. In addition, permutations of sequences
on single states with several different states may yield transitions from one state to another that are
not possible. Thus, simply permuting states to determine the distribution of final states in a set of
sequences creates two problems: 1) it alters the numbers of pairs of states, and 2) it can create
transitions that are not possible given the Markov chain. The question therefore becomes, given a
typical sequence, how many other sequences of the same length satisfy the following two constraints:

1) each sequence must have the same distribution of pairs, and

2) each sequence cannot have any impossible transitions.
It turns out that for combinations that have the same state at the beginning and end of the sequence, the number of different sequence that satisfies both constraints is equal to just the number of transitions in the sequence. For example, since the sequence A B B A A A is typical, only 4 other sequences satisfy the two constraints. To see this, first note the frequency of the pairs in this sequence as indicated in the table below.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>2</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
</tr>
<tr>
<td>BA</td>
<td>1</td>
</tr>
<tr>
<td>BB</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1

Thus, there are 2 transitions from state A to state A, 1 transition from state A to state B and so on. It is now possible to use a symbol to define each pair or digram such as in the table below.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Pair symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>W</td>
</tr>
<tr>
<td>AB</td>
<td>X</td>
</tr>
<tr>
<td>BA</td>
<td>Y</td>
</tr>
<tr>
<td>BB</td>
<td>Z</td>
</tr>
</tbody>
</table>

Table 2

The sequence A B B A A A can therefore be encoded using these pair symbols as X Z Y W W. With this approach, there will always be a one-to-one correspondence between a sequence and its pairwise representation. The advantage of doing this is that it allows us to permute these symbols without changing the distribution of pairs. Thus, constraint (1) above is always met for all permutations of the pair symbols. Constraint 2, on the other hand is still problematic. In fact, encoding a sequence illuminates how constraint 2 restricts the number of valid sequences.

Consider the sequence X Z Y W W. Note that, in general, symbols that are repeated, such as W in this case, can be interchanged without violating either constraint. But this results in exactly
the same sequence. For all cases where two different symbols are interchanged, an impossible transition results. For example, interchanging either W with Y is impossible. This is because in the permuted sequence X Z W W Y, the only symbol that can link, or be in between the Z and the W is a Y, by the definitions of these symbols. For these sequences, the only possible permutations are those accomplished by a rotation of the symbols, provided that the original sequence begins and ends with the same symbol. In sequences with different beginning and ending symbols, no rotations or permutations are possible.

To understand this caveat, notice that in the original sequence begins and ends with A. Thus, in our example, the following rotations of the pair-symbols and the corresponding rotations of the original symbols are possible.

**Correspondence Between Sequences and Transition Symbols**

<table>
<thead>
<tr>
<th>Transition-symbols</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>XZYWY</td>
<td>ABBAAA</td>
</tr>
<tr>
<td>ZYWXY</td>
<td>BBAAAB</td>
</tr>
<tr>
<td>YWWXZ</td>
<td>BAAABB</td>
</tr>
<tr>
<td>WWXZY</td>
<td>AAABBA</td>
</tr>
<tr>
<td>WXZYW</td>
<td>AABBA</td>
</tr>
</tbody>
</table>

Table 3

Note that in the symbols column, the frequency of the pairs of states in each of the listed sequences is the same even though the frequency of the individual states may vary. For example, in the first sequence there are 4 A's while in the second sequence there are only 3. In addition, each sequence in this column always begins and ends with the same symbol. Without loss of generality, consider for example the first sequence which has the symbol A at both ends. Removing the A at the beginning of the sequence, destroys a pair, specifically, the AB pair leaving a B at the beginning. This pair can, however, always be restored by putting the symbol that was in second place (and now in first, a B) at the end. Since the first sequence ended with an A, adding a B to the end of the sequence
restores the AB transition thus resulting in another sequence with the same symbol at both ends, a B in this case, while at the same time preserving the distribution of pairs.

Finally, note that in the sequences of pair-symbols, 2/5 of the symbols are W's and that 2/5 of the last states are also W's. This frequency correspondence is also somewhat more general. Note that in each of the sequences denoted with pair-symbols, 3/5 of the five symbols indicate transitions to the symbol A—the two W's and the Y. This is reflected in the fact that 3/5 of the sequences defined in terms of symbols end with an A. These properties lead to the following lemma which will allow us to approximate the expectation value of the final state in an annealing sequence and relate it to the entropy measure.

Lemma 1: Given a combination of state pairs in \( \Gamma_1 \) sequences, for any \( \varepsilon > 0 \) and \( r \) sufficiently large where \( r \) is the length of the subsequences, the maximum number of sequences with the same distribution of state pairs is \( O(r^{\varepsilon}) \).

Proof: See Appendix A. ■

Thus, knowledge of this distribution provides information on the distribution of the final states among all the typical sequence. Since these sequences have most of the probability in the entire sequence space (from Theorem 3.1), knowledge of this distribution provides insight into the expected value of the final state.

Now that we have some understanding of the distribution of the last pairs of states (and hence final state) in an SA experiment, the AEP can be utilized to show the connection between entropy and convergence. The asymptotic case is described in the next section. Latter subsections describe extensions to finite-time experiments.

4.4. **Entropy and Convergence in the Limit**

Now that the distribution of the last pairs and single states of typical annealing sequences is understood, it is possible to develop a theoretical relationship between the expected value of the final state in annealing sequences and the entropy measure \( H(m,n-2) \). This requires a method of
determining the number of typical sequences. This number can be determined by counting how many different combinations of pairs are in the typical group and how many sequences of each combination are possible. The latter can be done by operation of Lemma 1. The former is more difficult to determine, but can be approximated by considering the stationary distributions of states in SA and examining how these stationary distributions relate to the expected value of a finite-time run.

The stationary distribution in SA at time index \( k \) can be denoted by the vector \( \pi^{(k)} = \pi(t_k) = \left[ \pi_1(t_k), \pi_2(t_k), \ldots, \pi_s(t_k) \right] \), where each element is given by the following equation known as the Boltzmann distribution (Aarts and Korst [1989 pp. 14]):

\[
\pi^{(k)}_i = \pi_i(t_k) = \frac{e^{-\beta_i(t_k)}}{N_0(t_k)}
\]

(4)

where \( N_0(t_k) = \sum_{i=1}^{s} e^{-\beta_i(t_k)} \) is a normalization constant often referred to as the partition function in statistical mechanics (Bonomi et al. [1984]). Note that from its form, the closer the objective function value \( f_i \) is to the globally optimal value, the higher is its stationary probability \( \pi_i(t_k) \) at any fixed temperature \( t_k \). From (3.6), the number of \( ij \) transitions, \( m_{ij}(m,n-2) \), in the last subsequence of typical subsequences between time index \( m \) and time index \( n-2 \) for some \( \delta > 0 \) sufficiently large, is:

\[
\left[ \hat{E} \{ m_{ij}(m,n-2) \} - r\delta \right] \leq m_{ij}(m,n-2) \leq \left[ \hat{E} \{ m_{ij}(m,n-2) \} + r\delta \right]
\]

(5)

where \( r = n - m - 1 \), the length of the subsequence. Thus, the number of \( ij \) transitions in the final subsequence ranges between two integer values. Note that at low temperature values after a high number of iterations, the state distribution vector \( \mathbf{v} \) is close to the vector of stationary distributions \( \pi \) (see Aarts and Korst pp.18). For now, assume that these vectors are close in \( L^1 \) norm. The difference between them will be considered later. Thus,

\[
\hat{E} \{ m_{ij}(m,n-2) \} = v^{[m]}_{ij} p^{[m]}_{ij} + v^{[m+1]}_{ij} p^{[m+1]}_{ij} + \cdots + v^{[n-2]}_{ij} p^{[n-2]}_{ij}
\]

\[
= \pi^{[m]}_{ij} p^{[m]}_{ij} + \pi^{[m+1]}_{ij} p^{[m+1]}_{ij} + \cdots + \pi^{[n-2]}_{ij} p^{[n-2]}_{ij}
\]

(6)

This expression can be simplified by expanding its various components. First consider the case where there is a transition from state \( i \) to state \( j \). Recall that
\[ \Delta f^+_j = (f_j - f_i)^+ = \begin{cases} (f_j - f_i) & \text{for } f_j > f_i \\ 0 & \text{for } f_j \leq f_i \end{cases} \]

Thus, if \( f_j > f_i \), then using (4),

\[
\pi_i^{(k)} p_{ij}^{(k)} = \frac{e^{-f_i^k / k}}{N_0} \cdot \frac{1}{N} \cdot e^{-N_j^{(k)}/k} \\
= \frac{e^{-f_j^k / k}}{N_0} \cdot \frac{1}{N} \cdot e^{-f_i^k / k} \\
= \frac{e^{-f_j^k / k}}{N_0} \cdot \frac{1}{N} \\
= \frac{\pi_j^{(k)}}{N} 
\]

In general,

\[
\pi_i^{(k)} p_{ij}^{(k)} = \begin{cases} \frac{\pi_j^{(k)}}{N} & \text{for } f_j > f_i \\ \frac{\pi_i^{(k)}}{N} & \text{for } f_j \leq f_i \end{cases} \quad (7) 
\]

For the diagonal elements and using (7),

\[
\pi_i^{(k)} p_{ii}^{(k)} = \pi_i^{(k)} \left[ 1 - \sum_{l \neq i} \pi_l^{(k)} p_{il}^{(k)} \right] \\
= \pi_i^{(k)} - \sum_{l \neq i} \pi_i^{(k)} p_{il}^{(k)} \\
= \pi_i^{(k)} - \frac{1}{N} \left[ \alpha \pi_i^{(k)} + \sum_{l \neq i} \pi_l^{(k)} \right] 
\]

where \( \alpha \) is the number of neighbors that have objective function values less than or equal to \( f_i \).

Thus, it is possible to write the expression for (6) strictly in terms of the stationary probabilities. Using (4), (7), and (8), the upper and lower bounds for the frequency of occurrence of state pairs can be ranked in terms of their likelihood of occurrence. Thus, at low temperatures, the pair \( S_0 S_0 \) is the most likely pair followed by transitions to and from the global minimum to its next-to-least-cost neighbor \( S_0 S_1 \) and \( S_1 S_0 \) where \( S_n \) is the globally optimal state. \( S_1 \) is the next-to-least-cost
state, and so on. From (5), it is possible, however, that these frequencies may overlap depending on the temperature index. That is, the number of $S_y S_0$ pairs in a typical subsequence of length, say 10, may range from 7 to 9 while the number of $S_1 S_0$ pairs may range from 6 to 8. This indicates that there are numerous combinations of pairs that satisfy the requirements of (3.6).

Now we are getting close to understanding how to construct typical subsequences each of which has a certain number of transition/pairs based on (5). First of all, it is possible to calculate the numbers of transition/pairs in typical subsequences using (6). Because of (5), these numbers range between two integer values. Thus, there will be different combinations of these transition/pairs each of which has a number of rotations based on Lemma 1. For example, suppose a typical subsequence satisfying (6) can have 7 or 8 $ii$ transition/pairs and 6 to 7 $ij$ transition/pairs. Thus, one combination or class of subsequences will have 7 $ii$ pairs and 7 $ij$ pairs. This combination may have many rotations. Thus, this combination contributes several sequences to the typical group, though other combinations are possible. Another combination has 8 $ii$ pairs and 6 $ij$ pairs, also with several rotations that all become part of the typical group. Thus, for each combination, there are also different permutations. Lemma 1 allows us to approximate the number of sequences given a combination, but we still do not know how to calculate the number of different combinations. This problem can be illustrated using the diagram below where the large circles represent the sets of specific transition/pairs and the numbers in each large circle are the number of transition/pairs that satisfy (5):
In this example, (5) requires that the number of $S_0S_0$ transition/pairs range between 6 and 10. The number of $S_1S_1$ transition/pairs can range from 5 to 8, and so on. The problem of determining the number of different combinations of transition/pairs that satisfy (5) is equivalent to determining how many ways there are to pick one number from every bin (large circle) such that the sum of the numbers selected equals the length of the subsequence. Each different way of selecting these numbers is a unique combination of the transition/pairs that are part of the typical group. Each such combination may have a number of permutations bounded by $\rho^c$ as Lemma 1 suggests.

Although the problem of determining the number of different combinations of pairs in the typical group is difficult, all that is needed is an upper bound on the number of combinations. Therefore, some simplifying constraints will facilitate the subsequent analysis. These constraints are rather contrived and seem to simplify our problem to such an extent that the subsequent analysis appears to have no value. This is not the case, however, and these constraints will be removed later on. This will help to show some remarkable properties of combinatorial problems that lead us to the main result.

This simplification can be done by reducing the number of bins that must be considered in counting the number of combinations of transition/pairs. The number of bins can be reduced by reducing the expected number of certain transition/pairs in typical subsequences so that the upper
bound from (5) is 0. The ranking of the probability of occurrence of the transition/pairs makes it possible to reduce the expected number of certain pairs by examining the system at low temperature values. The temperature will be so low, in fact, that the mean objective function value is between the globally optimal value and the next-to-least-cost objective function value. The mean objective function value at temperature $t_4$ is defined by

$$\langle f(t_4) \rangle = \sum_{i \in S} \pi_i(t_4) f_i.$$  

(9)

Therefore, there exists some time index $m$ such that $f_{S_0} \leq \langle f(t_m) \rangle \leq f_{S_1}$. This is illustrated in Figure 3.

This state of the system has some interesting properties. Note that by (4), the stationary probability of a particular objective function value is a continuous function of $t$, the temperature parameter. Taking the derivative of the stationary probability with respect to temperature gives

$$\frac{\partial \pi_i(t)}{\partial t} = \frac{\pi_i(t)}{t^2} \left( f_i - \langle f(t) \rangle \right)$$  

(10)

(for a simple proof, see Aarts and Korst [1989 p. 22]). Thus, for globally optimal solutions at time index $k$, where $f_{S_0} \leq \langle f(t_k) \rangle \leq f_{S_1}$, $\partial \pi_i(t)/\partial t < 0$ while for the objective function value of the next-to-least-cost neighbor, $\partial \pi_i(t)/\partial t \geq 0$. This implies that with decreasing temperature (increasing time index $k$), the stationary probability of the next-to-least-cost neighbor's objective...
function value decreases and the stationary probability of the globally optimal solution increases. These conditions (when the system is at low temperatures), although contrived, permit the simplifications necessary to calculate the number of different combinations of state pairs and, hence, the number of typical sequences.

As indicated earlier, counting the number of different combinations of state pairs can be simplified by reducing the number of states that occur in typical sequences. This can be done at low temperatures and by choosing the value of $\delta$ in (3.6) small so that the upper bound on the number of transition/pairs involving the state $S_2$ is 0. Doing this limits the number of transition/pairs that need to be considered. Note that transitions to and from $S_2$ from states with lower objective function values equals $\pi_{S_2}^{[m]} / N$; that is, $\pi_{S_2}^{[m]} / \pi_{S_2}^{[m]} = \pi_{S_2}^{[m]} / N$ where $\alpha = 0, 1$. Since from (10), the probability of these transition/pairs decreases monotonically, it is possible to choose a time index $m$ such that

$$\frac{\pi_{S_2}^{[m]}}{N} < \delta$$

Then from (6) and (7),

$$\bar{E}\left[m_{S_0S_2} (m, m + r - 1)\right] = \frac{1}{N} \left( \pi_{S_2}^{[m]} + \pi_{S_2}^{[m+1]} + \cdots + \pi_{S_2}^{[m+r-1]} \right) \leq r \delta < \frac{1}{2}.$$  \hspace{1cm} (11)

This gives the upper bound in (5) to be 0. Consequently, a time index $m$ can be chosen such that only the ranges for the following transition/pairs need be considered in constructing typical subsequences: $S_0S_0$, $S_0S_1$, $S_1S_0$, $S_1S_1$. This means that only transitions between and among the globally optimal state and its next-to-least-cost neighbor need be considered. The following definitions are needed. Define the lower and upper bounds for the number of these transition/pairs based on (5) as follows:

$$l_{00} \leq m_{S_0S_0} (m, m + r - 1) \leq u_{00}$$
$$l_{01} \leq m_{S_0S_1} (m, m + r - 1) \leq u_{01}$$
$$l_{10} \leq m_{S_1S_0} (m, m + r - 1) \leq u_{10}$$
$$l_{11} \leq m_{S_1S_1} (m, m + r - 1) \leq u_{11}$$  \hspace{1cm} (12)
where the $l_i$ and $u_i$ are integers. Note that from (7) $l_{i_t} = l_{t_t}$, and $u_{i_t} = u_{t_t}$. Define $c_i = u_i - l_i + 1$.

This corresponds to the number of integer values for occurrences of a transition/pair in a typical subsequence. Figure 4 illustrates these 4 classes of transition/pairs that make up all of the typical subsequences when the system is at low temperatures.

**Four Classes of State Pairs**

![Diagram of Four Classes of State Pairs]

**Figure 4**

One question of interest is the rate at which the number of typical sequences increases as the length $r$ increases, where a globally optimal and/or next-to-least cost suboptimal state is the final state. This leads to the following lemma:

**Lemma 2:**

Given a temperature sufficiently low and a tolerance value $\delta$ sufficiently small, the number of typical annealing subsequences with states $S_i$ and $S_j$ as the final state, $a_{S_i} \cdot a_{S_j}$ respectively, in terms of their length $r$, is $O(r^{r+1})$ where $r > 0$ is small.

**Proof:** See Appendix A. □

These results lead to the following theorem:

**Theorem 1:** Given some time index $m$ and some $\varepsilon > 0$, then for the length of the final subsequence $r$ sufficiently large, such that $f_{S_i} \leq f(\ell_{m}) \leq f_{S_j}$, the expected value of the final state in an SA experiment $\langle f^{M+r-\varepsilon} \rangle$ at time index $m + r - 1$ is

$$\langle f^{M+r-\varepsilon} \rangle \leq k \left[ \frac{d_{S_i} f_{S_i} + d_{S_j} f_{S_j}}{e^{H_{M+M+\cdots+M} - H}} \right] = \frac{\omega}{e^{H_{M+M+\cdots+M}}}$$  \hspace{1cm} (13)
where \( \omega = \Theta r^{r+1} \), and \( H(m,m+r-2) \) is the entropy of the inhomogeneous Markov chain from time index \( m \) to time index \( m+r-2 \).

**Proof:** See Appendix A. 

This theorem hints at a relationship between a convergence rate and the mean objective function value of an SA experiment. It suggests that since the entropy measure is the exponent in the denominator, the higher the entropy, the lower the mean objective function value of the final state. This theorem has several significant limitations, however. Most significant is the fact that it holds only in the case for \( f_{\omega_n} \leq \langle f(t_m) \rangle \leq f_{\omega_i} \) (i.e., for temperature \( t \) sufficiently close to zero). In order to generalize this theorem so that its basic form holds true when the mean objective function value is larger, i.e., \( t \leq \langle f(t_m) \rangle \), certain scaling properties of SA must be employed. These are described in the next section.

### 4.5. Scaling Properties of SA

In this subsection, some interesting scaling properties associated with SA are developed that allows us to generalize the results of the preceding subsection. Recall that our previous analysis was somewhat contrived in that the number of optimal states in subsequences was based on the stationary distribution (see (6)) and also because the SA algorithm was allowed to run long enough so that the mean objective function value was somewhere between the globally optimal objective function value and the objective function value of its next-to-least-cost neighbor. That is, \( f_{\omega_n} \leq \langle f(t_n) \rangle \leq f_{\omega_i} \).

Suppose that we stop the algorithm earlier. How does this affect the results of the preceding subsection?

To illustrate the way in which the previous analysis is generalized, consider Figure 5 where the mean objective function value at time index \( m \) is such that \( f_{\omega_n} \leq \langle f(t_m) \rangle \leq f_{\omega_i} \).
State of the System at a Higher Temperature

\[ f(t_m) \]

\[ S_2 \]

\[ S_1 \quad \text{Next-to-least-cost neighbor} \]

\[ S_0 \quad \text{Globally optimal state} \]

Figure 5

It is possible to apply the same analysis as before by creating a super-node as a substitute for the globally optimal node of the prior analysis. This is done by lumping or combining a number of the least-cost nodes below the mean objective function value into one super-node as illustrated in Figure 6.

Illustration of a Super-Node

\[ f(t_m) \]

\[ S_2 \]

\[ S_1 \]

\[ S_0 \quad \text{Super-node} \]

Figure 6

To make use of this idea, the super-node must be given a certain property. With a suitable and reasonable definition for this property, we can apply the same analysis by treating the super-node as the new globally optimal state. This property naturally leads to reasonable definitions for the super-node's stationary probability and its objective function value. This property is stated simply as follows:
Property: The system is in the state $\hat{S}$, corresponding to the super-node, if it is in any state it contains.

In Figure 6, this means that the system is in state $\hat{S}$ if it is in either state $S_0$ or state $S_1$. This leads to a reasonable definition for the super-node's stationary probability. Since the system is in $\hat{S}$ if it is in any of its member states, then the stationary probability of the system being in state $\hat{S}$ should naturally be the sum of the stationary probabilities of its member states. The following definitions will be useful.

Definitions:

$\hat{S}_i = \{S_i : 0 \leq i \leq k\}$. This corresponds to the super-node that contains states $S_0$ to $S_k$. In Figure 6, therefore, $\hat{S}_1 = \{S_0, S_1\}$.

$\pi_s^{[m]} = \sum_{i=0}^{k} \pi_{S_i}^{[m]}$. This is the stationary probability of the super-node at time index $m$ and is the sum of the stationary probabilities of all the states in the super-node at time index $m$. In Figure 6, therefore, $\pi_1^{[m]} = \pi_{S_0}^{[m]} + \pi_{S_1}^{[m]}$.

All that remains to define is a reasonable objective function value for the state $\hat{S}$. A reasonable definition (owing to the properties described below) would be the weighted average of the objective function values of the states contained within the super-node.

Definition:

$$j_k^{[m]} = \frac{\sum_{i=0}^{k} \pi_{S_i}^{[m]} f_{S_i}}{\sum_{i=0}^{k} \pi_{S_i}^{[m]}}$$

the objective function value of the super-node and is based on the following equality:

$$\sum_{i=0}^{k} \pi_{S_i}^{[m]} f_{S_i} = \sum_{i=0}^{k} \pi_{S_i}^{[m]} j_k^{[m]}$$

Thus, in Figure 6 the objective function value of $\hat{S}$ is
\[ j_i^{[m]} = \frac{\pi_n^{[m]} f_{S_0} + \pi^*[n] f_{S_1}}{\pi_n^{[m]} + \pi^*[n]} \]

since

\[ \pi_n^{[m]} f_{S_0} + \pi^*[n] f_{S_1} = \pi_n^{[m]} j_i^{[m]} + \pi^*[n] j_i^{[m]} \]

\[ = \left( \pi_n^{[m]} + \pi^*[n] \right) j_i^{[m]} \]

\[ = \pi^*[n] j_i^{[m]} \]

Note that the objective function value of the super-node is a function of the time index.

This property of the super-node, its stationary probability, and its objective function value, gives rise to certain properties of the super-node that mimic the globally optimal state of the preceding analysis. These properties can therefore be used to show that the analysis of the preceding subsection can be applied directly to the general case where the time index \( m \) is arbitrary. These properties are described in the following theorems.

**Theorem 2:**

For all time indices \( m \), the objective function value of the super-node is less than the mean objective function value. Thus, \( j^{[m]}_k < \langle f(t_m) \rangle \).

**Proof:** See Appendix A. ■

Theorem 2 therefore shows that the objective function value of the super-node is always less than the mean objective function value, as is the case for the globally optimal objective function value. In this sense, the super-node behaves as if it were the globally optimal node.

**Theorem 3:**

The stationary probability of the super-node increases monotonically with decreasing values of the temperature parameter \( t \).

**Proof:** See Appendix A. ■

Theorems 2 and 3 show that the super-node shares all the properties of the globally optimal node in terms of all the mathematical properties upon which Theorem 1 is based. In fact, the derivative with respect to temperature of the stationary probability of the super-node has exactly the
same form as the derivative with respect to temperature of the globally optimal node (see the proof in Appendix A). Because of this, the results of Theorem 1 can be generalized to the more general case where the mean objective function value is greater than \( f^*_i \), simply by substituting the objective function value of the super-node into (13) and expanding this objective function value based on its definition. Doing this, however, introduces some errors. Recall that Theorem 1 is based on the fact that the state distribution vector \( v(m) \) and the stationary distribution vector \( \pi(m) \) are close in norm so that the approximation in (6) can be made. These vectors are arbitrarily close to each other in the case where \( f^*_m \leq \langle f_i \rangle \leq f^*_m \). This is easily seen from the fact that \( \lim_{m \to \infty} \pi(m) = \lim_{m \to \infty} v(m) = \text{e}^* \) (see Aarts and Korst [1989 pp.18] and Mitra [1986 pp. 763]). The following section shows that these vectors become further apart the earlier the SA algorithm is stopped.

### 4.5.1. The Magnitude of the Error

In this subsection, the difference in norm of the vectors \( v(m) \) and \( \pi(m) \) is explored. In a similar fashion as in Mitra et al. [1986], let

\[
v(m) - \pi(m) = \{ v(m) - \pi(0) \} P(0, m) + \{ \pi(0) P(0, m) - \pi(m) \}
\]

Therefore, based on the triangle inequality

\[
\|v(m) - \pi(m)\| \leq \|v(m) - \pi(0)\| P(0, m) + \|\pi(0) P(0, m) - \pi(m)\|.
\] (14)

Mitra provides bounds for the two terms on the right-hand side in (14) (see Mitra et al. [1986 pp. 763-65]). First and foremost, these terms become smaller as the time index \( m \) increases and are larger when the time index \( m \) is smaller. Also, these terms depend on the radius of the graph, although its overall effect is difficult to quantify. The main point here, however, is that the expression in Theorem 1 suggests a type of convergence that relates the entropy of inhomogeneous Markov chains to the expected value of the final state in annealing experiments. This relationship, however, becomes weaker the earlier the annealing experiments are stopped. This effect will be illustrated in the experiments described in Chapter 6.
CHAPTER 5

EXPERIMENTAL METHODOLOGY

So far, SA has been described in very abstract terms. The configuration space, neighborhood structure and cooling schedules, have all been described in very general terms, along with the necessary mathematical elements that define an SA implementation. In this chapter, the details of how the SA algorithm is implemented on combinatorial optimization problems, as well as idealized problems using generic configuration spaces (alluded to in Chapter 4) are presented. In addition, the method of how the entropy values and performance measures are calculated are also described. The purpose of these methods is demonstrate the theoretical relationships developed in Chapters 3 and 4. Two principal approaches are used to show these theoretical relationships, namely, the generic SA experiments, and transformation experiments on combinatorial optimization problems.

5.1. GENERIC ANNEALING EXPERIMENTS

To show the relationship between the finite-time performance of SA and the entropy measure described above, generic annealing experiments were developed. These experiments are based on a standard implementation of SA using what is referred to as a generic configuration space. The term generic is used because the algorithm used and the configuration spaces on which the algorithm is run, do not depend on any particular type of combinatorial optimization problem, but rather, are general in nature and, therefore, provide a good test-bed for the theory. The specifics of the generic configuration space are described below.
5.1.1. The Generic Configuration Space

The most important aspect of generic configuration spaces is that they allow each objective function value to be specified independently and stored in computer memory. These objective function values are then connected in a graph to any specified number of other objective function values. This number constitutes the neighborhood size. Thus, the configuration space can be created in a very flexible manner.

Specifying all possible objective function values in this manner is admittedly unrealistic since it is computationally equivalent to an exhaustive search of the entire configuration space (which would make SA useless). Furthermore, exhaustively specifying objective function values is costly and less efficient in terms of computer memory usage than using an actual combinatorial optimization problem (which is done later). In combinatorial optimization problems, an objective function value is associated with some combination of attributes peculiar to the type of combinatorial optimization problem in question. The SA algorithm moves from one objective function value to a neighbor by modifying the current combination of attributes associated with the problem. Thus, at any point in the algorithm, only the current combination of attributes and its associated objective function value need be stored in memory, whereas in the generic case, all objective functions must be stored. In effect, combinatorial optimization problems illustrate an efficient method of encoding all possible objective function values for these problems—if one desires, one can decode and list all objective function values by enumerating all the possible combinations of attributes. These attributes, their combinations, and some neighborhood structure are all that is necessary to uniquely define the entire configuration space in SA for standard combinatorial optimization problem. In the generic case, however, each objective function value must be uniquely defined in order to define the configuration space. So why use generic configuration spaces?

Although this method of testing SA is costly in terms of memory usage, its benefits are numerous and allows for many different properties to be tested. For example, properties of a
combinatorial optimization problem that may constrain objective function values to a certain range are not present and any desired topology for the configuration space can be created. The configuration space can be made "bumpy" or "smooth", have any specified limits on the magnitude of changes in neighboring objective function values, the neighborhood size can be specified and easily changed to many different values, and the number of solutions is limited only by available memory. In short, we have total flexibility to create any form for the configuration space and examine how the SA algorithm performs on it. Since the goal here is to test the finite-time performance of SA and its connection to the entropy measure, how efficiently the configuration space has been encoded is not a concern. The next section describes, in detail, how these configuration spaces are generated.

5.1.1.1. Creating Generic Objective Function Values

Since the purpose of generic configuration spaces is to facilitate experimentation, they must be simple to create and modify. These configuration spaces are all derived from a simple list of numerical values which constitute the objective function values (often referred to as the solution space). This list is created using the linear congruential random number generator of the standard math library for the C programming language (see Kernighan and Ritchie [1978]). A program was developed that takes four parameters: 1) an upper bound on objective function values, 2) a lower bound, 3) the number of objective function values to generate, and finally, 4) a random number seed. This program creates the solution space with objective function values within the specified range and saves it as a separate file that is then used by other programs to define the configuration space.

5.1.1.2. Creating the Neighborhood Structure

The neighborhood structure is superimposed on the solution space with all solutions having the same neighborhood size. This is accomplished by defining the neighboring solutions of a particular solution to be adjacent solutions in the solution list. To illustrate, let \( S_j \) be the \( i \)th value in the solution list which constitutes the solution space \( S \). Thus, \( S = \{ S_1, S_2, \ldots, S_n \} \) for a solution space
of size \( n \). The neighbors of \( S_i \) are \( S_{i+1} \) and \( S_{i-1} \) for a neighborhood size of 2. In the implementations here, the neighborhood size is always some multiple of 2. This facilitates creating the neighborhood structure in a simple manner without the complications caused by using odd numbers of neighbors. Thus, it is a simple matter to determine the neighbors of a particular solution given the solution list. In general, the set of neighbors of \( S_i \), denoted as \( N(i) \), can be defined as follows:

\[
N(i) = \left\{ S_j \mid \left( i - \frac{n}{2} \right) \mod n \leq j \leq \left( i + \frac{n}{2} \right) \mod n \right\} \land \left( j \neq i \right)
\]

The modulus notation indicates that the neighbors of solutions "wrap around". For example, when the neighborhood size is 2, the neighbors of \( S_n \) are \( S_1 \), and \( S_{n-1} \) and so on, while the neighbors of \( S_1 \) are \( S_n \) and \( S_2 \).

5.2. EXPERIMENTS ON COMBINATORIAL OPTIMIZATION PROBLEMS

One of the methods used to illustrate the effect on SA's performance resulting from changes to the configuration space is by transforming combinatorial optimization problems into other combinatorial optimization problems. This permits comparisons to be made between the performance of SA on the original problem and its performance on the transformed problem. Since both of these problems have the same optimal solution but different configuration spaces, the effect of the configuration space on performance becomes apparent. These issues are addressed in detail below in Subsection 5.2.1.

5.2.1. NP-completeness and NP-hard Optimization Problems

The theory of computational complexity states that there exists a class of decision problems for which solutions can be verified in polynomial time (these are problems for which the answers are either yes or no). In addition, polynomial-time transformation algorithms may exist between such problems. The general class of these problems is termed \( NP \) (non-deterministic polynomial). The class \( NP \)-complete \( \subseteq \) \( NP \) constitutes the hardest problems in \( NP \). This class is defined by all
problems $L_2$ (L represents a problem encoded on a Turing Machine (Garey and Johnson [1979] p.34).) such that for all $L_1 \in NP$, $L_1$ transforms to $L_2$.

The concept of polynomial transformations involves using an algorithm to transform a problem instance of one type of problem to a problem instance of another type of problem in such a way that the answers to both decision problems are always the same. This allows for the possibility of solving a decision problem by solving the problem it transforms to. Therefore, a polynomial time algorithm exists for any NP-complete problem if and only if $P = NP$, hence, the existence of polynomial time algorithms for all members of the class $NP$. Since it is unlikely that $P = NP$, deterministic and stochastic heuristics have been developed to obtain good solutions for NP-complete problems. For a complete exposition, see Garey and Johnson [1979 pp. 17–44].

The above description specifically refers to decision problems. The solution to such problems is either a "yes" or a "no". There are problems $\Pi$, however, in which $\Pi \not\in NP$ (problems that are not decision problems, such as optimization problems) and yet that have problems in $NP$ which reduce to them. The concept of reductions is a generalization of transformations that is necessary to relate difficult decision problems (NP-complete) to difficult optimization problems. Such optimization problems are said to be in the class $NP$-hard.

These concepts are useful in assessing how transformations and reductions impact the performance of SA. Such an approach has not, to date, been taken in the study of SA. The next section highlights the focus of prior research on SA.

5.2.2. Objective Functions and Neighborhood Size

In this subsection, the specifics of certain combinatorial optimization problems mentioned earlier and the issues of how the objective functions are defined, as well as how the neighborhood structure and size are determined are discussed. A series of combinatorial optimization problems were selected and appropriate software developed to perform SA experiments. SA algorithms were developed for optimization versions of the following combinatorial problems: the independent set,
the clique, the vertex cover, the 3-SAT and the Traveling Salesman problem. These problems were selected because of their relationships to each other and the ease of implementing transformation algorithms associated with them. The independent set, clique, vertex cover and 3-SAT problems, allow for an empirical study that may show how transformations from one problem to another affect the performance of the SA algorithm.

Below is a description of how configuration spaces for these problems were constructed. Central to constructing them is defining suitable objective functions that are easy to compute and neighborhood structures which facilitate the search through the configuration space. In order to do this, the following definitions will be useful for a graph $G$ with $V' \subseteq V$.

Definitions:

\[
\begin{align*}
E_c(V) &= \text{The number of edges connecting nodes in set } V' \\
&\quad \text{i.e., } \forall u, v \in V', \{u, v\} \in E. \\
E_n(V) &= \text{The number of edges not connecting nodes in set } V' \\
&\quad \text{i.e., } \forall u, v \in V', \{u, v\} \notin E. \\
E_v(V) &= \text{The number of edges, given set } V', \text{ where each such edge } \text{has neither end in set } V' \\
&\quad \text{i.e., } \forall u, v \in \forall V', \{u, v\} \in E. \\
v' &= \text{The number of nodes in set } V'.
\end{align*}
\]

5.2.2.1. Independent Set

In the optimization version of the independent set problem, the goal is to determine the maximum number of nodes in a graph such that no two nodes are adjacent. In applying the SA algorithm to this problem, a method of generating candidate solutions in the configuration space and calculating the objective function values corresponding to these solutions must be devised. Since solutions to this problem correspond to specific combinations of nodes, a reasonable way of defining the configuration space would be to have all combinations of nodes correspond to solutions. In this way, the neighborhood structure could be defined by allowing candidate solutions to be obtained by adding or subtracting one node from a given set thereby creating a new combination of nodes and, hence, a new solution.
It is often the case, as it is here, that generating candidate solutions that satisfy all the constraints in a particular problem may be difficult. For instance, in the independent set problem, including a node previously not in the set may have edges connecting it to nodes already in the set thus violating the constraint that no nodes in the set can be connected by an edge. One way around this problem is to use some penalty function for solutions that violate these constraints. One objective function formulation for the independent set suggested by Aarts and Korst [1989 pp.81], is the following: Let \( V' \) be the set of nodes corresponding to the current solution. \( E_I(V') \) is therefore the set of arcs connecting nodes in \( V' \) and, hence, violate the independence criteria. This objective function is defined as

\[
    f_{\text{indep set}}(V', G) = v' - \lambda E_I(V') 
\]

(1)

where \( \lambda \) is the penalty parameter and \( G \) is the graph in question. Since this is a maximization problem, we subtract the penalty term. As the SA algorithm proceeds and as the temperature parameter decreases, the probability of accepting a candidate solution violating the constraints is reduced. Typically, at termination of the SA algorithm, a solution which does not violate the constraints is produced.

By including solutions which violate the constraints imposed by the above problems through the use of a penalty term, the problem of selecting candidate solutions has been simplified. Without a penalty term, adding or deleting a node (flipping) from the combination of nodes corresponding to the current solution would require verification that the change does not violate the constraints of the problem. For any particular problem instance, only certain nodes from a particular combination would be viable candidates for flipping and each combination would likely have different numbers of neighbors. With the addition of a penalty term, all nodes can be flipped and each solution (combination of nodes in \( V \)) has \( n \) neighbors. Inclusion of a penalty term also increases the size of the configuration space. This clearly makes implementation easier. Therefore, for the graph problems described here, the number of different combinations of nodes is
Therefore, there are $2^n$ solutions, including the trivial solution.

5.2.2.2. Clique

In formulating an objective function for the optimization version of the clique problem, a similar approach to that of Aarts and Korst [1989 pp. 81] can be used. The optimal solution to the clique problem is the maximum number of nodes in a graph such that all such nodes are connected (the opposite objective of the independent set). In other words, we attempt to determine the maximum size of a complete subgraph (see Bondy and Murty [1986] for a complete discussion).

Whereas in the independent set problem we decrease or penalize the objective function by the number of arcs which should not be in the current set, in the clique problem it is reasonable to penalize the objective function by the number of arcs missing from the set, $E_1(V)$. Thus, we have formulated the objective function

$$f_{\text{clique}}(V', G) = v' - \lambda E_2(V').$$

Note that $E_2(V)$ can be related to $E_1(V)$ by

$$E_2(V') = v' \left( \frac{v' - 1}{2} \right) - E_1(V').$$

Therefore, $f_{\text{clique}}(V', G)$ can be written as

$$f_{\text{clique}}(V', G) = v' - \lambda \left[ v' \left( \frac{v' - 1}{2} \right) - E_1(V') \right].$$

5.2.2.3. Vertex Cover

In the optimization version of the vertex cover problem, the minimum number of nodes in a set such that all edges in the graph connect to at least one such node must be determined. If $E_3(V)$ is the number of uncovered edges in a graph given set $V'$, then a suitable objective function is

$$f_{\text{vertex cover}}(V', G) = v' + \lambda E_3(V').$$
where the penalty parameter is added to the value of $v'$ since the vertex cover is a minimization problem. In this case, $E_3(V)$ is related to $E_1(V)$ by

$$E_3(V') = E_1(V \setminus V').$$

(7)

### 5.2.2.4. 3-SAT

In the optimization version of the 3-SAT problem, the objective is to determine a truth assignment for a set of boolean variables that maximizes the number of boolean clauses that have a truth value of one, where each clause has exactly three variables (see Garey and Johnson [1979 pp. 259]). Thus, given a truth assignment $U$ over a set of clauses $C$, an objective function can be defined as

$$f_{3\text{-SAT}}(U, C) = \sum_{c \in C} c$$

where the values for the $c \in C$ are all determined by the truth assignment $U$. For the 3-SAT problem, the configuration space for $n = |U|$ boolean variables also provides for $2^n$ solutions. Since each solution can be perturbed by flipping or inverting the value of one of the boolean variables, each solution also has $n$ neighbors.

### 5.2.3. Implementation Schemes

In the actual implementation of SA, it is useful to develop a simple expression for $\Delta f$. For many problems, it is much easier to compute $\Delta f$ indirectly than by reevaluating the objective functions and taking the difference in the current solution and candidate solution. In the independent set problem, the following expression illustrates a simple expressions for $\Delta f$. If $u'$ is a candidate node in a graph which is being considered for inclusion in set $V'$, then the following expressions equal $\Delta f$ caused by the addition of node $u'$ to the set.
\[
\Delta f_{\text{indep set}} = \left[ \chi_{V-V'}(u') - \chi_{V'}(u') \right] \left( 1 - \lambda \sum_{[u', v] \notin E} 1 \right)
\]

where \( \chi_{a}(u) = \begin{cases} 1 & \text{if } u \in a, \\ 0 & \text{otherwise} \end{cases} \) Aarts and Korst [1989] p.82).

Using similar reasoning,

\[
\Delta f_{\text{clique}} = \left[ \chi_{V-V'}(u') - \chi_{V'}(u') \right] \left( 1 - \lambda \sum_{[u', v] \notin E} 1 \right),
\]

and for the vertex cover,

\[
\Delta f_{\text{vertex cover}} = \left[ \chi_{V'}(u') - \chi_{V-V'}(u') \right] \left( \lambda \sum_{[u', v] \notin E} 1 \right) = \Delta f_{\text{clique}}.
\]

These functions have been implemented in code. Using these functions, the values for \( \Delta f \) are based only on the node that is to be either included or excluded and the arcs it is connected to. This avoids reevaluating the objective function value based on rescanning the entire graph.

This principle of computing \( \Delta f \) is not as readily applied to the 3-SAT problem. In 3-SAT, the truth value of a set of clauses changes when a boolean variable is changed from zero to one or vice versa. Since only the clauses that contain that particular boolean variable are affected, only those clauses need to be evaluated. This idea is accomplished by using a table look-up method in which a table for each variable is compiled at run-time that lists all the clauses that variable is in. Thus, when a variable is flipped, the look-up table is used to check which clauses contain that variable and only these clauses are evaluated.
5.2.4. Effect of Transformations on the Configuration Space

For SA to work effectively, Johnson et al. [1989] observed that it is desirable to have neighborhoods with smooth structures, thereby permitting easy escape from local minima. Goldstein and Waterman [1988] suggest that it may be disadvantageous to have too many neighbors because it tends to smooth out the configurations space (by reducing the number of local optima) and allows easy escape from the global optima. This indicates two opposing views in the literature. Naturally, the neighborhood generation mechanism must provide the potential for all solutions to be explored. But the particular neighborhood size used in any particular problem seems to affect the performance of SA (Goldstein and Waterman [1988]). Given that a particular problem has numerous local minima of various depths (determined by the configuration space and neighborhood structure), the question arises whether a transformation to another problem can modify the configuration space in such a way that it improves the performance of SA. In Section 5.2.4.1, we explore transformations among the three graph problems cited earlier are explored. For these problems, the configuration spaces on transformed problems are equivalent. As we shall see, however, the transformation from the 3-SAT problem to the clique problem does modify the configuration space, hence the performance of SA.

5.2.4.1. Transformations between Independent Set, Clique and Vertex Cover

To create a problem instance of the clique problem that produces identical results as for the independent set problem, a graph transformation is necessary. The transformation operation is simple: if an edge connects any two nodes in the graph, remove the edge. Conversely, if any two nodes do not have an edge connecting them, link the two nodes with an edge. The resulting graph $G^c$ is referred to as the complement graph of $G$. This operation can be performed in polynomial time in the size of the problem instance.
In Section 5.2.2, $E_1(V)$ is defined as the number of edges connecting nodes in a set $V' \subseteq V$. Let $E^c_1(V')$ be the number of edges connecting nodes in a set $V'$ for a graph $G^c$. From the relationship between the independent set and the clique problems,

$$E^c_1(V') = V' \left( \frac{V' - 1}{2} \right) - E_1(V').$$

(12)

Substituting $E^c_1(V')$ for $E_1(V')$ in the equation for $f_{\text{clique}}(V', G)$ to obtain $f_{\text{clique}}(V', G^c)$ leads to

$$f_{\text{clique}}(V', G^c) = f_{\text{indep set}}(V', G).$$

(13)

Thus, for complement graphs, the number of nodes are the same, the number of possible solutions is the same, the size of the neighborhoods are the same and, more significantly, the values of the objective functions for a given set of nodes are the same (provided that the penalty parameters are the same). This indicates that the application of the annealing algorithm to two related versions of these problems would show no difference in performance, since the configuration spaces are identical.

The equivalence of the independent set clique problems to the vertex cover problem is less obvious. For one thing, the vertex cover is a minimization problem, whereas the other two are maximization problems. To determine whether the configuration spaces are equivalent, we need to examine each problem's configuration space. To do this, some basic graph theory will illuminate the ways in which the configuration spaces of the independent set and vertex cover are related.

For a given graph $G$, if $V'$ is a vertex cover, then $\forall V'$ is an independent set (Garey and Johnson, [1979 pp. 53]). This implies that if $\alpha$ is a maximum independent set and $\beta$ is a minimum vertex cover, then $\alpha + \beta = |V|$ (Bondy and Murty [1976 pp. 101]). While the above relationship is true for any graph, recall that the configuration spaces for these two problems are based on objective functions which permit violations of the constraints to each of these problems. Although the number of solutions in the configuration spaces are identical for the independent set and the vertex cover problems, the values of the associated objective functions may be different. Nonetheless, even with
the inclusion of the penalty term and the corresponding increase in the size of the configuration space. If the graphs and penalty parameters are identical, the configuration spaces are mirror images of each other. Using the identity in (7),

\[ f_{\text{indep set}}(V', G) + f_{\text{vertex cover}}(V \setminus V', G) = |V| \quad (14) \]

Consequently, the configuration spaces are equivalent and transforming one problem to the other would be of no benefit insofar as running the SA algorithm.

5.2.4.2. Transformation from 3-SAT to Clique

The transformation from the 3-SAT problem to the clique problem provides further insight in determining how a configuration space might be altered. In this case, a definite modification to the configuration space occurs. This is best illustrated by describing the transformation. For each boolean variable in the 3-SAT problem, there is a corresponding node in a graph for the clique problem. Each boolean variable in 3-SAT has a corresponding node in clique that is connected to all other variables (nodes) except nodes that 1) correspond to boolean variables in the same clause or, 2) that correspond to that variable's negation (inverse) (see, for example, Manber [1989] p.352). The following example illustrates this transformation:

3-SAT to Clique Transformation

![Diagram of 3-SAT to Clique Transformation]

Figure 1

In the example in Figure 1, the dashed rectangular boxes correspond to clauses, each of which has three boolean variables represented by circles. These circles correspond to nodes in a graph and the
lines correspond to the arcs in the graph resulting from a transformation according to the rules outlined above. The shaded nodes correspond to variables that have truth value of one. Thus, \( x = \bar{y} = 1 \) and three clauses are satisfied, or evaluate to 1. Consequently, the objective function value for the optimization version of the 3-SAT problem is three. Notice that for the variables that evaluate to one, and hence cause that clause to evaluate to one, the lines connecting them are heavier. These heavier lines form cliques each of which is the same size as the number of clauses that have truth value one. Here, four cliques of size three are apparent.

This transformation yields a configuration space for the clique problem that is significantly different from the configuration space corresponding to the original 3-SAT problem. In particular, for \( b \) boolean variables in a 3-SAT problem with \( c \) clauses, there are \( 2^b \) solutions. The transformation to clique changes the number of solutions to \( 2^{3c} \) and changes the neighborhood size from \( b \) to \( 3c \). If \( c > b/3 \), then increases in the neighborhood size and solution space are possible.

### 5.3. The Cooling Schedule

Recall that the transition probabilities in SA depend on what has been referred to as the temperature parameter. During the annealing process (both physical and simulated), this temperature parameter decreases such that, with each iteration of the algorithm, the state distribution vector approaches the optimum vector. The manner in which the temperature decreases is governed by a cooling schedule. This schedule determines the value of the temperature parameter during each iteration of the algorithm. Many different cooling schedules have been suggested for optimizing the performance of SA. The one used in the experiments described in Chapter 6 is presented in Mitra et al. [1986] (see (1.3)). This schedule ensures convergence in probability. For these experiments the cooling schedule has the form

\[
t_k = \frac{\gamma}{\log(c + k)}
\]

where \( \gamma \) and \( c \) are constants.
To maintain uniformity in the experiments and make comparisons of performance between different configuration spaces, identical cooling schedules are used. To facilitate this, three parameters are defined and used in the various problems instances for determining the cooling schedule: the initial temperature, $t_i$, the final temperature, $t_f$, and the number of iterations used in each experiment, $n$. Using these three parameters, it is possible to determine the appropriate values for $\gamma$ and $c$ so that at iteration 1 the temperature is $t_i$, and at iteration $n$ the temperature is $t_f$.

To see this, first note that from the definitions of $t_i$, $t_f$, $n$ and the cooling schedule defined in (15), the following equalities hold:

$$t_i = t_1 = \frac{\gamma}{\log(c+1)} \quad \text{and} \quad t_f = t_n = \frac{\gamma}{\log(c+n)}$$

Using the first equality, the value of $c$ can be defined in terms of the value of $\gamma$. Thus,

$$c = e^{\gamma/t_i} - 1.$$  \hspace{1cm} (16)

Using the second equality, the value of $\gamma$ can be written in terms of the final temperature, $t_f$ and the number of iterations, $n$. Thus, $\gamma = t_f \log(c+n)$. This expression for $\gamma$ can be substituted into (16) to yield

$$c = (c+n)^{t_f/t_i} - 1.$$  \hspace{1cm} (17)

Using (17), define the function

$$g(c) = (c+n)^{t_f/t_i} - 1.$$  \hspace{1cm} (18)

Now note that

$$\frac{\partial g}{\partial c} = \frac{t_f}{t_i} (c+n)^{t_f/t_i - 1}$$

Since $t_f < t_i$, then $t_f/t_i < 1$ and the exponent $t_f/t_i - 1 < 0$. Consequently, $0 < \partial g/\partial c < 1$. Thus, the fixed-point theorem holds and the recursion

$$c_{k+1} = (c_k + n)^{t_f/t_i} - 1.$$
converges quickly to the value of $c$ (see Haaser and Sullivan [1991 pp.93]. The value for $c$ thus obtained is then used to compute the appropriate value for $\gamma$ from the equation $\gamma = t_f \log(c + n)$.

5.4. Calculation of the Entropy Measure

In Chapter 3, the entropy measure for inhomogeneous Markov chains was defined as

$$H^*(m, n - 2) = \sum_{k=m}^{n-2} H(k)$$

where $H(k) = -\sum_{i=1}^{r} \sum_{j=1}^{s} v_i^{[k]} p_{ij}^{[k]} \ln p_{ij}^{[k]}$. These expressions require a determination of all the transition probabilities $p_{ij}^{[k]}$ as well as all the state probabilities $v_i^{[k]}$ at each time index from $m$ to $n-2$. Clearly, this computation is at least as difficult as a total enumeration of all the solutions. Again, the goal here is to demonstrate the connection between performance and entropy. Thus, the complexity of this computation is not a concern.

In determining these probabilities, however, certain factors are more problematic than others. Thus, in the transition probabilities, the probability of generating a neighbor is simple and for these problems, remains constant at $1/N$. The acceptance probability (see (1.1)), however, is more difficult as this involves determining objective function values of the neighboring candidate solutions.

5.4.1. Calculating $\Delta f$

The chief difficulty in calculating the $p_{ij}^{[k]}$ values is the determination of $\Delta f_{ji}^*$ for each neighbor $j$ of solution $i$, hence the objective function values for the neighbors as well as for the current solution. This is true for both the generic configuration spaces and the combinatorial optimization problems described above. Before these objective function values are determined, however, it is necessary to identify which of the solutions in the solution space are the neighbors of solution $i$. 
For the generic configuration spaces, the neighbors are easily determined from the list of solutions as described in Subsection 5.1.1.2. Thus, the $N$ neighbors of solution $i$ are simply the $N$ values in the list that are directly adjacent (in both directions) to the current solution.

For the combinatorial optimization problems, a list of solutions, similar to that in the generic annealing experiments, must be created. This is done by representing each solution's position in a list using binary notation. This serves as an address that is used to lookup the objective function value of a solution. The number of binary digits corresponds to the number of nodes in the graphs or boolean variables (note that this is also the neighborhood size). Thus, the number of solutions is $2^N$ where $N$ is the neighborhood size. Using this approach, the first solution in the list is represented by 00000000 in a solution space of size 256 where each solution has 8 neighbors.

Each of the digits in this binary notation correspond to a node in a graph or a boolean variable, as the case may be. The value of each digit, that is, whether it is a one or zero, is based on whether the corresponding node (or boolean variable) is in the current solution set (the clique, independent set, etc.). Thus, the binary number represents a combination of nodes or, in the case of the 3-SAT problem, boolean variables that have truth value of one or zero, that corresponds to a particular solution or address in the computer memory. The objective function value for a particular binary representation is calculated and that value placed in a list in the position corresponding to the value of the binary number. For example, suppose the combination of nodes represented by 00100100 has an objective function value of 165. Then the number 165 is placed in the 37th position in the solution list because the binary number corresponds to the value 36 in decimal notation (remember the 0th value in the list is first).

Once the solution list is created in the computer memory, the determination of neighboring solutions proceeds by determining the address of solutions that correspond to neighbors of the current solution. This is accomplished by flipping bits that correspond to nodes or variables, that is,
changing a 0 to a 1 and a 1 to 0 in each digit's position. In the binary example above, where the current solution is 00100100, the following constitute all of its neighbors:

<table>
<thead>
<tr>
<th>Neighbors to Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>00100100</td>
</tr>
<tr>
<td>00100111</td>
</tr>
<tr>
<td>00100000</td>
</tr>
<tr>
<td>01101100</td>
</tr>
<tr>
<td>01101000</td>
</tr>
<tr>
<td>01001100</td>
</tr>
<tr>
<td>01000100</td>
</tr>
<tr>
<td>11001100</td>
</tr>
<tr>
<td>10100100</td>
</tr>
</tbody>
</table>

Table 1

These solution addresses are determined by flipping the first bit for the first neighbor, the second bit for the second neighbor, and so on. Thus, the neighbors of solution #36 have the addresses: 37, 38, 32, 44, 52, 4, 100, and 164, respectively. Each of these neighbors has some objective function value, which is placed in the list in the position that corresponds to the binary value (plus 1) of its solution address.

These neighbors are determined in code by performing an exclusive-or (XOR) operation on the bits representing the current solution. For example, the third neighbor is obtained by putting a 1 in the 3rd binary position of a masking byte of data and performing the XOR instruction. This is indicated by the following:

\[
\begin{align*}
30100100 & \quad \text{current solution} \\
\oplus \quad 00000100 & \quad \text{3rd neighbor} \\
\Rightarrow & \quad 30100000
\end{align*}
\]

Thus, the 3rd neighbor of the current solution is solution #32. Its objective function value is placed in the 33rd place in the solution list.

Once the solution space list is created and these neighboring solutions are determined, it is a simple matter to determine the values of \( \Delta f^* \). The objective function value for the neighboring solution is looked up in the list after its position in the list is determined, and the objective function value of the current solution is then subtracted from it. If the difference is less than or equal to zero,
then $\Delta_{ji}^* = 0$; otherwise, it is the difference. The transition probability as defined in (1.2) can then be determined. In executing the software that uses the above methods, the solution list is determined only once and stored in memory. The transition probabilities are then determined as needed.

5.4.2. Calculation of the State Probabilities

The calculation of the state probabilities depends on the values of the transition probabilities. In effect, the state probabilities are updated by vector-matrix multiplication, that is, $v(m+1) = v(m)P(m)$. The state vector $v(m)$ is essentially an array or list in memory. The matrix $P(m)$ is not, however, stored in memory as such. The transition probabilities are calculated as needed, as indicated in Subsection 5.4.1. In performing the vector-matrix multiplication, note that a column of the transition matrix constitutes transition probabilities where the transition is to the current solution. Consider the following in which the current solution is state $i$.

$$
\begin{pmatrix}
\begin{array}{cccc}
    v_1^{(11)} & v_2^{(11)} & \cdots & v_{n1}^{(11)} \\
    p_{11}^{(11)} & p_{12}^{(11)} & \cdots & p_{1n}^{(11)} \\
    \vdots & \vdots & \ddots & \vdots \\
    p_{n1}^{(11)} & p_{n2}^{(11)} & \cdots & p_{nn}^{(11)}
\end{array}
\end{pmatrix}
\begin{pmatrix}
    v_1^{(11)} \\
    v_2^{(11)} \\
    \vdots \\
    v_{n1}^{(11)}
\end{pmatrix}
$$

Each state probability $v_i^{(11)}$ is updated to $v_i^{(11)}$ by taking the dot product of the state distribution vector and the $i^{th}$ column of the transition matrix. The resulting value is stored in a temporary array. After all elements have been updated, the temporary array is written over the old probability values.

This process is facilitated by the fact that the neighborhood structure is symmetric, that is, $j \in N(i) \Leftrightarrow i \in N(j)$. By using the methods in the Subsection 5.4.1, the neighbors of solution $i$ and their addresses can easily be determined. Once the $n^{th}$ neighbor’s address $\alpha_n$ and the objective function values are determined, the value of $\Delta_{ij}^* = \max \{ 0, f_i - f_j \}$ is determined, the corresponding
transition probability is calculated and then multiplied by the element in the $\alpha_i$th position in the vector $v(m)$. Note that the column of $P(m)$ is based on the values of $\Delta f_{ij}^+$, as opposed to the rows which are based on the values of $\Delta f_{ji}^+$ (the $i$ and $j$ are switched—this follows from the symmetry described above). The C code in Appendix B implements these ideas for updating the state probabilities. Once the state probabilities are determined, it is a simple matter to calculate the entropy $H(k)$ at time index $k$.

5.5. CALCULATION OF THE EXPECTED VALUE, VARIANCE, AND NORM

In Chapter 4, we indicated that one measure of performance of the SA algorithm is to determine how close the final state of an annealing experiment is to the globally optimal state. In effect, the final state serves as an estimator of the global optimum. In Section 4.4 the issues concerning the expected value and variance of this estimator were discussed. In this section, the methods used calculating these values for the generic and the combinatorial optimization problems is described.

5.5.1. The Expected Value and Variance

Recall from Section 4.3 that when the optimal objective function value is zero, the bias is simply the expected value of the final state, $E(X)$, where $X$ is a random objective function value of the final state. Rather than perform experiments and keep track of the random variable corresponding to the final state, the expected value can be determined exactly by using the the final state distribution vector $v(n-1)$, along with the solution list. These data contain all the information needed to calculate the bias and variance of the final state. Thus, the bias is

$$\text{Bias} = E(X) = \sum_{i=0}^{N-1} v_i^{(n-1)} f(i)$$  \hspace{1cm} (19)

where the $i$th objective function value, $f(i)$, is found in the solution list at address $i$. 
The calculation of the variance proceeds by first determining $E(X^2)$. This is accomplished by substituting $[f(i)]^2$ for $f(i)$ in (19). Once this is done, the variance is given by $\text{Var}(X) = E(X^2) - E(X)$.

5.5.2. The Norm

The rate of convergence shown by Mitra et al. [1986] is given in terms of the $L^1$ norm. This norm is used in the experiments described in Chapter 6 to measure of the distance between the state distribution vector and the optimum vector, that is, the vector in which all the probability is equally distributed among the globally optimal solutions. The $L^1$ norm used here is defined as

$$\|v(n-1) - q\| = \sum_{i=0}^{s-1} |v_i^{n-1} - q_i|$$

where $q = (q_0, q_1, \ldots, q_{n-1})$ is the optimum vector. This expression requires a determination of the optimum vector. This can be done by scanning the solution list and keeping track of the smallest values (or the largest as the case may be) and the number of times this value occurs. Thus, for some problems several optima exist. If $s^*$ is the set of globally optimal solutions, then the value for the $q_i$ that correspond to the optima is $1/s^*$ and all the other $q_i$ are all 0. Thus, (20) can be restated as

$$\|v(n-1) - q\| = \sum_{i \in s^*} |v_i^{n-1} - \frac{1}{s^*}| + \sum_{i \notin s^*} |v_i^{n-1}|.$$

These measures, the expected value, variance, and norm, will serve as a basis for comparing the performance of SA (see Chapter 6) after changes to the configuration space are implemented. Since these changes will also change the entropy measures of these problems, the effect of changes on the configuration space as measured by the entropy will be compared to these performance measures and illustrate the theory developed in Chapters 3 and 4.
CHAPTER 6

EXPERIMENTAL RESULTS

This chapter presents a number of computational results based on the methodology presented in Chapter 5. The experiments are based on calculations involving the Markov chain that models the SA algorithm. The state distribution vector is computed exactly, using vector-matrix multiplications rather than approximated by a series of actual simulation experiments and determining the frequency distribution of the final states. This exact approach effectively computes the state distribution vector of the inhomogeneous Markov chain that models the SA algorithm. Since the state distribution vector contains all the information regarding the behavior of SA at each time index, it can be used in calculating the performance measures for SA as well as to assess the relationship between such performance measures and the entropy measures.

The two major sections of this chapter describe the generic annealing experiments and the transformation experiments, respectively. A series of computational results and appropriate descriptions and analyses are presented in each subsection. These experiments were done using programs written in the C programming language (Kernighan and Ritchie [1978]) and compiled under the UNIX operating system on a SUN SLC Sparestation.

6.1. GENERIC EXPERIMENTS

This section describes the results of the generic annealing experiments. Two different generic solution spaces were used: one with objective function values ranging from 0 to 100 (denoted as solution space A), the second with objective function values ranging from 0 to 500 (denoted as solution space B). Both of these solution spaces have global minimum objective function values of 0. These objective function values were made to make comparisons of the performance measures and entropy measures between the two solution spaces. The larger range of objective function values in
the solution space B should result in the average value of $\Delta f$ being greater (compared to solution space A) and hence, affect the transition probabilities and entropy measures differently. From Subsection 3.2.1, the larger values of $\Delta f$ should decrease the entropy measure and hence, the performance of SA, given identical cooling schedules.

To demonstrate the convergence properties associated with higher entropy values and longer simulation runs, as described in Subsection 4.5.1, three experiments were performed using each solution space. Each of these experiments involved a larger number of iterations. It should be noted that the computer output indicates the three parameters (the number of iterations, the initial temperature, and the final temperature) used in the computations. These parameters are used to determine identical cooling schedule parameters (see Section 5.3) and do not necessarily reflect the actual number of iterations in the SA algorithm that are executed. The last two experiments for each solution space do, however, execute a larger number of iterations (the program was recompiled to do this). This number of iterations is clearly stated. The main routines from the code that generated these results are presented in Appendix B.

6.1.1. Solution Space A

The computer outputs for three experiments are presented below. Table 1 indicates the data obtained for the first experiment, where the number of actual iterations was 100. The time index $m$ indicates the iteration number that the entropy measure $H^*(m, n-2)$ is calculated from. The number of data points is 25 meaning that 25 values of neighborhood sizes—from 2 to 50 in multiples of 2—were computed. The values under N, H, EX, VAR, and NORM indicate the neighborhood size, the entropy measure $H^*(m, n-2)$, the expected value of the final state, the variance of the final state, and the $L^1$ norm, respectively (see Subsections 5.5.1-2). Figure 1 below graphs the entropy measure and the expected value of the final state as a function of the neighborhood size.

Computer Output for Experiment A1

The number of solutions is 5000
The number of iterations is 100
The number of data points is 25

The initial temperature is 10.000000
The cutoff temperature is 1.000000
The time index m is 80

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.724</td>
<td>20.818</td>
<td>295.019</td>
<td>1.927</td>
</tr>
<tr>
<td>4</td>
<td>0.772</td>
<td>13.673</td>
<td>153.275</td>
<td>1.888</td>
</tr>
<tr>
<td>6</td>
<td>1.016</td>
<td>9.925</td>
<td>87.052</td>
<td>1.828</td>
</tr>
<tr>
<td>8</td>
<td>1.229</td>
<td>7.747</td>
<td>59.401</td>
<td>1.794</td>
</tr>
<tr>
<td>10</td>
<td>1.234</td>
<td>6.172</td>
<td>42.493</td>
<td>1.753</td>
</tr>
<tr>
<td>12</td>
<td>1.27</td>
<td>5.237</td>
<td>32.394</td>
<td>1.724</td>
</tr>
<tr>
<td>14</td>
<td>1.306</td>
<td>4.591</td>
<td>26.02</td>
<td>1.689</td>
</tr>
<tr>
<td>16</td>
<td>1.572</td>
<td>4.092</td>
<td>21.269</td>
<td>1.65</td>
</tr>
<tr>
<td>18</td>
<td>1.639</td>
<td>3.693</td>
<td>16.596</td>
<td>1.617</td>
</tr>
<tr>
<td>20</td>
<td>1.696</td>
<td>3.324</td>
<td>12.502</td>
<td>1.595</td>
</tr>
<tr>
<td>22</td>
<td>1.774</td>
<td>3.02</td>
<td>11.064</td>
<td>1.553</td>
</tr>
<tr>
<td>24</td>
<td>1.788</td>
<td>2.866</td>
<td>10.081</td>
<td>1.528</td>
</tr>
<tr>
<td>26</td>
<td>1.862</td>
<td>2.652</td>
<td>8.363</td>
<td>1.498</td>
</tr>
<tr>
<td>28</td>
<td>1.896</td>
<td>2.545</td>
<td>8.125</td>
<td>1.464</td>
</tr>
<tr>
<td>30</td>
<td>1.974</td>
<td>2.36</td>
<td>6.967</td>
<td>1.429</td>
</tr>
<tr>
<td>32</td>
<td>2.084</td>
<td>2.245</td>
<td>6.376</td>
<td>1.419</td>
</tr>
<tr>
<td>34</td>
<td>2.109</td>
<td>2.173</td>
<td>6.182</td>
<td>1.402</td>
</tr>
<tr>
<td>36</td>
<td>2.073</td>
<td>2.112</td>
<td>5.738</td>
<td>1.398</td>
</tr>
<tr>
<td>38</td>
<td>2.099</td>
<td>2.052</td>
<td>5.54</td>
<td>1.38</td>
</tr>
<tr>
<td>40</td>
<td>2.209</td>
<td>1.97</td>
<td>5.229</td>
<td>1.361</td>
</tr>
<tr>
<td>42</td>
<td>2.196</td>
<td>1.91</td>
<td>4.902</td>
<td>1.338</td>
</tr>
<tr>
<td>44</td>
<td>2.22</td>
<td>1.85</td>
<td>4.73</td>
<td>1.318</td>
</tr>
<tr>
<td>46</td>
<td>2.265</td>
<td>1.762</td>
<td>4.284</td>
<td>1.295</td>
</tr>
<tr>
<td>48</td>
<td>2.404</td>
<td>1.719</td>
<td>4.026</td>
<td>1.288</td>
</tr>
<tr>
<td>50</td>
<td>2.417</td>
<td>1.666</td>
<td>3.745</td>
<td>1.273</td>
</tr>
</tbody>
</table>

Table 1
Expected Value and Entropy Versus Neighborhood Size

![Graph](image)

Figure 1

Notice that as the neighborhood size increases, the expected value of the final state and the entropy measure increases. Note that the data in Table 1 indicates that this entropy measure does not increase monotonically as suggested in Subsection 3.2.1. Notice also that the rate at which the expected value of the final state decreases diminishes with increasing neighborhood size. This type of change in rate is also apparent in the entropy measure; the rate of increase in the entropy also diminishes, although this effect is less apparent than it is in the expected value of the final state.

6.1.1.2. Experiment A2.

This experiment involves the same solution space and cooling schedule as for Experiment A1. The only difference is that the number of iterations executed by SA has been increased from 100 to 150 (again, the parameters indicated at the beginning of the computer output indicate only 100 iterations, but this number is used only in determining the cooling schedule's parameters). Figure 2 graphs the entropy measure and the expected value of the final state as a function of the neighborhood size.
Computer Output for Experiment A2

The number of solutions is 5000
The number of iterations is 100
The number of data points is 25

The initial temperature is 10.000000
The cutoff temperature is 1.000000
The time index m is 80

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.396</td>
<td>20.784</td>
<td>294.161</td>
<td>1.927</td>
</tr>
<tr>
<td>4</td>
<td>2.501</td>
<td>13.63</td>
<td>152.599</td>
<td>1.888</td>
</tr>
<tr>
<td>6</td>
<td>3.311</td>
<td>9.865</td>
<td>86.307</td>
<td>1.827</td>
</tr>
<tr>
<td>8</td>
<td>4.053</td>
<td>7.683</td>
<td>58.822</td>
<td>1.792</td>
</tr>
<tr>
<td>10</td>
<td>4.059</td>
<td>6.102</td>
<td>41.769</td>
<td>1.75</td>
</tr>
<tr>
<td>12</td>
<td>4.12</td>
<td>5.152</td>
<td>31.792</td>
<td>1.721</td>
</tr>
<tr>
<td>14</td>
<td>4.203</td>
<td>4.48</td>
<td>25.177</td>
<td>1.685</td>
</tr>
<tr>
<td>16</td>
<td>5.099</td>
<td>3.97</td>
<td>20.424</td>
<td>1.641</td>
</tr>
<tr>
<td>18</td>
<td>5.31</td>
<td>3.573</td>
<td>15.932</td>
<td>1.604</td>
</tr>
<tr>
<td>20</td>
<td>5.482</td>
<td>3.187</td>
<td>11.73</td>
<td>1.578</td>
</tr>
<tr>
<td>22</td>
<td>5.71</td>
<td>2.877</td>
<td>10.352</td>
<td>1.531</td>
</tr>
<tr>
<td>24</td>
<td>5.747</td>
<td>2.72</td>
<td>9.457</td>
<td>1.5</td>
</tr>
<tr>
<td>26</td>
<td>5.955</td>
<td>2.476</td>
<td>7.613</td>
<td>1.462</td>
</tr>
<tr>
<td>28</td>
<td>6.051</td>
<td>2.358</td>
<td>7.371</td>
<td>1.42</td>
</tr>
<tr>
<td>30</td>
<td>6.29</td>
<td>2.152</td>
<td>6.129</td>
<td>1.379</td>
</tr>
<tr>
<td>32</td>
<td>6.659</td>
<td>2.025</td>
<td>5.477</td>
<td>1.366</td>
</tr>
<tr>
<td>34</td>
<td>6.728</td>
<td>1.946</td>
<td>5.272</td>
<td>1.345</td>
</tr>
<tr>
<td>36</td>
<td>6.587</td>
<td>1.893</td>
<td>4.877</td>
<td>1.341</td>
</tr>
<tr>
<td>38</td>
<td>6.661</td>
<td>1.835</td>
<td>4.718</td>
<td>1.322</td>
</tr>
<tr>
<td>40</td>
<td>7.041</td>
<td>1.743</td>
<td>4.409</td>
<td>1.297</td>
</tr>
<tr>
<td>42</td>
<td>6.982</td>
<td>1.673</td>
<td>3.975</td>
<td>1.272</td>
</tr>
<tr>
<td>44</td>
<td>7.047</td>
<td>1.605</td>
<td>3.825</td>
<td>1.245</td>
</tr>
<tr>
<td>46</td>
<td>7.202</td>
<td>1.508</td>
<td>3.363</td>
<td>1.217</td>
</tr>
<tr>
<td>48</td>
<td>7.688</td>
<td>1.457</td>
<td>3.095</td>
<td>1.206</td>
</tr>
<tr>
<td>50</td>
<td>7.724</td>
<td>1.398</td>
<td>2.803</td>
<td>1.187</td>
</tr>
</tbody>
</table>

Table 2
Expected Value and Entropy Versus Neighborhood Size

Notice that the expected value of the final state for each neighborhood size is slightly lower for Experiment A2 relative to Experiment A1. This is consistent with the basic theory in SA, that longer simulation runs yield better results. In Experiment A2, the number of executed iterations has been increased by 50% over the number of executed iterations in Experiment A1. Also notice that the corresponding entropy values are larger in Experiment A2 than in Experiment A1, owing to the larger number of iterations. The variance and norm also indicate improved performance for the longer SA run. Finally, notice that the general shape of the entropy curve in Figure 2 closely mirrors the shape of the curve depicting the expected value of the final state.

6.1.1.3. Experiment A3.

This experiment again involves the same solution space and cooling schedule as the previous two experiments: the only difference is that the number of executed iterations has been increased from 100 to 500. Table 3 contains the data while Figure 3 graphs the entropy measure and the expected value of the final state as a function of the neighborhood size. Also, in Figure 3, the entropy measure was rescaled by dividing it by 2 in order to make visual comparisons.
Computer Output for Experiment A3

The number of solutions is 5000
The number of iterations is 100
The number of data points is 25

The initial temperature is 10.000000
The cutoff temperature is 1.000000
The time index m is 80

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12.056</td>
<td>20.724</td>
<td>292.584</td>
<td>1.927</td>
</tr>
<tr>
<td>4</td>
<td>11.75</td>
<td>13.564</td>
<td>151.517</td>
<td>1.887</td>
</tr>
<tr>
<td>6</td>
<td>15.84</td>
<td>9.761</td>
<td>84.997</td>
<td>1.825</td>
</tr>
<tr>
<td>8</td>
<td>20.524</td>
<td>7.57</td>
<td>57.887</td>
<td>1.787</td>
</tr>
<tr>
<td>10</td>
<td>20.465</td>
<td>5.968</td>
<td>40.35</td>
<td>1.745</td>
</tr>
<tr>
<td>12</td>
<td>19.878</td>
<td>4.992</td>
<td>30.605</td>
<td>1.713</td>
</tr>
<tr>
<td>14</td>
<td>19.914</td>
<td>4.308</td>
<td>23.958</td>
<td>1.677</td>
</tr>
<tr>
<td>16</td>
<td>24.962</td>
<td>3.791</td>
<td>19.289</td>
<td>1.627</td>
</tr>
<tr>
<td>18</td>
<td>26.009</td>
<td>3.415</td>
<td>15.222</td>
<td>1.583</td>
</tr>
<tr>
<td>20</td>
<td>26.519</td>
<td>2.976</td>
<td>10.65</td>
<td>1.546</td>
</tr>
<tr>
<td>22</td>
<td>27.382</td>
<td>2.666</td>
<td>9.367</td>
<td>1.493</td>
</tr>
<tr>
<td>24</td>
<td>27.626</td>
<td>2.509</td>
<td>8.66</td>
<td>1.448</td>
</tr>
<tr>
<td>26</td>
<td>28.406</td>
<td>2.238</td>
<td>6.645</td>
<td>1.405</td>
</tr>
<tr>
<td>28</td>
<td>28.52</td>
<td>2.104</td>
<td>6.409</td>
<td>1.345</td>
</tr>
<tr>
<td>30</td>
<td>29.442</td>
<td>1.877</td>
<td>5.265</td>
<td>1.288</td>
</tr>
<tr>
<td>32</td>
<td>31.574</td>
<td>1.71</td>
<td>4.243</td>
<td>1.27</td>
</tr>
<tr>
<td>34</td>
<td>31.797</td>
<td>1.621</td>
<td>3.991</td>
<td>1.243</td>
</tr>
<tr>
<td>36</td>
<td>30.82</td>
<td>1.592</td>
<td>3.81</td>
<td>1.236</td>
</tr>
<tr>
<td>38</td>
<td>31.125</td>
<td>1.553</td>
<td>3.769</td>
<td>1.213</td>
</tr>
<tr>
<td>40</td>
<td>33.361</td>
<td>1.452</td>
<td>3.477</td>
<td>1.183</td>
</tr>
<tr>
<td>42</td>
<td>32.9</td>
<td>1.371</td>
<td>2.918</td>
<td>1.161</td>
</tr>
<tr>
<td>44</td>
<td>32.968</td>
<td>1.269</td>
<td>2.7</td>
<td>1.12</td>
</tr>
<tr>
<td>46</td>
<td>33.875</td>
<td>1.165</td>
<td>2.162</td>
<td>1.091</td>
</tr>
<tr>
<td>48</td>
<td>36.678</td>
<td>1.122</td>
<td>2.025</td>
<td>1.075</td>
</tr>
<tr>
<td>50</td>
<td>36.808</td>
<td>1.058</td>
<td>1.797</td>
<td>1.047</td>
</tr>
</tbody>
</table>

Table 3
The relationship between the expected value of the final state and entropy continue to indicate that, in general, as entropy increases, the expected value of the final state decreases. Notice that for given values of the neighborhood size, the values of the expected value are lower and the entropy measure higher in this experiment than in either Experiments A1 or A2 due to the larger number of iterations. The next set of experiments show the effect of larger values of $\Delta f$. These values are, on average, larger in solution space B relative to solution space A because of the larger range of objective function values.

6.1.2. Solution Space B

The following are the computer outputs for three experiments using solution space B. Although these experiments are run using a different solution space, the cooling schedule's parameters are the same as in Experiments A1, A2, and A3. The experiments in this subsection are, therefore run in the same manner as described for the experiments in Subsection 6.1.1, i.e., the number of executed iterations for experiments B1, B2, and B3 were 100, 150, and 500, respectively.

Table 4 contains the computer generated data while Figure 4 below graphs the entropy measure and
the expected value of the final state as a function of the neighborhood size. The entropy values were rescaled by multiplying them by 100 in order to facilitate visual comparison.

### 6.1.2.1. Experiment B1.

**Computer Output for Experiment B1**

<table>
<thead>
<tr>
<th>The number of solutions is</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of iterations is</td>
<td>100</td>
</tr>
<tr>
<td>The number of data points is</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The initial temperature is</th>
<th>10,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>The cutoff temperature is</td>
<td>1,000,000</td>
</tr>
<tr>
<td>The time index m is</td>
<td>80</td>
</tr>
<tr>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
</tr>
<tr>
<td>2</td>
<td>0.198</td>
</tr>
<tr>
<td>4</td>
<td>0.173</td>
</tr>
<tr>
<td>6</td>
<td>0.198</td>
</tr>
<tr>
<td>8</td>
<td>0.267</td>
</tr>
<tr>
<td>10</td>
<td>0.263</td>
</tr>
<tr>
<td>12</td>
<td>0.26</td>
</tr>
<tr>
<td>14</td>
<td>0.317</td>
</tr>
<tr>
<td>16</td>
<td>0.346</td>
</tr>
<tr>
<td>18</td>
<td>0.412</td>
</tr>
<tr>
<td>20</td>
<td>0.431</td>
</tr>
<tr>
<td>22</td>
<td>0.453</td>
</tr>
<tr>
<td>24</td>
<td>0.465</td>
</tr>
<tr>
<td>26</td>
<td>0.492</td>
</tr>
<tr>
<td>28</td>
<td>0.535</td>
</tr>
<tr>
<td>30</td>
<td>0.545</td>
</tr>
<tr>
<td>32</td>
<td>0.628</td>
</tr>
<tr>
<td>34</td>
<td>0.659</td>
</tr>
<tr>
<td>36</td>
<td>0.666</td>
</tr>
<tr>
<td>38</td>
<td>0.736</td>
</tr>
<tr>
<td>40</td>
<td>0.719</td>
</tr>
<tr>
<td>42</td>
<td>0.748</td>
</tr>
<tr>
<td>44</td>
<td>0.756</td>
</tr>
<tr>
<td>46</td>
<td>0.767</td>
</tr>
<tr>
<td>48</td>
<td>0.773</td>
</tr>
<tr>
<td>50</td>
<td>0.816</td>
</tr>
</tbody>
</table>

Table 4
From Figure 4, notice that the trend of the entropy measures again increases with increasing neighborhood size while the expected value of the final state tends to decrease. The curve depicting the entropy measure is also more jagged than in the experiments in Section 6.1.1 illustrating that the entropy measure does not increase monotonically with neighborhood size as discussed in Section 3.2.1.

Besides showing the general relationship between expected values of the final state and entropy via changes to the configurations caused by changes in the neighborhood size, this relationship is also apparent in comparing the data in Experiments A1 and B1. Notice that the entropy measures for each neighborhood size in Experiment B1 are less than the corresponding entropy measures in Experiment A1. This is due to the larger values for $\Delta f$. This lower entropy configuration space also leads to poorer performance in that the expected values of the final state for each neighborhood size in Experiment A1 are closer to the global minimum than the corresponding expected values in Experiment B1.
The next experiment was run using the same solution space and cooling schedule parameters as in Experiment B1, but the number of executed iterations was increased from 100 to 150. Figure 5 below graphs the entropy measure and the expected value of the final state as a function of the neighborhood size. The entropy measures were rescaled by a factor of 50.

6.1.2.2. Experiment B2.

Computer Output for Experiment B2

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>H</td>
<td>EX</td>
<td>VAR</td>
<td>NORM</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.662</td>
<td>109.839</td>
<td>8006.421</td>
<td>1.981</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.564</td>
<td>71.034</td>
<td>3918.052</td>
<td>1.966</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.645</td>
<td>53.811</td>
<td>2528.321</td>
<td>1.953</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.892</td>
<td>41.251</td>
<td>1453.677</td>
<td>1.946</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.865</td>
<td>33.696</td>
<td>1002.624</td>
<td>1.931</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.842</td>
<td>28.787</td>
<td>766.996</td>
<td>1.906</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1.004</td>
<td>24.596</td>
<td>560.724</td>
<td>1.898</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>1.077</td>
<td>22.184</td>
<td>484.946</td>
<td>1.88</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1.258</td>
<td>20.191</td>
<td>404.082</td>
<td>1.865</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.289</td>
<td>18.82</td>
<td>347.389</td>
<td>1.855</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>1.327</td>
<td>17.499</td>
<td>301.621</td>
<td>1.854</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>1.323</td>
<td>15.587</td>
<td>249.335</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>1.374</td>
<td>14.602</td>
<td>230.439</td>
<td>1.826</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>1.492</td>
<td>13.733</td>
<td>205.184</td>
<td>1.825</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1.517</td>
<td>13.114</td>
<td>194.988</td>
<td>1.817</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>1.752</td>
<td>12.211</td>
<td>182.727</td>
<td>1.798</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>1.843</td>
<td>11.48</td>
<td>159.766</td>
<td>1.797</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>1.858</td>
<td>11.051</td>
<td>142.343</td>
<td>1.801</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>2.083</td>
<td>10.611</td>
<td>133.846</td>
<td>1.794</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1.993</td>
<td>10.254</td>
<td>132.604</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>2.081</td>
<td>9.787</td>
<td>112.904</td>
<td>1.763</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>2.093</td>
<td>9.488</td>
<td>107.567</td>
<td>1.754</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>2.111</td>
<td>9.193</td>
<td>103.535</td>
<td>1.748</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>2.129</td>
<td>9.235</td>
<td>104.171</td>
<td>1.744</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>2.254</td>
<td>8.869</td>
<td>95.438</td>
<td>1.738</td>
</tr>
</tbody>
</table>

Table 5
In Figure 5, the entropy curve is higher than in Figure 4 because of the larger number of iterations. A comparison of the data in Tables 4 and 5 reveals that the expected values of the final state have decreased slightly for each value of the neighborhood size due to the larger number of iterations in Experiment B2 relative to B1. Also, the entropy measures have increased for each neighborhood size also due to the larger number of iterations. The general trend indicating of better performance associated with increases in neighborhood size is matched by generally higher entropy values. Furthermore, comparison of the data generated for the two configuration spaces in Experiment A2 and Experiment B2 also shows that the configuration space in A2 is associated with higher entropy values relative to B2 for each neighborhood size. Again, the performance, as measured by the expected value of the final state, is associated with the higher entropy configuration space in Experiment A2. For example, for a neighborhood size of 16, the entropy in A2 is 5.099 versus 1.077 in B2, while the expected value of the final state in A2 is 3.97, versus 22.184 in B2, much worse than in A2. Thus, the configuration space with higher entropy values shows superior performance.
In the last experiment of this subsection, the number of iterations has been increased to 500.

Figure 6 below graphs the entropy measure (× 10) and the expected value of the final state as a function of the neighborhood size.

6.1.2.3. Experiment B3.

Computer Output for Experiment B3

The number of solutions is 5000
The number of iterations is 100
The number of data points is 25

The initial temperature is 10.00000
The cutoff temperature is 1.00000
The time index m is 80

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.436</td>
<td>109.787</td>
<td>7995.145</td>
<td>1.981</td>
</tr>
<tr>
<td>4</td>
<td>2.739</td>
<td>70.973</td>
<td>3913.786</td>
<td>1.966</td>
</tr>
<tr>
<td>6</td>
<td>3.14</td>
<td>53.734</td>
<td>2524.62</td>
<td>1.952</td>
</tr>
<tr>
<td>8</td>
<td>4.713</td>
<td>41.175</td>
<td>1444.911</td>
<td>1.946</td>
</tr>
<tr>
<td>10</td>
<td>4.456</td>
<td>33.648</td>
<td>997.191</td>
<td>1.93</td>
</tr>
<tr>
<td>12</td>
<td>4.247</td>
<td>28.749</td>
<td>762.337</td>
<td>1.905</td>
</tr>
<tr>
<td>14</td>
<td>4.898</td>
<td>24.422</td>
<td>549.696</td>
<td>1.895</td>
</tr>
<tr>
<td>16</td>
<td>5.134</td>
<td>22.015</td>
<td>474.45</td>
<td>1.876</td>
</tr>
<tr>
<td>18</td>
<td>5.868</td>
<td>20.016</td>
<td>393.753</td>
<td>1.862</td>
</tr>
<tr>
<td>20</td>
<td>5.834</td>
<td>18.691</td>
<td>341.199</td>
<td>1.85</td>
</tr>
<tr>
<td>22</td>
<td>5.787</td>
<td>17.311</td>
<td>293.602</td>
<td>1.849</td>
</tr>
<tr>
<td>24</td>
<td>5.403</td>
<td>15.412</td>
<td>247.354</td>
<td>1.823</td>
</tr>
<tr>
<td>26</td>
<td>5.417</td>
<td>14.436</td>
<td>227.85</td>
<td>1.82</td>
</tr>
<tr>
<td>28</td>
<td>5.778</td>
<td>13.556</td>
<td>203.665</td>
<td>1.818</td>
</tr>
<tr>
<td>30</td>
<td>5.867</td>
<td>12.894</td>
<td>192.336</td>
<td>1.812</td>
</tr>
<tr>
<td>32</td>
<td>6.659</td>
<td>11.958</td>
<td>180.37</td>
<td>1.79</td>
</tr>
<tr>
<td>34</td>
<td>6.874</td>
<td>11.112</td>
<td>154.674</td>
<td>1.79</td>
</tr>
<tr>
<td>36</td>
<td>6.9</td>
<td>10.573</td>
<td>132.585</td>
<td>1.793</td>
</tr>
<tr>
<td>38</td>
<td>7.948</td>
<td>10.203</td>
<td>128.369</td>
<td>1.778</td>
</tr>
<tr>
<td>40</td>
<td>7.066</td>
<td>9.793</td>
<td>129.108</td>
<td>1.745</td>
</tr>
<tr>
<td>42</td>
<td>7.387</td>
<td>9.272</td>
<td>107.894</td>
<td>1.741</td>
</tr>
<tr>
<td>44</td>
<td>7.319</td>
<td>8.895</td>
<td>102.409</td>
<td>1.724</td>
</tr>
<tr>
<td>46</td>
<td>7.298</td>
<td>8.557</td>
<td>97.441</td>
<td>1.717</td>
</tr>
<tr>
<td>48</td>
<td>7.354</td>
<td>8.593</td>
<td>98.846</td>
<td>1.711</td>
</tr>
<tr>
<td>50</td>
<td>7.616</td>
<td>8.085</td>
<td>86.335</td>
<td>1.705</td>
</tr>
</tbody>
</table>

Table 6
A comparison of the data in Tables 4 and 6 reveals even further improvements in performance (the differences in these measures has increased) as well as increases in entropy measures. Comparing the data from Experiment A3 with that of B3, the same relationships hold as in the comparisons of A1 and B1 and A2 and B2. The experiments using solution space B had lower entropy values for each neighborhood size than the corresponding experiments using solution space A where the cooling schedules and number of executed iterations were the same. Furthermore, the experiments using solution space B showed poorer performance as measure by the expected value of the final state than the performance in the corresponding experiments using solution space A.

In each of these experiments, when the expected value of the final state and the entropy were plotted against neighborhood size, improved performance was associated with increased neighborhood size and a trend of increasing entropy values. In the next subsection, combinatorial experiments are used to demonstrate similar relationships between the expected value of the final state and entropy.
6.2. COMBINATORIAL EXPERIMENTS

These experiments are based on the methodology described in Chapter 5 regarding the 3-SAT and clique problems. Two sets of clauses were used to run the 3-SAT annealing experiments. Each of these were then transformed to their related clique problems by the method described in Subsection 5.2.4.2. The SA algorithm was then applied to these problems using identical cooling schedules making it possible to compare the performance of SA on these two problems.

Another method to illustrate the theory in Chapter 5 that relates entropy to the expected value of the final state, is to create several solution spaces for the clique problems using different values of the penalty parameter \( \lambda \). Recall from Subsection 5.2.2.1, the purpose of using a penalty parameter is to facilitate the construction of the configuration space by allowing for arcs that violate the constraints of the problem. Inclusion of such arcs in the clique problem have a penalty associated with it. Thus, a candidate solution that has such arcs has an objective function value that is functionally dependent on this penalty parameter. Thus different values of this penalty parameter change the transition probabilities of the associated Markov chain and, hence, change the configuration space. These experiments therefore not only show the effect of transformations on the configuration space, but also the effect of the penalty parameters on the configuration space. These two methods of changing a configuration space provide further avenues to illustrate the theory that relates the entropy measure to the expected value of the final state. The following discussion describes all the details used in performing these experiments.

6.2.1. Transformation Experiment A

This set of clauses was created using random number schemes similar to that used in creating the generic solution space. Four clauses were used which translates to a clique problem with \( 2^{12} = 4096 \) solutions—comparable in size to the generic solution space. For this set of clauses, 12 boolean variables were used to create a solution space similar in size to the generic problems. The following table give the values obtained from running the experiment in a similar fashion as the
previous tables in Section 6.1. The only difference is that the neighborhood size is fixed (this is why a transformation is done) for a given 3-SAT problem instance. In this case, the neighborhood size is equal to the number of boolean variables (see Subsections 5.2.4.2. and 5.4.1). The following tables contain the computer output for the 3-SAT problem and its associated clique problem.

**Computer Output for 3-SAT A**

| The number of clauses is | 4 |
| The number of variables is | 12 |
| The number of iterations is | 10 |
| The initial temperature is | 1.000000 |
| The cutoff temperature is | 0.010000 |
| The time index m is | 6 |

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>Var</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>9.55</td>
<td>3.96</td>
<td>0.04</td>
<td>0.121</td>
</tr>
</tbody>
</table>

Table 7

**Computer Output for Clique A**

| The number of solutions is | 4096 |
| The number of nodes is | 12 |
| The number of iterations is | 10 |
| The initial temperature is | 1.000000 |
| The cutoff temperature is | 0.010000 |
| The time index m is | 6 |

<table>
<thead>
<tr>
<th>Lambda</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.10</td>
<td>2.660</td>
<td>3.706</td>
<td>0.281</td>
<td>1.028</td>
</tr>
<tr>
<td>1.30</td>
<td>2.603</td>
<td>3.588</td>
<td>0.416</td>
<td>1.011</td>
</tr>
<tr>
<td>1.50</td>
<td>2.572</td>
<td>3.472</td>
<td>0.606</td>
<td>1.000</td>
</tr>
<tr>
<td>1.70</td>
<td>2.549</td>
<td>3.358</td>
<td>0.851</td>
<td>0.993</td>
</tr>
<tr>
<td>1.90</td>
<td>2.531</td>
<td>3.246</td>
<td>1.151</td>
<td>0.987</td>
</tr>
<tr>
<td>2.10</td>
<td>2.516</td>
<td>3.133</td>
<td>1.505</td>
<td>0.982</td>
</tr>
<tr>
<td>2.30</td>
<td>2.505</td>
<td>3.022</td>
<td>1.914</td>
<td>0.978</td>
</tr>
<tr>
<td>2.5C</td>
<td>2.496</td>
<td>2.910</td>
<td>2.378</td>
<td>0.975</td>
</tr>
<tr>
<td>2.70</td>
<td>2.489</td>
<td>2.799</td>
<td>2.897</td>
<td>0.973</td>
</tr>
<tr>
<td>2.90</td>
<td>2.483</td>
<td>2.687</td>
<td>3.471</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Table 8

Comparing the performance of SA on these two related problems, the data in Tables 7 and 8 indicate that the entropy measures of the associated clique problem are all less than the entropy of the 3-SAT problem. The highest entropy value in the clique problem of 2.66 occurred for $\lambda = 1.10$
compared to the entropy value of 9.55 for the 3-SAT problem. Notice that the expected value of the final state of the 3-SAT problem is also higher than the expected value of the final state of the associated clique problems—3.96 versus 3.706, respectively. This again indicates that higher entropy values are associated with better performance as measured by the expected value of the final state. This fact is also indicated in Table 8 itself which shows the relationships between the penalty parameter $\lambda$, the entropy measure, and the expected value of the final state. This data is depicted in Figure 7.

From Table 8 and illustrated in Figure 7, the entropy values decrease with increasing penalty parameter values while the expected value of the final state also tends to decrease with increasing penalty parameter values, thus indicating that poorer performance is associated with lower entropy values.
6.2.2. Transformation Experiment B

In this experiment, only 11 boolean variables were used in creating a 3-SAT problem instance. This decreased the size of the solution space for 3-SAT by a factor of 2. Thus, the solution space size is $2^{11} = 2048$. Tables 9 and 10 contain the computer output for this 3-SAT experiment and its related clique experiment, respectively.

**Computer Output For 3-SAT B**

| The number of clauses is | 4 |
| The number of variables is | 11 |
| The number of iterations is | 10 |
| The initial temperature is | 1.000000 |
| The cutoff temperature is | 0.010000 |
| The time index m is | 6 |

<table>
<thead>
<tr>
<th>N</th>
<th>H</th>
<th>EX</th>
<th>Var</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>9.17</td>
<td>3.96</td>
<td>0.04</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Table 9

**Computer Output For Clique B**

| The number of solutions is | 4096 |
| The number of nodes is | 12 |
| The number of iterations is | 10 |
| The initial temperature is | 1.000000 |
| The cutoff temperature is | 0.010000 |
| The time index m is | 6 |

<table>
<thead>
<tr>
<th>Lambda</th>
<th>H</th>
<th>EX</th>
<th>VAR</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.10</td>
<td>2.823</td>
<td>3.603</td>
<td>0.371</td>
<td>1.107</td>
</tr>
<tr>
<td>1.30</td>
<td>2.769</td>
<td>3.478</td>
<td>0.533</td>
<td>1.091</td>
</tr>
<tr>
<td>1.50</td>
<td>2.741</td>
<td>3.355</td>
<td>0.756</td>
<td>1.082</td>
</tr>
<tr>
<td>1.70</td>
<td>2.719</td>
<td>3.233</td>
<td>1.038</td>
<td>1.075</td>
</tr>
<tr>
<td>1.90</td>
<td>2.702</td>
<td>3.113</td>
<td>1.381</td>
<td>1.070</td>
</tr>
<tr>
<td>2.10</td>
<td>2.689</td>
<td>2.993</td>
<td>1.785</td>
<td>1.066</td>
</tr>
<tr>
<td>2.30</td>
<td>2.679</td>
<td>2.873</td>
<td>2.248</td>
<td>1.062</td>
</tr>
<tr>
<td>2.50</td>
<td>2.671</td>
<td>2.754</td>
<td>2.773</td>
<td>1.060</td>
</tr>
<tr>
<td>2.70</td>
<td>2.665</td>
<td>2.634</td>
<td>3.358</td>
<td>1.058</td>
</tr>
<tr>
<td>2.90</td>
<td>2.660</td>
<td>2.515</td>
<td>4.004</td>
<td>1.056</td>
</tr>
</tbody>
</table>

Table 10

Notice that the entropy measure in 3-SAT B, 9.17, is slightly less than in 3-SAT A due to the smaller solution space size in 3-SAT B—2048 versus 4096. Notice that the the norm is also
smaller indicating better convergence. In both experiments, however, the expected value of the final state is very close to the optimum value of 4. namely 3.96.

Table 10 contains the computer output for the associated clique problem. Notice that the number of nodes, hence the neighborhood size, is still 12 as it was in Clique A. This illustrates the effect of the transformation on the neighborhood size as discussed in Subsection 5.2.4.2. This data shows similar relationships as the data in Table 8. Thus, the performance of the clique problem is worse than in its related 3-SAT problem. In the 3-SAT problem, the expected value of the final state is 3.96, while in this related clique problem, the maximum expected value of the final state is 3.603. Once again, the entropy as well as the performance decreased in the transformed problem (the entropy in 3-SAT B is 9.17 while in its related clique problem, the entropy is 2.823).

In the clique problem itself, the expected value of the final state is also related to the entropy, both affected by the value of the penalty parameter. Rather than plot the expected value of the final state and entropy as a function of the penalty parameter, since a very similar graph as depicted in Figure 7 would result, only the entropy value is graphed in the figure below in order to better illustrate how the entropy measure decreases with increasing values of the penalty parameter.

![Entropy Versus Penalty Parameter](image)

Figure 8
6.3. CONCLUSION

The four methods described above all test the theory that the expected value of the final state of annealing experiments is closer to the globally optimal solution when the entropy measure is higher and vice versa. The generic experiments do indicate some errors which were described in Subsection 4.5.1, but the trend of increasing entropy values as a function of the neighborhood size was consistent with the decreasing expected value of the final state values as a function of the neighborhood size in these minimization problems.

It was also possible to make comparisons between the experiments on the two different generic configuration spaces. Both solution spaces A and B had the same number of solutions and the same globally minimum objective function value, but the ranges of their objective function values were different. This difference was designed to cause the entropy measures and expected value of the final state to be different. For given neighborhood sizes, the solution space A had consistently higher entropy values than the corresponding entropy measure in solution space B. Again, the higher entropy values were associated with superior performance.

The combinatorial experiments also provided two methods for testing the theory. The first by transformation of 3-SAT problem instances to clique problem instances. Both such problems have the same optimal solution, but different configuration spaces. This allowed for comparisons to be made of entropy measures and the expected value of the final state of these related configuration spaces. The experiments show that the transformations of two 3-SAT problems, one with 12 boolean variables, the other with 11 boolean variables, have a decreased entropy measure in their related clique problems. As the theory suggests, this decrease in entropy measure is associated with a decrease in the performance of SA. The maximum expected value of the final state in the clique problems were less than in their related 3-SAT problems.
In the clique problems themselves, different configuration spaces were created by changing the penalty parameter \( \lambda \). Again, the changes in entropy measures matched the changes in the expected value of the final state measures when plotted as a function of the penalty parameters.

All these data show a basic consistency with the theory that the expected value of the final state in an annealing experiment is related to the value of the entropy measure. In both minimization problems (the generic experiments) and maximization problems (the combinatorial experiments), higher entropy values are associated with improved performance of the SA algorithm.
CHAPTER 7

SIMULATED ANNEALING: PRESENT AND FUTURE

7.1. SUMMARY OF THIS RESEARCH

This section summarizes the basic elements of the research presented in this dissertation. Two subsections are described. The first focuses on the theoretical aspects of this research involving information theory and the relationship between entropy and the performance of the SA algorithm. The second subsection focuses on the experimental results which demonstrate the theoretical aspects of this research.

7.1.1. Modeling SA Using Information Theory

The contribution of this research stems from modeling the SA algorithm as a Markov information source. The notion that the succession of states generated by the SA algorithm produces information that can be measured is a new approach in the study of the SA algorithm. In viewing SA as a Markov information source, many of the concepts used in information theory can be applied to gain insights into the behavior of the SA algorithm. Specifically, the issue of measuring how much information is produced and its relationship to performance can be addressed using information theoretic concepts.

Among these information theoretic concepts is the AEP—a property associated with ergodic information sources. The AEP relates the sequences of states generated by an information source to an entropy measure. The literature discussing the AEP, however, deals either with homogeneous Markov chains, or more generally and abstractly, with ergodic information sources. It was therefore necessary to extend the concepts associated with the AEP to the type of information sources that the SA algorithm is modeled by, namely, strongly ergodic, inhomogeneous Markov chains.
7.1.1.1. Extensions of The AEP

In trying to apply the concepts associated with the AEP to SA, it was necessary to fashion a number of mathematical concepts to fit the context of SA. The AEP uses the concept of a set of typical sequences. In extending the AEP to the inhomogeneous case, it was necessary to define two sets of typical sequences—one based on the notion of convergence in probability and the other based on the frequency of pairs of states. Using these two sets, it was possible to show a relationship between the entropy measure of an annealing experiment and the number of sequences that have pairs of states close to their expected values. This was accomplished by using the intersection set of the two sets to relate the total probability of typical sequences to the entropy measure and also by showing that these two sets become the same set asymptotically (i.e., as the length of the sequences increases).

7.1.1.2. Using the Final State as an Estimator of The Optimal Solution

The reason for basing the definition of one of the typical sets on the frequency of pairs of states, was to relate the final state in an annealing experiment to the entropy measure associated with an annealing experiment. This was done because the final state serves as a useful measure of performance of the SA algorithm. Since the SA algorithm converges in probability, the frequency of generating globally optimal states should increase as the algorithm proceeds in its search of the configuration space. Thus, the final state can serve as an estimator of the globally optimal solution. The longer the algorithm runs, the closer the estimator is to the globally optimal value.

By relating the expected value of the final state to an entropy measure, a new theoretical foundation for the analysis of SA's performance has been achieved. This was done by showing that the expected value of the final state is related, asymptotically, to the entropy measure and that the higher this entropy measure is, the closer the expected value of the final state is to the globally optimal state. The asymptotic case involves a state of the system in which the mean objective function value (as determined using the stationary probability) is between the globally optimal solution and its next-to-least-cost neighbor.
7.1.1.3. **Development of Scaling Properties of SA**

Because the relationship of the expected value of the final state and entropy was limited to the asymptotic case, some method to extend this result to the more general case where the mean objective function value is arbitrary was needed. The approach used here was to develop a scaling property of SA. This scaling property is based on lumping the lowest cost states together into one supernode. This node has all of the mathematical properties of the globally optimal node upon which the asymptotic case is based. This made it possible to extend the analysis of the asymptotic case, in a very direct way, to the more general case.

In extending the asymptotic case to the general case, however, errors are introduced in the basic relationship between the mean objective function value of the final state and the entropy measure. This error is caused by differences between the vector of stationary probabilities in SA and the vector of actual state probabilities. Because the asymptotic analysis was derived using the stationary probabilities and not the actual state distribution, differences in these two vectors weaken the asymptotic relationship. The earlier the SA algorithm is stopped, the larger is this error since the difference in $L^1$ norm of these two vectors is larger.

By using the stationary probabilities, the mathematical analysis was simplified. Moreover, the error caused by doing this can be made arbitrarily small in the asymptotic case—when the mean objective function value is very close to the optimal objective function value. This is because the difference between the state distribution vector and the vector of stationary probabilities tends to zero in $L^1$ norm. The magnitude of this error was analyzed in Subsection 4.5.1 by a method involving a decomposition of the differences in these vectors and then using results from Mitra [1986].

7.1.1.4. **Theoretical Association of Entropy and the Finite-Time Performance of SA**

The relationship described above shows how the expected value of the final state in an annealing experiment is affected by the entropy of the associated Markov chain. This relationship shows that the expected value of the final state is bounded above (in a minimization problem) by an
expression in which the denominator has an exponential term involving the entropy measure while the numerator has an expression involving a polynomial in terms of the length of the annealing experiment. As such, it demonstrates that higher entropy values tend to lead to better finite-time performance of the SA algorithm.

7.1.2. Experimental Validation of the Theory

In demonstrating the theoretical relationship between the entropy measure and the finite-time performance of SA, four different experimental approaches were utilized. The first two methods were based on the concept of generic configuration spaces. The next two methods involved transformations and modifications of combinatorial optimization problems. These four different methods to test the theory all produced results consistent with the theory.

7.1.2.1. Generic Experiments

The generic experiments were based on using generic configuration spaces. These configuration spaces were not based on any particular combinatorial problem and are comprised of a list of objective function values and a set of rules used to determine the neighborhood structure. The rules used in these experiments were that neighbors of any particular solution were those objective function values in the solution list adjacent to that particular solution. Two solution spaces were created both of which had the same globally optimal objective function value.

Using generic configuration spaces permitted a great deal of flexibility to change the configuration space while preserving the same globally optimal objective function value. Changes to the configuration space were effected by changing the neighborhood size and by changing the range of objective function values. Calculations to determine the entropy measure and the expected value of the final state were then made. For each generic experiment, increasing the neighborhood size tended to increase the entropy measure, although these increases were not monotone. The trend in increasing entropy measures as a function of neighborhood size was mimicked by a decreasing trend in the expected value of the final state. This relationship also held in comparing the results of the two
solution spaces both having the same global minimum. The configuration space with the higher entropy values for a given neighborhood size showed superior performance when the SA algorithm was run on it than when the SA algorithm was run on the other configuration space.

7.1.2.2. Combinatorial Experiments

Combinatorial optimization problems were also used to demonstrate the theoretical relationship between entropy and the expected value of the final state. Two ways of demonstrating this relationship were employed. The first method was based on the use of a polynomial transformation algorithm from the theory of NP-completeness. This algorithm transformed two 3-SAT optimization problems to two clique optimization problems. Thus, for each problem instance in 3-SAT, another configuration space, with the same globally optimal solution, was created and the SA algorithm run on both problems using identical cooling schedules. For both sets, the transformation to clique degraded SA's performance in that the expected value in the clique problem was less than in the related 3-SAT problem (both are maximization problems). As expected, however, the entropy measure associated with the clique problem was also less than in the related 3-SAT problem. Thus, higher entropy values were associated with superior performance.

Within the clique problems themselves, different configuration spaces were created by using different penalty parameters. The higher penalty parameters were associated with lower performance and lower entropy values. As in the generic and the transformation experiments, higher entropy values were associated with superior performance.

7.1.3. Applications

The theories and experimental methodologies developed lay the foundation for the intuition that a smoother configuration space improves performance of the SA algorithm. In addition to providing insight into the behavior of SA, these ideas can also help to focus future research into answering the most significant issue regarding SA: how to ensure the effectiveness of SA. The theories show this can be done by increasing the entropy measure associated with a configuration
space. Consequently, the question of how best to achieve high entropy configuration spaces can now be addressed. The theories and experimental methodology used here not only provoke new questions, but also suggest practical methods that tend to increase the entropy values. For example, knowledge of how penalty parameters may affect the entropy value can be utilized to select appropriate or even optimal values for such parameters. With future research, it may even be possible to define a certain class of problems or transformation algorithms that tend to be associated with high entropy measures. This, in turn, can save a practitioner time by enabling him or her to decide \textit{a priori} whether SA is an appropriate method to use on certain problems. Moreover, the theories and experiments suggest how to modify a problem so that the effectiveness of SA is improved—by increasing the neighborhood size or decreasing the average value of $\Delta f$.

7.1.4. Conclusion

In analyzing the SA algorithm, the perspective of information theory suggested new tools and ideas to use in investigating the finite-time performance of this optimization technique. Some of these tools and ideas required modification in order to deal with the inhomogeneous Markov chain that embodies SA. The AEP, for example, is a well developed concept that had not been applied to strongly ergodic, inhomogeneous Markov chains of the type encountered in SA. By extending the AEP to the inhomogeneous case, a theoretical relationship between the number and total probability of typical annealing sequences and the entropy measure associated with those sequences was developed.

This theoretical relationship was then utilized to develop another theoretical relationship between the expected value of the final state in an annealing experiment and an entropy measure for the special case where the mean objective function value is arbitrarily close to the globally optimal objective function value. By developing scaling properties of SA, this relationship was then extended to the more general case of an arbitrary mean objective function value. Doing this introduced some errors in the relationship which were discussed and for which an upper bound is available.
SA has therefore been analyzed from a new perspective that facilitated a new theoretical understanding of its performance. As an information source, the greater the information produced (higher entropy) in the algorithm's search for the global optimum, the more likely it is that the final state of an experiment will be close to the globally optimal solution. By viewing a search algorithm as an information producing mechanism, a whole new approach in viewing optimization problems has been developed. A host of possibilities for future research therefore exists, and are discussed below.

7.2. DIRECTIONS FOR FUTURE RESEARCH

7.2.1. Estimating the Entropy

The main result of this dissertation shows that the performance of SA is related to an entropy measure. This entropy measure may therefore be useful in predicting when SA will perform well on a given problem, if some efficient method of determining the entropy measure can be devised. In the experiments in Chapter 6, the entropy measure was calculated exactly. This required all the data necessary to perform calculations on the Markov chain. To do this, the entire solution space had to be examined. This is equivalent in complexity to performing an exhaustive search of the solution space.

Clearly then, calculating the entropy measure exactly would be counterproductive in any practical setting. Consequently, methods of estimating the entropy are desirable. At present, it is unknown whether this is feasible. It may be that the entropy measure is too sensitive to certain transition probabilities in the transition matrix for this to be accomplished with any accuracy. That is, extrapolating the entropy measure for the entire Markov chain based on the entropy associated with a sample of solutions may not be reasonable. Some research is therefore necessary to determine the robustness of an estimate of the entropy measure when given a relatively small sample of solutions from the solution space.
7.2.2. The AEP

In extending the AEP to the inhomogeneous case, one of the lemmas used requires that the time index \( m \) be sufficiently large (see Lemma 3.2). This is based on a function (see the proof of Lemma 3.2) that diverges to \(-\infty\) as \( m \to \infty \). This function is depicted below.

**Function Used in the Inhomogeneous AEP**

![Function diagram](image)

Figure 1

Since this function diverges to \(-\infty\) for small and large temperatures, the requirement that the time index \( m \) be sufficiently large (temperature sufficiently small) may be unnecessary. It may be possible to determine a maximum value of this function and utilize it in generalizing the lemma statement.

7.2.3. Counting the Combinations of State Pairs

In Subsection 4.4, an effort was made to determine the number of different combinations of state pairs that met the criterion for inclusion in the typical group of sequences. One of the reasons for developing the asymptotic analysis was to simplify this counting problem. This was done by reducing the number of possible state pairs that needed to be counted so as to involve only the global optimum and its next-to-least-cost neighbor. It may be possible that an efficient method for counting all the relevant combinations of state pairs could be developed and, in some way, obviate the necessity of the simplifications used in the asymptotic analysis.
7.2.4. **Bounding the Error in Using the Scaling Property**

In Subsection 4.5.1, an error expression for the difference between the state distribution and the vector of stationary probabilities was decomposed and analyzed. This error term was ignored in Subsections 4.4 and 4.5 in order to simplify the analysis. These error terms are caused by using the vector of stationary probabilities instead of the more accurate state distribution vector. Since Mitra et al. [1986] has upper bounds for these error terms, it is possible to incorporate these terms directly in the analysis and develop a more precise relationship between the expected value of the final state and the entropy measure.

7.2.5. **Relating the Eigenvalues and Entropy**

The relationship connecting the mean objective function value to the entropy measure illustrates a type of convergence—the mean objective function value is closer to the global optimum when the entropy measure is higher. One interesting possibility, therefore, would be to relate the convergence rate of a Markov chain to an entropy measure. This has not been done even for the homogeneous case.

By using the approach in Subsections 4.4 and 4.5 for a homogeneous Markov chain, some expression relating convergence rates to entropy may be possible. The difficulty in doing this for the inhomogeneous case is that the expression for the bounds of the errors, described in the preceding section, is quite cumbersome, yet relates the expected value of the final state to an entropy measure. By limiting the same analysis to homogeneous Markov chains, this error bound can be written in terms of the second largest eigenvalue of a stationary transition matrix. This may therefore provide a way of expressing the convergence rate in terms of an entropy measure and the second largest eigenvalue for a transition matrix in a Markov chain of the type used in analyzing SA where the temperature parameter is held constant. Such an expression would be new and demonstrate, at least in the homogeneous case, how entropy affects the convergence of a Markov chain to a steady-state.
7.2.6. **Noisy Information Sources**

This dissertation evolved entirely from viewing the SA algorithm as an information source. One possible embellishment of this perspective would be to model the SA algorithm as a *noisy* information source. Recall that at the beginning of an experiment, when the temperature parameter is high, the states in the configuration space are visited in a highly random manner. As the temperature is lowered and the algorithm converges in probability, the states become more predictable. One could therefore make an analogy that the information source is always transmitting the globally optimal state, but that these states are perturbed or altered by noise. This noise level is higher when the temperature parameter is higher. A great deal of literature exists on noisy information sources (see for example Shannon [1948]) and using some of the approaches in this dissertation, modeling the SA algorithm as a noisy information source may provide even further insights into SA's finite-time behavior.

7.2.7. **Parallel Implementations**

Some of the combinatorial optimization problems presented in this dissertation provide some interesting implications for running SA in parallel. For example, if two SA algorithms are run for two related graph theoretic problems using common random number streams where the solutions to both add to some constant $|\Gamma|$ at some time $k^*$, then the solutions of each program for all iterations after time index $k^*$ ($\forall k > k^*$) will always add to $|\Gamma|$. For a simple proof, see Appendix B. This may be useful in developing an adaptive scheme for controlling the temperature parameter.

Consider the following example: two SA algorithms operate in parallel on the same graph—one for solving the independent set problem, the other for solving the vertex cover problem. Each computer uses the same set of nodes for the initial solution. Thus, if $\alpha$ is the initial set of nodes for independent set and $\beta$ the initial set for vertex cover then $\alpha = \beta$. As the SA algorithm proceeds, each set of nodes will tend toward the optimal set for each program. Thus, the nodes will tend to *separate* as they migrate toward the optimal sets and eventually become mutually exclusive based on
(5.15). One adaptive scheme would be to set the temperature parameter equal to some function of the intersection set of the nodes. For example,

\[ t_k = \frac{|\alpha \cap \beta|}{\log(c + k)}. \]

Once \( \alpha \cap \beta = \emptyset \), then the SA would freeze (i.e., temperature becomes zero), remain frozen (see the proof in Appendix B), and proceed based on local search. Since at low temperatures there is a greater probability that solutions will be in the neighborhood of the global solution, local search at that point should yield the global optimum. A reasonable conjecture is that the variance of the final solution based on the above procedure would be significantly reduced.

For such a system, application of the theory presented in this dissertation would have to be modified. For instance, in the graph problems described above, when SA is applied, the algorithm can be modeled as a Markov chain as before, but with random transition probabilities. Consequently, in developing a relationship between the performance of this adaptive scheme and entropy, the appropriate entropy measure would be the expected value of the entropy of the Markov chain.

7.2.8. Other Transformation Experiments

The theory of \( NP \)-completeness shows that there are a host of \( NP \)-hard optimization problems that can be reduced (transformed) to other \( NP \)-hard optimization problems (see Garey and Johnson [1979]). A number of problems other than those selected in Chapter 6 are amenable to SA and can therefore be selected and experiments performed on them similar to those in Chapter 6. Different problems may, however, have significantly different entropy measures and performance characteristics when SA is applied to them. Although \( NP \)-hard problems that are transformed from other such problems are known to be "at least as hard" to solve as the original problem (see Garey and Johnson [1979 pp.109]), application of the theories presented here may shed new light on how difficult transformed problems are and it may now be possible to actually quantify this difficulty in a new way.
APPENDIX A

Proof of Theorem 3.1:

Define, in similar fashion as in Goldie and Pinch [1991 pp.53], the random variables $\xi_m = -\ln \Pr \{ Z_m \}/n$, $\xi_m = -\ln r_{m-1}^{[m-1]}$ for $m \geq 1$, and $\xi_0 = -\ln v_0^{[0]}$ where the $z_m$ are the randomly generated states of the random sequence $Z_n$. Therefore,

$$E\{\xi_m\} = -\sum_{i=1}^{s} \sum_{j=1}^{s} \Pr \{ Z_m = j, Z_{m-1} = i \} \ln p_{ij}^{[m-1]}.$$  

Note that $\Pr \{ Z_m = j, Z_{m-1} = i \} = \sum_{k \in S} v_k^{[0]} p_{kj}^{[0,m-2]} p_i^{[m-1]}$. But $\sum_{k \in S} v_k^{[0]} p_{kj}^{[0,m-2]} = v_i^{[m-1]}$. (In SA, this term can be written as $q_i + O(1/k^{\min(a,b)})$ where $q_i$ is the vector element corresponding to the optimal state distribution vector and $k$, $a$, and $b$ are as in Mitra et al. [1986] for any arbitrary starting vector.) Therefore,

$$E\{\xi_m\} = -\sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m-1]} p_{ij}^{[m-1]} \ln p_{ij}^{[m-1]} = H(m-1). \tag{1}$$

Define $\lim_{m \to \infty} p_{ij}^{[m]} \equiv p_{ij}^{[\infty]}$. In SA, the transition probabilities converge to

$$p_{ij}^{[\infty]} = \begin{cases} 
0 & \text{if } \Delta f_{ji}^* > 0 \\
\frac{1}{|N(i)|} & \text{if } \Delta f_{ji}^* = 0, j \in N(i) \\
1 - \sum_{l \neq i} p_{il}^{[\infty]} & \text{if } j = i \\
0 & \text{otherwise}
\end{cases} \tag{2}$$

and the state distribution converges to $v_i^{[\infty]} = q_i = \begin{cases} 
\frac{1}{s^*} & \text{if } i \in S^* \\
0 & \text{otherwise}
\end{cases}$. 

120
Therefore,

$$
\lim_{n \to \infty} E\{\xi_n\} = \lim_{n \to \infty} \sum_{j=1}^{s} \ln \sum_{j=1}^{s} p_j^{[m-1]} \ln p_j^{[m-1]}
$$

$$
\leq - \sum_{j=1}^{s} v_j^{[m-1]} \ln v_j^{[m-1]} = - \sum_{j=1}^{s} q_j \ln q_j
$$

$$
= \ln |s^{[m]}|  \tag{3}
$$

where the $\ln p_j^{[m-1]}$ is replaced with $\ln v_j^{[m-1]}$ to obtain the inequality from Gibb's Inequality (see the corollary to Theorem 2.1). From the definitions of $\xi_n$, $\xi_0$, and $\xi_m$, and recalling that $Pr\{Z_n\} = v_{\xi_0}^{[0]} p_{\xi_1}^{[1]} p_{\xi_2}^{[1]} \ldots p_{\xi_{n-1}}^{[n-2]}$, then $\bar{\xi}_n = (\xi_0 + \xi_1 + \ldots + \xi_{n-1})/n$. Therefore,

$$
E\{\bar{\xi}_n\} = \frac{E\{\xi_0\} + E\{\xi_1\} + \ldots + E\{\xi_{n-1}\}}{n}
$$

$$
= \frac{E\{\xi_0\} + \sum_{k=0}^{n-2} H(k)}{n}
$$

$$
= \frac{E\{\xi_0\} + H^*(0, n - 2)}{n}
$$

Since the limit of a sequence equals the limit of its Cesàro averages (Goldie and Pinch [1991 pp. 54]), then

$$
\lim_{n \to \infty} E\{\bar{\xi}_n\} = \lim_{n \to \infty} \frac{H^*(0, n - 2)}{n} \leq \ln |s^{[m]}|.
$$

Q.E.D.

Proof of Corollary 1 to Theorem 3.1:

Using the same definitions as above and defining $\tilde{\xi}_{(m,n-1)} = - \ln Pr\{Z_{(m,n-1)}\}/(n-m)$ and $\tilde{\xi}_n' = - \ln v^{[m]}_n$, then from the definition of $Z_{(m,n-1)},$

$$
\tilde{\xi}_{(m,n-1)} = \frac{\tilde{\xi}_n + \tilde{\xi}_{n+1} + \tilde{\xi}_{n+2} + \ldots + \tilde{\xi}_{n-1}}{n-m}
$$
Therefore,
\[ E \left\{ \tilde{\xi}_{m,n-1} \right\} = \frac{E \{ \xi_m \} + E \{ \xi_{m+1} \} + E \{ \xi_{m+2} \} + \ldots + E \{ \xi_{n-1} \}}{n-m} \]
\[ = \frac{E \{ \xi_m \} + H(m) + H(m+1) + \ldots + H(n-2)}{n-m} \]
\[ = \frac{E \{ \xi_m \} + H^*(m,n-2)}{n-m}. \]

Consequently,
\[ \lim_{n \to \infty} E \left\{ \tilde{\xi}_{m,n-1} \right\} = \lim_{n \to \infty} \frac{H^*(m,n-2)}{n-m} \leq \ln|s^*| \]

for the same reasons given in the case where \( m = 0. \)

Q.E.D.

Proof of Corollary 2 to Theorem 3.1:

In similar fashion as in Goldie and Pinch [1991 pp.41], we show that
\[ \tilde{\xi}(m,n-1) \xrightarrow{t} H^*(m,n-2) \]
and that therefore,
\[ \tilde{\xi}(m,n-1) \xrightarrow{P} H^*(m,n-2) \] (by Chebychev's Inequality; see Goldie and Pinch [1991 pp.53]). To do this, it must be shown that \( \text{var} \tilde{\xi}(m,n-1) \to 0 \) as \( n \to \infty. \) Using the definition of \( \tilde{\xi}(m,n-1) \)
\[ \text{var} \left[ \tilde{\xi}(m,n-1) \right] = \frac{\sum_{k=m}^{n-1} \sum_{j=m}^{n-1} \text{cov}(\xi_k, \xi_j)}{(n-m)^2} \] (4)

with the \( \xi'_m \) term denoted by \( \tilde{\xi}. \) To determine bounds for these covariances, the following result is needed from Mitra \textit{et al.} [1986] and the theory of strongly ergodic Markov chains (see Isaacson and Madsen [1976 pp. 157]). For fixed \( k \) and for all \( i \) and \( j, \)
\[ |p_j^{[k_{a^+, b^+}]} - v^{[a^+, b^+]}_j| = O \left( \frac{1}{r_{\min(a,b)}} \right). \] (5)

where the \( a \) and \( b \) are defined in Mitra \textit{et al.} [1986]. But in SA \( p_j^{[k_{a^+, b^+}]} = p_j^{[0, k_{a^+, b^+}]} \). Therefore, as in (5),
\[ |p^{(i)}_{x} - v^{(i)}_{x}| = \frac{1}{k_{max(x,y)}} \]

The terms of (4) can be decomposed into five separate parts based on the indices \( k \) and \( r \) in the \( \text{cov}(z_{k}, z_{k+r}) \) terms. Thus,

\[
\text{var} \left( \bar{z}(m, n-1) \right) = \frac{\text{cov}(\bar{z}_{m}, \bar{z}_{m})}{(n-m)^{2}} + \frac{\sum_{k=m+1}^{n-1} \text{cov}(\bar{z}_{k}, \bar{z}_{k})}{(n-m)^{2}} + \frac{2 \sum_{r=1}^{n-m-1} \text{cov}(\bar{z}_{m}, \bar{z}_{m+r})}{(n-m)^{2}}

+ \frac{2 \sum_{k=m+1}^{n-2} \text{cov}(\bar{z}_{k}, \bar{z}_{k+1})}{(n-m)^{2}} + \frac{2 \sum_{k=m+1}^{n-2} \sum_{r=1}^{n-k-1} \text{cov}(\bar{z}_{k}, \bar{z}_{k+r})}{(n-m)^{2}}.
\]

(7)

Part 1 is fixed in terms of \( m \), i.e., \( k = m \) and \( r = 0 \). Thus,

\[ \lim_{n \to \infty} \text{cov}(\bar{z}_{m}, \bar{z}_{m})/(n-m)^{2} = 0. \]

Part 2 consists of terms with \( k > m \) and \( r = 0 \). This corresponds to the diagonal elements of the associated variance/covariance matrix. Thus, for \( k > m \) and \( r = 0 \),

\[ E(\bar{z}_{k} \bar{z}_{k}) = \sum_{i=1}^{n} \sum_{i=1}^{n} \bar{v}_{i}^{(k)} \bar{v}_{i}^{(k)} \left[ -\ln p^{(k)}_{y} \right]^{2} \]

\[ \leq C \]

for \( C = \max \sum_{i=1}^{n} \bar{v}_{i}^{(k)} \left[ -\ln p^{(k)}_{y} \right]^{2} \). Therefore, for \( k > m \) and \( r = 0 \),

\[ \text{cov}(\bar{z}_{k}, \bar{z}_{k}) \leq C = \rho(n) \]

Thus, for Part 2,

\[ \frac{\sum_{k=m+1}^{n-1} \text{cov}(\bar{z}_{k}, \bar{z}_{k})}{(n-m)^{2}} = \frac{(n-m-1)\rho(n)}{(n-m)^{2}} \to 0 \text{ as } n \to \infty. \]

Part 3 involves terms where \( k = m \) and \( r > 0 \), i.e., that have terms with \( \bar{z}_{m}^{(r)} \). For each such term note that the following inequality holds:

\[ \text{cov}(\bar{z}_{m}^{(r)}, \bar{z}_{m+r}) = E(\bar{z}_{m}^{(r)} \bar{z}_{m+r}) - E(\bar{z}_{m}^{(r)})E(\bar{z}_{m+r}) \leq E(\bar{z}_{m}^{(r)} \bar{z}_{m+r}). \]
To evaluate bounds for the $E_i(z_{m+1}^r, z_{m+1})$, this expectation can be written based on the following expression:

$$\Pr\{Z_m = i, Z_{m+1} = j^*, Z_{m+1} = i\} = \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} \Pr\{Z_m = i, Z_{m+1} = j^*, Z_{m+1} = i\} \ln v_i^{(m)} \ln p_{i,j}^{(m^2+1)}$$

Thereafter,

$$E_i(z_{m+1}^r, z_{m+1}) = \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} \Pr\{Z_m = i, Z_{m+1} = j^*, Z_{m+1} = i\} \ln v_i^{(m)} \ln p_{i,j}^{(m^2+1)}$$

Substituting in for the probability expression,

$$E_i(z_{m+1}^r, z_{m+1}) = \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} \sum_{r=1}^{r_{\lambda}} v_i^{(m)} p_{i,j}^{(m^2+1)} p_{i,j}^{(m^2+1)} [\ln v_i^{(m)} - \ln p_{i,j}^{(m^2+1)}]$$

$$= \sum_{i=1}^{\lambda} v_i^{(m)} [\ln v_i^{(m)}] \sum_{j=1}^{\lambda} \sum_{r=1}^{r_{\lambda}} p_{i,j}^{(m^2+1)} [\ln p_{i,j}^{(m^2+1)}]$$

$$= \sum_{i=1}^{\lambda} v_i^{(m)} [\ln v_i^{(m)}] \sum_{r=1}^{r_{\lambda}} \sum_{j=1}^{\lambda} [v_i^{(m^2+1)}] + O(1/r^m) \sum_{j=1}^{\lambda} p_{i,j}^{(m^2+1)} [\ln p_{i,j}^{(m^2+1)}]$$

$$= \tilde{H}(m) [H(m + r - 1) + O(1/r^m)]$$

where $\tilde{H}(m) = \sum_{i=1}^{\lambda} v_i^{(m)} [\ln v_i^{(m)}]$. But,

$$E_i(z_{m+1}^r) = -\sum_{i=1}^{\lambda} \Pr\{Z_m = i\} \ln v_i^{(m)}$$

$$= \sum_{i=1}^{\lambda} v_i^{(m)} \ln v_i^{(m)}$$

$$= \tilde{H}(m).$$

Consequently,

$$\text{cov}(z_{m+1}^r, z_{m+1}) = E(z_{m+1}^r, z_{m+1}) - E(z_{m+1}^r)E(z_{m+1})$$

$$= \tilde{H}(m) [H(m + r - 1) + O(1/r^m)] - \tilde{H}(m)H(m + r - 1)$$

$$= O(1/r^m).$$

Thus, $\lim_{r\to\infty} \text{cov}(z_{m+1}^r, z_{m+1}) = 0$. Therefore these terms are $o(1)$. Consequently, for Part 3...
\[
\frac{2}{(n-m)^2} \sum_{r=1}^{n-m-1} \text{cov}\{\xi'_r, \xi'_{m+r}\} = \frac{2(n-m-1)\sigma(1)}{(n-m)^2} \to 0 \text{ as } n \to \infty.
\]

Part 4 consists of terms with \( k > m \) and \( r = 1 \). To evaluate \( E\{\xi_k \xi_{k+1}\} \) note that in this case, the weighting probability is

\[
\Pr\{Z_{k+1} = i', \tau_k = j, Z_{k-1} = i\} = \sum\limits_{i \in S} v_i^{(1)} p_i^{(0, \lambda-1)} p_j^{(1)} p_{i'}^{(1)},
\]

Therefore,

\[
E\{\xi_k \xi_{k+1}\} = \sum\limits_{i \in S} \sum\limits_{j \in S} \sum\limits_{i' \in S} v_i^{(1)} p_i^{(1)} \left[ -\ln p_j^{(1)} \right] \left[ -\ln p_{i'}^{(1)} \right] = H(k-1) \sum\limits_{i \in S} p_i^{(1)} \left[ -\ln p_i^{(1)} \right] \leq H(k-1) \ln s \leq (\ln s)^2
\]

since \( |\ln s| \leq \ln s \). Therefore, for all covariance terms where \( k > m \) and \( r = 1 \),

\[
\text{cov}\{\xi_k, \xi_{k+1}\} \leq (\ln s)^2 = o(n).
\]

Thus, for Part 4

\[
\frac{2}{(n-m)^2} \sum_{k=m+1}^{n-2} \text{cov}\{\xi_k, \xi_{k+1}\} = \frac{2(n-m-2)\sigma(n)}{(n-m)^2} \to 0 \text{ as } n \to \infty.
\]

Finally, for Part 5, to determine bounds for \( E\{\xi_k \xi_{k+r}\} \) note that

\[
\Pr\{Z_k = j, Z_{k+r} = i, Z_{k+r-1} = i'\} = \sum\limits_{i \in S} v_i^{(0)} p_i^{(0, \lambda-1)} p_j^{(1)} p_{i'}^{(1)} p_{i'}^{(1)},
\]

Therefore,

\[
E\{\xi_k \xi_{k+r}\} = \sum\limits_{i \in S} \sum\limits_{j \in S} \sum\limits_{i' \in S} \Pr\{Z_k = j, Z_{k-1} = i, Z_{k+r} = j', Z_{k+r-1} = i'\} \ln p_j^{(1)} \ln p_{i'}^{(1)}.
\]

Substituting (8) into (9).
\[ E(\xi_k z_{k+r}) = \sum_{i=1}^{x} \sum_{j=1}^{x} \sum_{l=1}^{x} v_i^{[0]} p_u^{[0, a-1]} p_g^{[k-1]} p_i^{[k+a-2]} p_{ij}^{[k+r-2]} \ln p_y^{[k-1]} \ln p_{ij}^{[k+r-1]} \cdot \]

Noting that \( \sum_{i \in S} v_i^{[0]} p_u^{[0, a-1]} = v_i^{[k-1]} \) and from (6) that \( p_{ij}^{[k+a-2]} = v_i^{[k+r-1]} + O(1/k^{\min(a,b)}) \) and rearranging,

\[ E(\xi_k z_{k+r}) = \sum_{i=1}^{x} \sum_{j=1}^{x} \sum_{l=1}^{x} v_i^{[k-1]} p_g^{[k-1]} p_i^{[k+a-2]} p_{ij}^{[k+r-1]} \left[-\ln p_y^{[k-1]}\right] \left[-\ln p_{ij}^{[k+r-1]}\right] \]

\[ = \sum_{i=1}^{x} \sum_{j=1}^{x} v_i^{[k-1]} \left[-\ln p_y^{[k-1]}\right] \left[\sum_{l=1}^{x} p_i^{[k+a-2]} p_{ij}^{[k+r-1]} \left[-\ln p_{ij}^{[k+r-1]}\right]\right] \]

\[ = H(k-1) \left[\sum_{i=1}^{x} \sum_{j=1}^{x} v_i^{[k+r-1]} + O(1/k^{\min(a,b)}) \right] p_{ij}^{[k+r-1]} \left[-\ln p_{ij}^{[k+r-1]}\right] \]

\[ = H(k-1) \left[H(k+r-1) + O(1/k^{\min(a,b)})\right]. \]

Consequently,

\[ \text{cov}(\xi_k, z_{k+r}) = E(\xi_k z_{k+r}) - E(\xi_k) E(z_{k+r}) \]

\[ = H(k-1) \left[H(k+r-1) + O(1/k^{\min(a,b)})\right] - H(k-1)H(k+r-1) \]

\[ = O(1/k^{\min(a,b)}). \]

Thus, for all \( k > m \) and \( r > 1 \), \( \lim_{k \to \infty} \text{cov}(\xi_k, z_{k+r}) = 0 \) and therefore these terms are \( o(1) \).

Consequently, for Part 5,

\[ 2 \sum_{k=m+1}^{n-3} \sum_{r=2}^{n-k-1} \text{cov}(\xi_k, z_{k+r}) = \frac{(n-m-3)(n-m-4) o(1)}{(n-m)^2} \to 0 \text{ as } n \to \infty. \]

From Chebychev's Inequality, for any \( \delta > 0 \),

\[ \Pr \left( \left| \frac{-\ln \Pr(Z_{(m,n-1)}) - H^*(m,n-2)}{n-m} \right| \leq \delta \right) \geq 1 - \frac{\text{var} \xi(m,n-1)}{\delta^2}. \]

Since all the parts of (7) converge to zero, i.e., \( \text{var} \xi(m,n-1) \to 0 \) as \( n \to \infty \), then for any \( \varepsilon > 0 \) there exists an \( n \) such that \( \text{var} \xi(m,n-1)/\delta^2 \leq \varepsilon \) and the theorem is proved.

\[ Q.E.D. \]
Proof of Lemma 3.1:

Let \( k_u \) denote the temperature index associated with the \( u^{th} \) transition in the subsequence of transitions. Then, for the non-diagonal elements of the transition matrix, the \( p_{ij}(t_{k_u}) \) equal the transition probabilities of the actual \( ij \) transitions occurring in the sequence between time indices \( m \) and \( n \). With the cooling schedule in (1.3), the non-diagonal elements of the transition matrix constitute a monotonically decreasing sequence of reals in \((0,1)\). Therefore, \( p_{ij}(t_m) \geq p_{ij}(t_{k_1}) > p_{ij}(t_{k_2}) > \cdots > p_{ij}(t_{k_r}) \geq p_{ij}(t_n) \).

Consequently, \( p_{ij}(t_{k_1})p_{ij}(t_{k_2}) \cdots p_{ij}(t_{k_r}) \leq \left(p_{ij}(t_m)\right)^r \), where \( r = m_{ij}(m,n) \).

Therefore, \( \left[p_{ij}(t_{k_1})p_{ij}(t_{k_2}) \cdots p_{ij}(t_{k_r})\right]^{\frac{1}{r}} \leq \left[\left(p_{ij}(t_m)\right)^r\right]^{\frac{1}{r}} = p_{ij}(t_m) \).

But, \( \left[p_{ij}(t_{k_1})p_{ij}(t_{k_2}) \cdots p_{ij}(t_{k_r})\right]^{\frac{1}{r}} = \hat{p}_{ij}(m,n) \).

Therefore, \( \hat{p}_{ij}(m,n) = C_{ij} \leq p_{ij}(t_m) \).

Using similar arguments for the lower bound, gives

\[
p_{ij}(t_n) \leq \hat{p}_{ij}(m,n) \leq p_{ij}(t_m).
\]

Since \( p_{ij}(t) \) is a continuous function in \( t \), defined on a compact interval \( [t_n, t_m] \), by the Intermediate Value Theorem (Salas et al. [1974 pp.71]), there exists some constant \( C_{ij} \) such that \( p_{ij}(t_n) \leq C_{ij} \leq p_{ij}(t_m) \), then there exists a \( t_{ij} \in [t_n, t_m] \) such that \( p_{ij}(t_{ij}) = C_{ij} = \hat{p}_{ij}(m,n) \).

Using similar reasoning for the diagonal elements which are a monotonically increasing sequence of reals, the same conclusion can be reached (in SA, for the nondiagonal elements the value of \( t_{ij} \) can be explicitly stated).

\(Q.E.D.\)
Proof of Lemma 3.2:

Since the outer summation is over index \( i \), it is sufficient to show that for all \( i \), there exists a \( t_i > 0 \) such that

\[
\sum_{j=1}^{i} m_{ij}(m,n) \ln p_{ij}(t_{ij}) = C_i(m,n) = \sum_{j=1}^{i} m_{ij}(m,n) \ln p_{ij}(t_i)
\]  

(10)

where the existence of \( t_{ij} \) is established in Lemma 3.1.

There are three cases to consider; the first two are trivial. The first is where all the transitions in the subsequence are to the same state, \( i.e. \) the sequence has \( m-n+1 \) \( ii \) transitions and zero transitions to other states. In this case, there is only one temperature value \( t_{ii} \) obtained from the application of Lemma 3.1. Therefore, \( t_i = t_{ii} \) and (10) is satisfied.

In the second case, for some state \( i \), there are transitions only from state \( i \) to its neighbors and zero \( ii \) transitions. From (1.2), this occurs if \( \Delta f_{ij}^+ = 0 \) such as when state \( i \) is a local maximum. In this case, the transition probabilities are constant and independent of the temperature parameter. Thus, any nonzero value for the temperature parameter satisfies (10).

In the third case, there are transitions to state \( i \) and transitions to other states as well. \( C_i(m,n) \) assumes larger values when the transitions to other states occur at the beginning of the subsequence and transitions to the same state occur at the end of the subsequence. This is because the \( ij \) transitions have larger probabilities at higher temperature values (early in the subsequence) and the \( ii \) transitions have larger probabilities at low temperature values (later in the subsequence, see (1.2) and (1.3)). Thus, for a subsequence of \( r \) transitions from time index \( m \) to time index \( n \), \( C_i(m,n) \) is largest for subsequences of the form

\[
P_{ij} \cdot P_{ij} \cdot P_{ij} \cdot P_{ij} \cdot \ldots \cdot P_{ij} \cdot \ldots \cdot P_{ij}
\]

where

- \( P_{ij} \) transitions from \( i \) to other states
- \( P_{ii} \) transitions from \( i \) to \( i \)
Therefore, for all subsequences with $m_{ij} > 0$ $ij$ transitions and $m_{ii}$ $ii$ transitions and for any state $j$ and for all time indices between time index $m$ and time index $n$, define $\hat{C}_i(m, n, j)$ from

$$\max_{m_{ij} \geq 0} C_i(m, n) \leq \min_j \left[ m_{ij} \ln p_{ij}(t_m) + m_{ii} \ln p_{ii}(t_n) \right]$$

$$< \min_j \left[ m_{ij} \ln p_{ij}(t_m) + m_{ii} \ln p_{ii}(t_n) \right]$$

$$= \hat{C}_i(m, n, j)$$

Note also that

$$\hat{C}_i(m, n, j) = m_{ij}(m, n) \ln p_{ij}(t_m) + m_{ii}(m, n) \ln p_{ii}(t_n)$$

$$< m_{ij}(m, n) \ln p_{ij}(t_m).$$

Consequently, for all $M < 0$ and for time indices $m > m^*(n)$ sufficiently large,

$$\max_{m_{ij}} C_i(m, n) < \hat{C}_i(m, n, j) < m_{ij}(m, n) \ln p_{ij}(t_m) < M$$

since $p_{ij}(t) \to 0$ and $\ln p_{ij}(t_m) \to -\infty$ monotonically as $t \to 0^+$. Thus, it is always possible to find an $m^*(n)$ such that for all $m > m^*(n)$, $m_{ij}(m, n) \ln p_{ij}(t_m) \leq M$ provided that $\Delta f_{ji}^+ > 0$. Now for each subsequence of transitions and for each state $i$, define the function

$$F_i(t) = \sum_{j \neq i} m_{ij}(m, n) \ln p_{ij}(t)$$

$$= \sum_{j \neq i} m_{ij}(m, n) \ln p_{ij}(t) + m_{ii}(m, n) \ln p_{ii}(t)$$

where $p_{ii}(t) = 1 - \sum_{j \neq i} p_{ij}(t)$ and corresponds to the diagonal elements of the transition matrix (see (1.2)). Choose $M$ such that $M = \max_{0 < t < \infty} F_i(t)$. Therefore, for $m$ sufficiently large,

$$C_i(m, n) \leq \max_{0 < t < \infty} F_i(t).$$
The next step is to show that the function $F_i(t)$ is increasing in the neighborhood of zero in order to show that a lower bound exists. Observe that

$$\lim_{t \to 0^+} F_i(t) = \lim_{t \to 0^+} \sum_{j \neq i} m_{ij}(m,n) \ln p_j(t) + \sum_{j \neq i} m_{ij}(m,n) \ln p_i(t)$$

diverges to $-\infty$, provided $m_{ij} > 0$ for some $j \neq i$. Note also that

$$\frac{\partial}{\partial t} F_i(t) = F'_i(t) = \sum_{j \neq i} \frac{\partial}{\partial t} \left( m_{ij}(m,n) \ln p_j(t) \right) + \frac{\partial}{\partial t} \left( m_{ij}(m,n) \ln p_i(t) \right).$$

Since $p_{ii}(t) = 1 - \sum_{j \neq i} p_{ij}(t)$, then

$$\frac{\partial}{\partial t} F_i(t) = \sum_{j \neq i} \frac{m_{ij}}{p_{ij}(t)} \frac{\partial p_{ij}(t)}{\partial t} - \frac{m_{ij}}{p_{ii}(t)} \left[ \sum_{j \neq i} \frac{\partial p_{ij}(t)}{\partial t} \right]$$

$$= \sum_{j \neq i} \frac{\partial p_{ij}(t)}{\partial t} \left[ \frac{m_{ij}}{p_{ij}(t)} - \frac{m_{ij}}{p_{ii}(t)} \right].$$

By the cooling schedule in (1.3), $\frac{\partial}{\partial t} p_{ij}(t) = p_{ij}(t) \frac{\Delta f^i_j}{t^2} > 0$ for all $t > 0$, $\Delta f^i_j > 0$ and $j \neq i$.

Consequently, as $t \to 0^+$, $p_{ij}(t) \to 0$, $p_{ii}(t) \to 1$ and $\frac{m_{ij}}{p_{ij}(t)}$ diverges to $\infty$. Therefore $F_i(t) \to -\infty$ as $t \to 0$ (provided there is at least one non-diagonal transition, i.e., $m_{ij} > 0$ for some $ij$). Thus, for small $t > 0$, $F_i(t)$ is an increasing function. Since $p_{ij}(t) \leq 1$, then $\ln p_{ij}(t) \leq 0$, and consequently $F_i(t) < 0$ for all $t$. But since $F_i(t)$ is continuous for all $t > 0$, then by the Extreme Value Theorem (Salas et al. [1974 pp.72]) there exist values $t''$ and $M$ such that $\max_{t''} F_i(t) = F_i(t'') = M$. Recall from (11), that for $m$ sufficiently large, $C_i(m,n) \leq M$. Since $\lim_{t \to 0^+} F_i(t)$ diverges to $-\infty$, there exists some value $t'$ such that $F_i(t') \leq C_i(m,n)$. Therefore, for $m$ sufficiently large, there exists values
\( t' \) and \( t'' \) such that \( F_i(t') \leq C_i(m^*(i), n) \leq F_i(t'') \). By the Intermediate Value Theorem, there exists some \( t_i \in [t', t''] \) such that \( F_i(t_i) = C_i(m, n) \) for each state \( i \). Therefore, for all \( m > m^*(i) \), there exists some \( t_i \) such that

\[
\sum_{j=1}^{r} m_{ij}(m, n) \ln p_{y_j}(t_{ij}) = \sum_{j=1}^{r} m_{ij}(m, n) \ln p_{y_j}(t_i) \]

The final step then is to choose the maximum time index \( m^* = \max[m^*(i)] \). Thus, for all \( m > m^* \),

\[
\sum_{j=1}^{r} \sum_{i=1}^{s} m_{ij}(m, n) \ln p_{y_j}(t_{ij}) = \sum_{j=1}^{r} \sum_{i=1}^{s} m_{ij}(m, n) \ln p_{y_j}(t_i) \]

Q.E.D.

Proof of Theorem 3.2:

Note that (3.12) can be as:

\[
- \sum_{j=1}^{r} \sum_{i=1}^{s} \left[ v_i^{(m)} p_{y_j}^{(m)} \ln p_{y_j}(t_{ij}) + v_i^{(m+1)} p_{y_j}^{(m+1)} \ln p_{y_j}(t_{ij}) + \ldots + v_i^{(m+r_{c-1})} p_{y_j}^{(m+r_{c-1})} \ln p_{y_j}(t_{ij}) \right]
\]

\[
= - \sum_{j=1}^{r} \sum_{i=1}^{s} v_i p_{y_j}(t_{ij}) - \ldots - \sum_{j=1}^{r} \sum_{i=1}^{s} v_i^{(m+r_{c-1})} p_{y_j}^{(m+r_{c-1})} \ln p_{y_j}(t_{ij})
\]

Thus writing (3.12) in terms of \( G^*(m, m + r_{c-1}) \) defined in Subsection 3.1 leads to the equation

\[
- \ln \Pr\{T_{\tau}\} = G^*(m, m + r_{c-1}) - \sum_{j=1}^{r} \sum_{i=1}^{s} r_{c} \delta_{y_j}(m, m + r_{c-1}) \ln p_{y_j}(t_{ij}) \tag{12}
\]

where \( G^*(m, m + r_{c-1}) \) refers to the \( c \)th subsequence of length \( r_{c} \). To simplify the notation, define \( G^*_c = G^*(m, m + r_{c-1}) \). In a similar manner, \( \theta_y(\tau) = \theta_y(m, m + r_{c-1}) \). Therefore,

\[
- \ln \Pr\{T_{\tau}\} = G^*_c - \sum_{j=1}^{r} \sum_{i=1}^{s} r_{c} \delta_{y_j}(\tau) \ln p_{y_j}(t_{ij})
\]

Dividing (12) by \( r_{c} \).
\[
\frac{-\ln \Pr \{ T_\tau \}}{r_\tau} = \frac{G_\tau^*}{r_\tau} - \sum_{i=1}^{s} \sum_{j=1}^{s} \delta \theta_{y_i}(\tau) \ln \beta_{y_j}(t_i)
\]

where \( \eta_\tau' = -\sum_{i=1}^{s} \sum_{j=1}^{s} \delta \theta_{y_i}(\tau) \ln \beta_{y_j}(t_i) \). Therefore,

\[
\frac{-\ln \Pr \{ T_\tau \}}{r_\tau} = \frac{G_\tau^*}{r_\tau} = \eta_\tau'.
\]

Note that \( \eta_\tau' \) is a function of \( t_\tau \) and hence depends on the frequency and the ordering of the states in subsequence \( \tau \). In order to establish a bound that is independent of the particular subsequence, define

\[
\eta_\tau = \max_{t \in T_1} \left\{ -\sum_{i=1}^{s} \sum_{j=1}^{s} \delta \ln \beta_{y_j}(t_i) \right\}
\]

where the \( \theta_{y_i}(\tau) \) has been removed and the max is over all \( \tau \)th subsequences of \( T_1 \) sequences. Since \( |\theta_{y_i}(\tau)| \leq 1 \), then \( |\eta_\tau'| \leq |\eta_\tau| \) and therefore,

\[
\left| \frac{-\ln \Pr \{ T_\tau \}}{r_\tau} - \frac{G_\tau^*}{r_\tau} \right| \leq \eta_\tau
\]

Q.E.D.
Corollary to Theorem 3.2:

Taking the log of the products in (3.8) and from (12)

\[-\ln \Pr \{\Sigma_n\} = -\ln \left(v_{\sigma_0} \prod_{\tau} \Pr \{T_{\tau}\}\right)\]

\[= -\ln v_{\sigma_0} - \sum_{\tau} \ln \Pr \{T_{\tau}\}\]

\[= -\ln v_{\sigma_0} + \sum_{\tau} G^*_{\tau} - \sum_{\tau} \sum_{i=1}^T \sum_{j=1}^T r_{\tau} \delta \theta_{ij}(\tau) \ln p_{ij}(t_i)\]

\[= -\ln v_{\sigma_0} + \sum_{\tau} G^*_{\tau} + \sum_{\tau} r_{\tau} \eta'_{\tau}\]

If \(N\) is the length of the entire sequence of transitions, that is \(N = \sum_{\tau} r_{\tau}\), then dividing through by \(N\),

\[-\frac{\ln \Pr \{\Sigma_n\}}{N} - \frac{\sum_{\tau} G^*_{\tau}}{N} = -\frac{\ln v_{\sigma_0}}{N} + \frac{\sum_{\tau} r_{\tau} \eta'_{\tau}}{N}.\]

Consequently,

\[-\frac{\ln \Pr \{\Sigma_n\}}{N} - \frac{\sum_{\tau} G^*_{\tau}}{N} = \eta'_{\Sigma}\]

where \(\eta'_{\Sigma} = -\frac{\ln v_{\sigma_0}}{N} + \frac{\sum_{\tau} r_{\tau} \eta'_{\tau}}{N}\). Define \(\eta_{\Sigma} = -\frac{\ln v_{\sigma_0}}{N} + \frac{\sum_{\tau} r_{\tau} \eta_{\tau}}{N}\). Since \(\eta_{\tau}\) is independent of the ordering in each subsequence, \(\eta_{\Sigma}\) is also independent of the ordering in each subsequence. Also, since \(|\eta'_{\tau}| \leq |\eta_{\tau}|\), it follows that \(|\eta'_{\Sigma}| \leq |\eta_{\Sigma}|\). Therefore,

\[-\frac{\ln \Pr \{\Sigma_n\}}{N} - \frac{\sum_{\tau} G^*_{\tau}}{N} \leq \eta_{\Sigma}.\]

\[Q.E.D.\]
Proof of Theorem 3.3:

From the corollary to Theorem 1, recall that \(-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i\) for probability ensembles \(\{p_i\}\) and \(\{q_i\}\). From the definitions of \(g(k)\) and \(h(k)\) and using Lemma 3.2, then for all \(i\), \(g_i(k) \geq h_i(k)\). Consequently, for any vector \(v \geq 0\) and for any \(k\), \(v \cdot g(k) \geq v \cdot h(k)\).

Therefore, for state distribution vector \(v(k)\), \(\sum_{k=m}^{m+r-c-1} v(k) \cdot g(k) \geq \sum_{k=m}^{m+r-c-1} v(k) \cdot h(k)\). Since these summations correspond to the definitions of \(G^*(m,m+r_c-1)\) and \(H^*(m,m+r_c-1)\) respectively, \(G^*(m,m+r_c-1) \geq H^*(m,m+r_c-1)\). Since \(H^*(m,m+r_c-1)\) is dependent only on the transition matrix and cooling schedule and initial state distribution vector, it is constant among all realizations of the Markov process using a fixed cooling schedule and initial state distribution vector in the interval defined by time indices \(m\) and \(m+r_c-1\).

Q.E.D.

Proof of Corollary to 3.3:

Recall from Lemma 3.2 that

\[
\sum_{j=1}^{s} \sum_{i=1}^{s} m_{ij} (m+m+r_c-1) \ln p_{ij}(t) = \sum_{j=1}^{s} \sum_{i=1}^{s} m_{ij} (m+m+r_c-1) \ln p_{ij}(t)
\]

Substituting in (3.10) on both sides of this expression leads to the equation

\[
-\sum_{j=1}^{s} \sum_{i=1}^{s} \left[ v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m+r_c-1]} p_{ij}^{[m+r_c-1]} + r_c \delta_{ij} (m+m+r_c-1) \right] \ln p_{ij}(t)
\]

\[
= -\sum_{j=1}^{s} \sum_{i=1}^{s} \left[ v_i^{[m]} p_{ij}^{[m]} + \cdots + v_i^{[m+r_c-1]} p_{ij}^{[m+r_c-1]} + r_c \delta_{ij} (m+m+r_c-1) \right] \ln p_{ij}(t) \tag{13}
\]

From (12), the right-hand side of (13) can be written in terms of \(G^*(m,m+r_c-1)\) as

\[
G^*(m,m+r_c-1) - \sum_{i=1}^{s} \sum_{j=1}^{s} r_c \delta_{ij} (m+m+r_c-1) \ln p_{ij}(t) .
\]

The left-hand side of (13) can be expanded as
\[
\left[ -\sum_{i=1}^{s} \sum_{j=1}^{i} v_i^{[m]} p_{ij}^{[m]} \ln p_{ij}(t_y) - \ldots - \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m+r-1]} p_{ij}^{[m+r-1]} \ln p_{ij}(t_y) \right] - \sum_{i=1}^{s} \sum_{j=1}^{s} r_{ij} \delta \theta_j^{(m,m+r-1)} \ln p_{ij}(t_y)
\]

Recall from Lemma 3.1 that \( t_y \in [t_{a}, t_{m}] \). Consequently, as \( m \to \infty \), \( t_{m} \to 0 \) and, therefore, \( t_y \to 0 \). Therefore, the transition probabilities approach their limiting values. Therefore, taking the limit as \( m \to \infty \), the left-hand side of (13) can be written as:

\[
\lim_{m \to \infty} \left[ -\sum_{i=1}^{s} \sum_{j=1}^{i} v_i^{[m]} p_{ij}^{[m]} \ln p_{ij}(t_y) - \ldots - \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m+r-1]} p_{ij}^{[m+r-1]} \ln p_{ij}(t_y) \right] - \sum_{i=1}^{s} \sum_{j=1}^{s} r_{ij} \delta \theta_j^{(m,m+r-1)} \ln p_{ij}(t_y)
\]

\[
= \lim_{m \to \infty} \left[ -\sum_{i=1}^{s} \sum_{j=1}^{i} v_i^{[m]} p_{ij}^{[m]} \ln p_{ij}(t_y) - \ldots - \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m+r-1]} p_{ij}^{[m+r-1]} \ln p_{ij}(t_y) \right] - \sum_{i=1}^{s} \sum_{j=1}^{s} r_{ij} \delta \theta_j^{(m,m+r-1)} \ln p_{ij}(t_y)
\]

\[
= \lim_{m \to \infty} \sum_{i=1}^{s} \sum_{j=1}^{s} r_{ij} \delta \theta_j^{(m,m+r-1)} \ln p_{ij}(t_y) = -r_{\tau} \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m]} p_{ij}^{[m]} \ln p_{ij}^{[m]}
\]

But from the definitions of \( H(k) \) and \( H^{*}(m,m+r-1) \),

\[
\lim_{m \to \infty} H^{*}(m,m+r-1) = -r_{\tau} \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{[m]} p_{ij}^{[m]} \ln p_{ij}^{[m]}
\]

Substituting this into (13), the left-hand side of (13) becomes

\[
\lim_{m \to \infty} H^{*}(m,m+r-1) - \sum_{i=1}^{s} \sum_{j=1}^{s} r_{ij} \delta \theta_j^{(m,m+r-1)} \ln p_{ij}^{[m]}
\]

Therefore, taking the limit on both sides of (13) and dividing through by \( r_{\tau} \) leads to

\[
\lim_{m \to \infty} H^{*}(m,m+r-1) \over r_{\tau} - \sum_{i=1}^{s} \sum_{j=1}^{s} \theta_j^{(m,m+r-1)} \ln p_{ij}^{[m]} =
\]

\[
\lim_{m \to \infty} \frac{G^{*}(m,m+r-1)}{r_{\tau}} - \lim_{m \to \infty} \frac{\sum_{i=1}^{s} \sum_{j=1}^{s} \theta_j^{(m,m+r-1)} \ln p_{ij}(t_y)}{r_{\tau}}
\]

Since \( \delta \) can be set arbitrarily small,

\[
\lim_{m \to \infty} \frac{H^{*}(m,m+r-1)}{r_{\tau}} = \lim_{m \to \infty} \frac{G^{*}(m,m+r-1)}{r_{\tau}}
\]

\[Q.E.D.]
Proof of Theorem 3.4:

From Theorems 3.2 and 3.3, the following relationships hold for all sequences which exhibit convergence in distribution:

\[ e^{-(G_{min}^{*} + \eta)} \leq e^{-(H^{*}(m,n-2)+\eta)} \leq \Pr\{\Sigma(m,n-1)\} \leq e^{-(G_{min}^{*} + \eta)} \leq e^{-(H^{*}(m,n-2)+\eta)} \]  

(14)

where \( G_{min}^{*} = \min_{\Sigma_{m,n-1} \in \Gamma_{1}} G^{*}(m,n-2) \) and \( G_{max}^{*} = \max_{\Sigma_{m,n-1} \in \Gamma_{1}} G^{*}(m,n-2) \). Define the set \( \Gamma'_{1} = \left\{ \Sigma_{n} \mid e^{-(H^{*}(m,n-2)+\eta)} \leq \Pr\{\Sigma(m,n-1)\} \leq e^{-(G_{min}^{*} + \eta)} \right\} \), a subset of the set of annealing sequences where the last \( n-m \) states of a sequence of length \( n \) exhibits convergence in distribution.

Thus, \( \Gamma'_{1} \subseteq \Gamma_{1} \).

Define the set \( \Omega_{1} = \left\{ \Sigma_{n} \mid e^{-(H^{*}(m,n-2)+\eta)} \leq \Pr\{\Sigma(m,n-1)\} \leq e^{-(H^{*}(m,n-2)+\eta)} \right\} \) from the AEP for strongly ergodic inhomogeneous Markov chains (see Theorem 3.1). Therefore, \( \Gamma'_{1} \subseteq \Omega_{1} \) and \( \Gamma_{1} = \Gamma_{1} \cap \Omega_{1} \).

Since \( G^{*} \rightarrow H^{*} \) as \( m \rightarrow \infty \), (from the corollary to Theorem 3.3) this implies that as \( m \rightarrow \infty \), \( |\Omega_{1} \setminus (\Gamma'_{1} \cap \Omega_{1})| \rightarrow 0 \), i.e., the two sets become the same set. Therefore, considering the probability of each set, \( \lim_{m \rightarrow \infty} \Pr\{Z_{n} \in \Gamma'_{1}\} = \lim_{m \rightarrow \infty} \Pr\{Z_{n} \in \Omega_{1}\} \), or for any \( \epsilon' > 0 \), there exists a time index \( m \) sufficiently large such that \( \Pr\{Z_{n} \in \Omega_{1}\} - \Pr\{Z_{n} \in \Gamma'_{1}\} \leq \epsilon' \). But from the corollary to Theorem 3.1 and the criterion for membership in \( \Omega_{1} \), for any \( \epsilon > 0 \), there exists a time index \( n \) sufficiently large such that \( \Pr\{Z_{n} \in \Omega_{1}\} \geq 1 - \epsilon \). Therefore, for any \( \epsilon \) and \( \epsilon' > 0 \), there exists time indices \( m \) and \( n \) sufficiently large, such that \( \Pr\{Z_{n} \in \Gamma'_{1}\} \geq 1 - \epsilon - \epsilon' \).

In general, for any ensemble of events, the total probability of that ensemble is equal to the number of elements in the ensemble times the average probability of these elements. If \( \overline{P}_{\Gamma_{1}} \) is the average probability of the sequences in \( \Gamma_{1} \), then

\[ |\Gamma'_{1}| \overline{P}_{\Gamma_{1}} = \Pr\{Z(m,n-1) \in \Gamma'_{1}\} \geq 1 - \epsilon - \epsilon'. \]  

(15)
If a value larger than $\bar{P}_{t_1}$ is substituted into (15), then a lower bound on the size of the set $\Gamma_1'$ can be determined. From (14) and (15),

$$|\Gamma_1'| e^{-H^*(m,n-2)-n} \geq 1 - \varepsilon - \varepsilon' .$$

Consequently, for any $\varepsilon$ and $\varepsilon' > 0$, there exist time indices $m$ and $n$ sufficiently large such that

$$|\Gamma_1'| \geq (1 - \varepsilon - \varepsilon')e^{H^*(m,n-2)-n} = e^{H^*(m,n-2)+\delta'}$$

where $e^{\delta'} = (1 - \varepsilon - \varepsilon')e^{-n}$ and $\delta' > 0$ is small.

Q.E.D.

Proof of Theorem 3.5:

Define a connection matrix $A$ based on the transition matrix $P$ defined by (1.2) such that

$$a_{ij} = \begin{cases} 1 & \text{if } p_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Since all states have the same number of neighbors $N$, each row in the probability transition matrix defined by (1.2) all have $N+1$ positive elements—the $N$ neighbors plus the diagonal element. Therefore, the matrix $A$ has the same number of 1's in each row, namely, $N+1$ of them.

Define a new transition matrix $P'$ where for each 1 in $A$, there is some corresponding transition probability in $P'$ such that the entropy of $P'$ is maximized relative to matrix $P$. From Theorem 1, the maximum entropy for $N+1$ possible events occurs when all events have equal probability. Thus, for each 1 in $A$, and therefore each positive element in $P$, there is a transition probability in $P'$ equal to $1/(N+1)$. Therefore, the entropy $H'(k)$ for the matrix $P'$ is
\[ H'(k) = - \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{(k)} p_i^{(k)} \ln p_i^{(k)} \]

\[ = - \sum_{i=1}^{s} \sum_{j=1}^{s} v_i^{(k)} \frac{1}{(N+1)} \ln \frac{1}{(N+1)} \]

\[ = -(N+1) \sum_{i=1}^{s} v_i^{(k)} \ln (N+1) \]

\[ = \ln(N+1) \]

Since there is a correspondence between matrices \( \mathbf{P} \) and \( \mathbf{P}' \) in terms of positive values, and since the entropy associated with \( \mathbf{P}' \) has been maximized, then the entropy \( H(k) \) of \( \mathbf{P} \) is

\[ H(k) \leq H'(k) = \ln(N+1). \]

\[ Q.E.D. \]

**Proof of Theorem 3.6:**

This follows from the proof in Ash [1965 pp.210]. Let \( N(n) \) be the total number of possible sequences of length \( n \). In other words, \( N(n) \) is the number of sequences such that \( \Pr[\Sigma_n] > 0 \). Let \( N_f(n) \) equal the number of possible sequences when the initial state is \( s_i \).

Therefore, the following difference equation gives the number of possible sequences of length \( n+1 \) when the initial state is \( s_i \):

\[ N_i(n+1) = \sum_{j=1}^{s} a_{ij} N_j(n) = \mathbf{A} \mathbf{N} \]

where the \( a_{ij} \) and \( \mathbf{A} \) are defined in Theorem 3.5 and \( \mathbf{N} = (N_1(n), N_2(n), \ldots, N_s(n))^T \). This constitutes a system of linear, homogeneous difference equations with constant coefficients (see Goldberg [1986 pp.121]). Consequently, the solution for each \( N_i(n) \) is a linear combination of exponentials \( \lambda^n \). Thus, assuming a solution of the form

\[ N_i(n) = y_j \lambda^n, \]

then

\[ \lambda^n (\lambda y_j) = (\lambda 
\sum_{j=1}^{s} a_{ij} y_j = (\lambda \mathbf{A} y. \]
Therefore, \( \lambda y = Ay \) and the \( \lambda \)'s are the eigenvalues of \( A \). Since each row of \( A \) has \( N+1 \) ones, the value of \( \lambda \) is \( N+1 \) from the Perron-Frobenius Theorem on non-negative matrices (see Marcus and Minc [1964 pp.124]). Since

\[
N(n) \leq \sum_{i=1}^{r} N_i(n) = \sum_{i=1}^{r} y_i \lambda^n = \sum_{i=1}^{r} y_i (N+1)^n
\]

then there is some number \( K \) such that \( N(n) \leq K(N+1)^n \).

\[Q.E.D.\]

Proof of Lemma 4.1:

For a sequence of states of length \( r \) with a particular distribution of state pairs, the number of such state pairs is \( r-1 \), the number of transitions in the sequence. Therefore, the number of state-pair symbols is \( r-1 \). The number of distinct permutations of these symbols is given by the multinomial coefficient. Thus, if \( m_{ij} \) is the number of \( ij \) pairs, then the maximum number of distinct permutations \( \rho \) is given by

\[
\frac{(r-1)!}{\prod_{i,j} (m_{ij}(m_j+r-1))!} \leq r^{r-\max\{m_{ij}(m_j+r-1)\}} \tag{16}
\]

where \( \sum_{i,j} m_{ij}(m_j+r-1) = r-1 \). At this point, the fact that the stationary probabilities converge is useful. Note that

\[
\forall i > 0, \lim_{t \to 0} \pi_{S_i}^k(t) = \lim_{k \to \infty} \pi_{S_i}^{(k)} = 0
\]

and

\[
\lim_{t \to 0} \pi_{S_0}^k(t) = \lim_{k \to \infty} \pi_{S_0}^{(k)} = 1 \tag{17}
\]

(see Aarts and Korst [1989 pp.18].) Consequently, from (4.5), (4.6), and (4.8) and for small \( |\delta| > 0 \).
\[
\lim_{r \to \infty} \frac{m_{S_0S_0}(m, m+r-1)}{r} = \lim_{r \to \infty} \frac{\bar{E}\left\{m_{S_0S_0}(m, m+r-1)\right\} + r\delta}{r}
\]

\[= \lim_{r \to \infty} \frac{\sum_{k=m}^{m+r-1} \mathbb{P}^{[k]}(k) \left( \sum_{i=1}^{\mathbb{P}^{[k]}(k)} \left( \sum_{i=1}^{\mathbb{P}^{[k]}(k)} \mathbb{P}^{[k]}(k) \right) + r\delta \right)}{r}
\]

\[= \lim_{r \to \infty} \frac{\sum_{k=m}^{m+r-1} \mathbb{P}^{[k]}(k) \left( \sum_{i=1}^{\mathbb{P}^{[k]}(k)} \mathbb{P}^{[k]}(k) \right) + r\delta}{r}
\]  

(18)

Since the limit of a sequence equals the limit of its Cesàro averages and using (17), the first term in (18) is 1. Using (17), the second term is 0. The third term is a constant \(\delta\) which can be made arbitrarily small. Therefore, for \(r\) sufficiently large in SA, the value of

\[\max_{i,j} \{m_{ij}(m, m+r-1)\} = m_{S_0S_0}(m, m+r-1)\]

and

\[\lim_{r \to \infty} \frac{m_{S_0S_0}(m, m+r-1)}{r} = 1.
\]

Therefore, for any \(\varepsilon > 0\), there exists an \(r\) such that \(m_{S_0S_0}(m, m+r-1) \geq r(1-\varepsilon)\). Substituting this into (16), \(r \leq r^{-\varepsilon \ln(1-\varepsilon)} = r^\varepsilon\).

\[Q.E.D.\]

**Proof of Lemma 4.2:**

From the definition of \(c_{ij}\), it is possible to define an upper bound on the number of combinations of transition/pairs. This number is an upper bound because the constraint that the sum of the numbers selected from each bin add up to the length of the subsequence is ignored. Thus, an upper bound on maximum number of different combinations, \(C\), is the product of all the \(c_{ij}\)'s. Therefore, \(C = c_{00} \cdot c_{01} \cdot c_{10} \cdot c_{11}\) In general then, the maximum number of combinations can be stated as

\[C = \prod_{i} \prod_{j} c_{ij} = \prod_{i} \prod_{j} (u_{ij} - l_{ij} + 1).
\]  

(19)
Each of these combinations therefore has a specific number of state pairs. For example, there are $c_{01} \cdot c_{10} \cdot c_{11}$ combinations with $l_{00}$ pairs of $S_0 S_0$. Each of these combinations has, at most, $r^{r^e}$ permutations by virtue of Lemma 4.1, and therefore correspond to $r^{r^e}$ different typical subsequences with, at most, $l_{00} \rho/r - 1$ of them with $S_0 S_0$ as the final transition/pair. Counting all the possible combinations with $S_0 S_0$ state pairs and then choosing $\rho/(r - 1)$ of them as sequences with $S_0 S_0$ as the final state pair, an upper bound $f_{00}$ on the total number of such subsequences with $S_0 S_0$ as the final transition/pair is

\[
f_{00} = \frac{\rho}{r - 1} \sum_{i=0}^{\infty} (i) \cdot c_{01} \cdot c_{10} \cdot c_{11} \]

\[
= \frac{\rho}{r - 1} (c_{01} \cdot c_{10} \cdot c_{11}) \sum_{i=0}^{\infty} i \]

\[
= \frac{\rho}{r - 1} (c_{01} \cdot c_{10} \cdot c_{11}) \left[ \frac{(l_{00} + \mu_{00})(\mu_{00} - l_{00} + 1)}{2} \right] \]

\[
= \frac{\rho C}{r - 1} \left( \frac{l_{00} + \mu_{00}}{2} \right) \quad (20) \]

Similarly, the number of subsequences with $S_1 S_0$ as the final transition/pair is

\[
f_{10} = \frac{\rho}{r - 1} \sum_{i=0}^{\infty} i \cdot c_{00} \cdot c_{01} \cdot c_{11} \]

\[
= \frac{\rho C}{r - 1} \left[ \frac{l_{10} + \mu_{10}}{2} \right] \quad (21) \]

Thus, the upper bound on the number of subsequences with $S_0$ as the final state, is therefore $a_{S_0} = f_{00} + f_{10}$. Similarly, an upper bound on the number of subsequences with $S_1$ as the final state is $a_{S_1} = f_{01} + f_{11}$. Since $a_{S_0}, a_{S_1}$ are functions of the $f_{00}, f_{01}, f_{10}, f_{11}$, the order of the $f_{00}, f_{01}, f_{10}, f_{11}$ must be determined. These in turn, are functions of the $c_{00}, c_{01}, c_{10}, c_{11}$. From its definition and using (4.5),
\[ c_y = \left\lfloor \left\lfloor E\{ m_y (m, m+r-1) \} + r\delta \right\rfloor - \left\lfloor E\{ m_y (m, m+r-1) \} - r\delta \right\rfloor \right\rfloor + 1 \]
\[ \leq \left\lfloor E\{ m_y (m, m+r-1) \} \right\rfloor + \left\lfloor r\delta \right\rfloor - \left\lfloor \left\lfloor E\{ m_y (m, m+r-1) \} \right\rfloor - r\delta \right\rfloor + 1 \]
\[ \leq \left\lfloor r\delta \right\rfloor + \left\lfloor r\delta \right\rfloor + 1 \]

Consequently, the \( c_{00}, c_{01}, c_{10}, c_{11} \) are \( O(r) \) and therefore \( C \in O(r^4) \). Now consider the order of the \( l_{00}, \ldots, l_{11} \). From their definitions, and (4.5) and (4.6),

\[ l_{01} = l_{10} = \left\lfloor E\{ m_{S_{0}S_{0}} (m, m+r-1) \} - r\delta \right\rfloor \]
\[ = \left\lfloor \pi_{1i}^{[m]} p_{S_{1}S_{0}}^{[m]} + \pi_{1i}^{[m+1]} p_{S_{1}S_{0}}^{[m+1]} + \ldots + \pi_{1i}^{[m+r-1]} p_{S_{1}S_{0}}^{[m+r-1]} - r\delta \right\rfloor \]
\[ = \left\lfloor \frac{1}{N} \left( \pi_{1i}^{[m]} + \pi_{1i}^{[m+1]} + \ldots + \pi_{1i}^{[m+r-1]} \right) - r\delta \right\rfloor \]

The complexity of these factors is therefore related to the sum of the stationary probabilities at increasing time indices. Using (17), (4.7) and (4.8), \( l_{00}, l_{01}, l_{10}, l_{11} \in O(r) \) and using similar argument, \( u_{00}, u_{01}, u_{10}, u_{11} \in O(r) \). Consequently, from (20) and (21), the \( f_{00}, f_{01}, f_{10}, f_{11} \in O(r^{ec+4}) \). Since \( a_{s_0} = f_{00} + f_{10} \) and \( a_{s_1} = f_{01} + f_{11} \), then \( a_{s_0}, a_{s_1} \in O(r^{ec+4}) \).

Proof of Theorem 4.1:

Define the following:

\[ a'_{s_0} = \text{the number of sequences in set } \Gamma'_i \text{ that has } S_0 \text{ as the final state.} \]

\[ a'_{s_1} = \text{the number of sequences in set } \Gamma'_i \text{ that has } S_1 \text{ as the final state.} \]

Using these definitions, the average value of the final states in set \( \Gamma'_i \) can be expressed as

\[ \left( f^{[m+r-1]} \right)_{\Gamma'_i} = \frac{a'_{s_0} f_{s_0} + a'_{s_1} f_{s_1}}{e^{H(m, m+r-1)+\delta}} \]

Using the above analysis and Theorem 3.4 and for some number \( k > 1 \) and \( f \) finite and equal to the mean value of the final state of the atypical sequence, an upper bound on the expected value of the final state can be expressed as
\[ \left( f^{(m-\xi+1)} \right) \leq \frac{a_{s_k} f_{s_0} + a_{s_k} f_{s_k}}{e^{H^m(m, a_{s_k} + \delta)}} (1 - e^{-e'}) + f'(e' + e') \leq k \left( \frac{a_{s_k} f_{s_0} + a_{s_k} f_{s_k}}{e^{H^m(m, a_{s_k} + \delta)}} \right) \] (22)

Since the numerator of the right-hand side is \( O(r^{e' + t}) \), the theorem is proved.

Q.E.D.

Proof of Theorem 4.2:

From the definition of the mean objective function value, we have
\[
\left( f(t_m) \right) = \sum_{i=0}^{S_i} \pi_{s_i}^{[m]} f_{s_i} = \sum_{i \leq k} \pi_{s_i}^{[m]} f_{s_i} + \sum_{i > k} \pi_{s_i}^{[m]} f_{s_i} \text{ for some } k, \tag{23}
\]

Thus, the mean objective function is some convex combination of all the objective function values. Recall that the objective function values are ordered, that is \( f_{s_0} \leq f_{s_1} \leq \cdots \leq f_{s_k} \leq f_{s_{k+1}} \leq \cdots \leq f_{s_{S_i}} \). Thus, dividing the first summation by \( \sum_{i \leq k} \pi_{s_i}^{[m]} \) makes the first term a convex combination of the first \( k \) objective function values. But
\[
\sum_{i \leq k} \pi_{s_i}^{[m]} f_{s_i} \sum_{i \leq k} \pi_{s_i}^{[m]} = f_{s_k}^{[m]}
\]

Therefore, since a convex combination of values is intermediate of the extreme values of that combination, \( f_{s_k}^{[m]} < f_{s_k} \). In similar fashion, define
\[
\tilde{f}_{s_k}^{[m]} = \sum_{i > k} \pi_{s_i}^{[m]} f_{s_i} \sum_{i > k} \pi_{s_i}^{[m]} \]

Thus, \( f_{s_k} < \tilde{f}_{s_k}^{[m]} \). Define \( \tilde{\pi}_{s_k}^{[m]} = \sum_{i > k} \pi_{s_i}^{[m]} \). Thus, each summation in (23) can be written as the convex combination of two values. Therefore,
\[
\left( f(t_m) \right) = \sum_{i \leq k} \pi_{s_i}^{[m]} f_{s_i} + \sum_{i > k} \pi_{s_i}^{[m]} f_{s_i} = \tilde{\pi}_{s_k}^{[m]} f_{s_k}^{[m]} + \tilde{\pi}_{s_k}^{[m]} \tilde{f}_{s_k}^{[m]} \text{ where } \tilde{\pi}_{s_k}^{[m]} + \tilde{\pi}_{s_k}^{[m]} = 1.
\]

Consequently.
\[ \langle f(t_m) \rangle - \bar{x}_k = \bar{x}_k \bar{x}_k^* + \bar{x}_k \bar{x}_k^* - \bar{x}_k \]

\[ = (\bar{x}_k^* - 1)\bar{x}_k + \bar{x}_k \bar{x}_k^* \]

\[ = \bar{x}_k \bar{x}_k^* - \bar{x}_k \bar{x}_k^* \]

\[ = \bar{x}_k \bar{x}_k^* \]

Since \( \bar{x}_k^* > \bar{x}_k^* \), then

\[ \langle f(t_m) \rangle - \bar{x}_k = \bar{x}_k \bar{x}_k^* (\bar{x}_k - \bar{x}_k^*) > 0 \]

\[ (f(t_m)) > \bar{x}_k^* \]

Q.E.D.

Proof of Theorem 4.3:

From the definition of the stationary probability of the super-node,

\[ \frac{\partial \bar{x}_k}{\partial t} = \frac{\partial}{\partial t} \sum_{i \in S} \pi_i \]

\[ = \sum_{i \in S} \pi_i \frac{\partial}{\partial t} \left( f_{Si} - \langle f(t_m) \rangle \right) \]

from (4.10),

\[ = \sum_{i \in S} \pi_i \left( \frac{f_{Si}^*}{t_m^2} \right) - \sum_{i \in S} \pi_i \left( \frac{\langle f(t_m) \rangle}{t_m^2} \right) \]

\[ = \frac{\bar{x}_k^*}{t_m^2} \]

\[ = \frac{\bar{x}_k^*}{t_m^2} \]

\[ = \frac{\bar{x}_k^*}{t_m^2} \]

(24)

Since from Theorem 4.2, \( \bar{x}_k^* < \langle f(t_m) \rangle \), then \( \partial \bar{x}_k / \partial t_m < 0 \) and therefore \( \bar{x}_k^* \) increases with decreasing \( t_m \).

Q.E.D.
APPENDIX B

The following code written in C implements these ideas for updating the state probabilities. The value of curr_index corresponds to the address in the solution list of the current solution, and neighbor_index corresponds to the address of neighbor number i in which the neighbor numbers range from 1 to nhood_size, the neighborhood size.

```c
float update_state_vector_element(curr_index, nhood_size, temp)
int curr_index, nhood_size;
float temp;
{
int i, neighbor_index;
float delta_row, delta_col;
float diag_element, pij, pji;
float sum=0.0;
float dot_prod=0.0;
for(i=1; i<nhood_size; i++)
{
neighbor_index = calc_neigh_index(i, curr_index);
delta_row = *(soin_ptr + neighbor_index) - *(soin_ptr + curr_index);
delta_row = (delta_row >= 0.0) ? 0.0 : delta_row;
delta_col = *(soin_ptr + curr_index) - *(soin_ptr + neighbor_index);
delta_col = (delta_col >= 0.0) ? 0.0 : delta_col;
pij = exp(delta_row/temp)/((float)nhood_size;
pji = exp(delta_col/temp)/((float)nhood_size;
sum *= pij;
dot_prod += *(state_ptr + neighbor_index)*pij;
}
diaq_element = 1.0 - sum;
dot_prod += *(state_ptr + curr_index)*diaq_element;
return(dot_prod);
}
```

Proof Noted in Subsection 7.2.7:

Consider a graph $G$ in which the current set of nodes $\alpha = V'$ constitutes the solution to the independent set problem run on one computer and $\beta = VV'$ is the current set of nodes for the vertex cover problem being run on a second computer running in parallel and both using the same random number streams (common random numbers). Thus, $\alpha \cap \beta = \emptyset$. Let $v* \in \alpha$ be a candidate node selected by both computers. Thus, $v* \in \beta$. Let $m$ edges connect $v'$ to nodes in $\alpha$ and $n$ edges connect to nodes in $\beta$. This is depicted below.
Therefore, $f_{\text{indep set}}(V) = |\alpha| - \lambda E_j(\alpha)$ and $f_{\text{vertex cover}}(V) = |\beta| + \lambda E_j(\beta)$ by the identity in (5.7). The values of these objective functions if node $v'$ is flipped are

$$f_{\text{indep set}}(V) = |\alpha| + 1 - \lambda E_j(\alpha) - \lambda m,$$

$$f_{\text{vertex cover}}(V) = |\beta| - 1 + \lambda E_j(\alpha) + \lambda m.$$ 

Therefore, the value of $\Delta f_{\text{indep set}}(V) = 1 - \lambda m$ and the value of $\Delta f_{\text{vertex cover}}(V) = -1 + \lambda m$. Recall from (1.1) that for a minimization problem, the candidate solution will be accepted with probability $e^{-\Delta f/T}$, but for maximization problems, this probability of acceptance is $e^{\Delta f/T}$, where

$$\Delta f = \begin{cases} \Delta f & \text{if } \Delta f < 0 \\ 0 & \text{otherwise} \end{cases}.$$ 

Therefore, for the above problems, $v'$ will be accepted into the current solution for the independent set with probability

$$Pr(\text{accept into independent set}) = \begin{cases} e^{(1 - \lambda m)/T} & \forall m > 0 \\ 1 & \text{otherwise} \end{cases} \quad \text{where } \lambda > 1.$$ 

For the vertex cover, the node will leave the current solution with probability

$$Pr(\text{leaving from vertex cover}) = \begin{cases} e^{(-1 + \lambda m)/T} & \forall m > 0 \\ 1 & \text{otherwise} \end{cases} \quad \text{where } \lambda > 1.$$ 

Since $Pr(\text{accept into independent set}) = Pr(\text{leaving from vertex cover})$, the node $v'$ will leave one set and enter the other with equal probability. By similar reasoning, we have $Pr(\text{leaving from independent set}) =$
$Pr(\text{accept into vertex cover})$. Since both computers are using common random numbers, the node $v'$ will either remain in the vertex cover and not enter the independent set, or it will move from one set to the other. Thus, if $\alpha \cap \beta = \emptyset$ at time $k$, then $\alpha \cap \beta = \emptyset$ at time $k + 1$.

Q.E.D.
REFERENCES


148


