Proceedings of 24th Conference
on Decision and Control
Ft. Lauderdale, FL • December 1985

CONVERGENCE AND FINITE-TIME BEHAVIOR OF SIMULATED ANNEALING

Debasis Mitra
AT&T Bell Laboratories
Murray Hill, NJ 07974

Fabio Romeo and Alberto Sangiovanni-Vincentelli
Department of EECS, University of California
Berkeley, CA 94720

Abstract

Simulated Annealing is a randomized algorithm which has been proposed for finding globally optimum least-cost configurations in large NP-complete problems with cost functions which may have many local minima. A theoretical analysis of Simulated Annealing based on its precise model, a time-inhomogeneous Markov chain, is presented. An annealing schedule is given for which the Markov chain is strongly ergodic and the algorithm converges to a global optimum. The finite-time behavior of Simulated Annealing is also analyzed and a bound obtained on the departure of the probability distribution of the state at finite time from the optimum. This bound gives an estimate of the rate of convergence and insights into the conditions on the annealing schedule which gives optimum performance.

1. Introduction

Many combinatorial optimization problems belong to the class of NP-complete problems [1]. Heuristic search algorithms for these problems often terminate at a local minimum. To avoid this behavior, a class of randomized algorithms (e.g. [2]) have been devised which generate the next configuration randomly, and which can "climb hills", i.e., moves that generate configurations of higher cost than the present one are accepted. Simulated annealing as proposed by Kirkpatrick et al. [3-5], allows "hill climbing" moves but these moves are accepted according to a certain criterion which takes the state of the search process into consideration in a manner unlike other randomized algorithms.

In applications to combinatorial optimization, this method starts from an arbitrary configuration and, given that the simulation is at configuration i at time m, $m = 0, 1, 2, \ldots$, a new configuration j is randomly generated from an admissible set of neighbouring configurations $N(i)$ and a check is made to determine whether the cost of the new configuration satisfies an acceptance criterion based on the temperature, a controlling parameter, at time m, T_m. If the cost decreases, the simulation accepts the move. Otherwise, a random number uniformly distributed over $[0,1]$ is picked and compared with $e^{-[c(j)-c(i)]/T_m}$, where $c(\cdot)$ is the cost function on configurations. If the random number is smaller, the simulation accepts the move, otherwise it discards the move. In any case, time is incremented. Note that the higher the temperature is, the more likely it is that a "hill climbing" move is accepted. The initial temperature, the number of moves generated at each temperature and the rate of decrease of temperature are all important parameters that affect the speed of the algorithm and the quality of the final configuration.

Experimental results [3, 6-8] show that Simulated Annealing produces very good results when compared to other techniques for the solution of combinatorial optimization problems such as those arising from the layout of integrated circuits, at the expense, however, of large computation time (a 1,500 standard cell placement problem can take as much as 24 hours of a VAX 11/780 [7]). This has emphasized the need for a better theoretical understanding of Simulated Annealing.

The analysis in this paper is based on time-inhomogeneous Markov chains [9-11]. We prove that for an arbitrary but bounded cost function, for annealing schedules of the form

$$T_m = \frac{\gamma}{\log(m + m_0 + 1)}, \quad m = 0,1,2,\ldots \tag{1.1}$$

where m_0 is any parameter satisfying $1 \leqslant m_0 < \infty$, the Markov chain is *strongly* ergodic if

$$\gamma \geqslant r\,L,$$

where r is the radius of the graph underlying the chain and L is a Lipschitz-like constant of the cost function. Strong ergodicity implies that, for any starting probability vector, the state probability vector converges component-wise to a constant vector $\mathbf{e}^*$. Furthermore, we show that $\mathbf{e}^*$ is the *optimum* vector, i.e., the vector in which all elements are zero except those with the indices of the global least-cost configurations. Our other main result is on finite-time behavior and rate of convergence. We give a bound on the departure of the state vector from the optimum vector after a *finite* number of iterations. This bound indicates how the annealing schedule must be balanced between contrary requirements for optimum performance. A simple Corollary to this result states that for large number of iterations k, the L_1-norm of the difference of the state vector from the optimum vector is $O(1/k^{\min(a,b)})$, where a and b respectively increase and decrease with increasing γ.

We also obtain a set of results on distributions which we call quasi-stationary. These constructs are the equilibrium distributions of time-homogeneous Markov chains obtained from Simulated Annealing by holding the temperature fixed at various values. The dependence of the quasi-stationary distributions on temperature is shown to have a number of desirable properties. These properties are essential for our analysis of the time-inhomogeneous Markov chains obtained from annealing schedules given in (1.1). In addition, they are of independent interest since they hold for annealing schedules considerably more general than (1.1). This may be important in the future if, as we expect, it becomes possible to design schedules matched to special properties of the cost function.

Early analyses [12,13] using time-homogeneous Markov chains made certain unrealistic assumptions on the number of iterations taken at each temperature. As in this paper, recent works [14-16] based on time-inhomogeneous Markov chains take full account of the time dependence of the temperature. The reader will find in ref. 20 the proofs of various assertions which have had to be omitted here.

2. Preliminaries

In this section, we describe the basic structure of the Simulated Annealing algorithm and we introduce a Markov chain model for it.

In the algorithm structure [20] three functions play a fundamental role: **accept**, **generate** and **update**. While several **accept** functions can be used [13], in this paper we restrict our attention to the one proposed in ref. 3 which has been described in Section 1.

The **generate** function selects a new configuration. In Simulated Annealing, a new configuration is generated randomly from a set of possible configurations. To completely specify this function, a set of configurations accessible from a given configuration and the probability of generating one of these has to be given.

The **update** function, also called the *annealing* schedule or *cooling* schedule, produces a new value for the temperature. This function is most important to determine the convergence properties of the algorithm. We focus on **update** functions which return monotonically decreasing values of the temperature, i.e. $\forall m \geqslant 0$, $T_{m+1} < T_m$ and $\lim_{m\to\infty} T_m = 0$. The function is completely specified when the explicit dependency of T on m is given. This paper is devoted to the study of **update** functions that guarantee convergence of the algorithm to the optimum vector.

CH2245-9/85/0000-0761 $1.00 © 1985 IEEE

It is easy to see that Simulated Annealing can be represented by a Markov chain, whose connectivity is fully specified by the **generate** function and whose transition probabilities are determined by the **accept** and by the **generate** functions.

The underlying directed graph, which we denote by G, is determined as follows. There is a bijective correspondence between the elements of S, the set of all the possible configurations of the optimization problem, and the nodes of the graph. Given two different elements, say i and j, of S, there is an arc from i to j if j can be generated starting from i. The two nodes are said to be *neighbors*. We define $N(i)$ to be the set of all the neighbors of i. We assume that $i \notin N(i)$. In several applications of the Simulated Annealing algorithm, the probability of generating a particular neighboring configuration starting from i is simply given by $1/|N(i)|$ where $|N(i)|$ is the cardinality of $N(i)$. However, in certain applications such as placement of integrated circuits [7], it is important to generate certain neighbors with higher probability. For this reason, we assume that the probability of generating j from i is given by

$$g(i,j)/g(i) \tag{2.1}$$

where $g(i,j)$ gives the "weights" for each of the neighbors of i and $g(i)$ is a normalizing function which ensures that $\frac{1}{g(i)} \sum_{j \in N(i)} g(i,j) = 1$. The directed graph G is assumed to be connected.

The one-step transition probabilities of the Markov chain are represented as weights on the edges of the directed graph G defined above and are determined by the product of the probability of generating a given configuration and the probability of accepting it. We define first a one-parameter family of transition probabilities:

$$\mathbf{P}_{ij}(T) = \begin{cases} 0 & \text{if } j \notin N(i) \text{ and } j \neq i \\ \frac{g(i,j)}{g(i)} \min[1, e^{-\{c(j)-c(i)\}/T}] & \text{if } j \in N(i) \end{cases} \tag{2.2}$$

and

$$\mathbf{P}_{ii}(T) = 1 - \sum_{j \in N(i)} \mathbf{P}_{ij}(T) \tag{2.3}$$

The transition probabilities of the time-inhomogeneous Markov chain in which m denotes discrete time are obtained from the above and the annealing schedule which specifies $T = T_m$, $m = 0, 1, 2, \ldots$.

3. Quasi-Stationary Probability Distributions and Their Properties

We have shown that a mathematical model for Simulated Annealing is a time-inhomogeneous Markov chain. However, if the temperature is frozen at a particular value T, then we obtain a time-homogeneous Markov chain. To prove the convergence of Simulated Annealing, it is important to study this Markov chain. In particular, we show here that this chain has a *stationary* probability distribution, which we call the *quasi-stationary* probability distribution of the time-inhomogeneous Markov chain. In addition, we show that the stationary probability distributions have a limit when T goes to zero, i.e. when m goes to infinity, and that this limit is the optimum vector $\mathbf{e}^*$.

3.1 The Quasi-Stationary Probabilities.

For $i \in S$ define

$$\boldsymbol{\pi}_i(T) \triangleq \frac{g(i)e^{-c(i)/T}}{G(T)} \tag{3.1}$$

where $G(T)$ is a scaling factor such that $||\boldsymbol{\pi}(T)|| = 1$ where

$$||\boldsymbol{\pi}(T)|| \triangleq \sum_{i=1}^{s} \boldsymbol{\pi}_i(T).$$

and $s = |S|$.

The role of $G(T)$ is similar to that of the partition function in statistical mechanics and stochastic networks [17].

We now show that $\boldsymbol{\pi}(T)$ is the stationary probability distribution for the time-homogeneous Markov chain. For this to be true we need to assume that the function $g(i,j)$ is symmetric, i.e.,

$$g(i,j) = g(j,i), \quad \forall i, j \in S. \tag{3.2}$$

This is a mild restriction which is easy to satisfy in implementations of Simulated Annealing. In particular, symmetry exists in the case where all neighbours of each configuration are given equal weights.

Proposition 3.1. If (3.2) holds, then $\{\boldsymbol{\pi}_i(T)\}$ defined by (3.1), satisfies

$$\boldsymbol{\pi}(T)\mathbf{P}(T) = \boldsymbol{\pi}(T), \quad m = 0,1,\ldots \tag{3.3}$$

where $\mathbf{P}(T)$ is the one-step transition probability matrix of the Markov chain defined in (2.2) and (2.3). □

The proof is in [20]. It is of some interest to note that detailed balance is equivalent to the time-reversibility [17] of the time-homogeneous Markov chain.

3.2 Asymptotic Quasi-Stationary Probabilities.

The results in this section and Sections 3.3-3.4. hold for any **update** function in which

$$T_m > T_{m+1}, \quad \forall m \geqslant 0, \tag{3.4.a}$$

$$\lim_{m \to \infty} T_m = 0. \tag{3.4.b}$$

It should be emphasized that here and in Sections 3.3-3.4 we are investigating the dependence of $\boldsymbol{\pi}(T_m)$ on T_m where $\{T_m\}$ behaves as in (3.4), and that $\boldsymbol{\pi}(T_m)$ is a construct and not the distribution obtained from Simulated Annealing.

The following result has a straightforward proof.

Proposition 3.2. If the **update** function satisfies (3.4.b), then the *quasi-stationary probability vector* $\boldsymbol{\pi}(T_m)$ defined in (3.1) converges, as $m \to \infty$, to the optimal vector $\mathbf{e}^*$

$$\mathbf{e}^*_i = \begin{cases} g(i)/g(*) & i \in S^* \\ 0 & i \notin S \end{cases} \tag{3.5}$$

where S^* is the set of indices of global least-cost configurations, i.e.

$$S^* \triangleq \{i \in S | c(i) \leqslant c(j) \; \forall j \in S\}$$

and $g(*) \triangleq \sum_{j \in S^*} g(j)$. □

3.3 Monotonicity of the Quasi-Stationary Probabilities

The convergence of the quasi-stationary distributions to $\mathbf{e}^*$ displays remarkable monotonicity properties. This property is insightful and also an essential element of the analysis of the asymptotic and finite-time behavior of Simulated Annealing. The proof of the following Proposition is in [20].

We will need to identify the "weighted mean cost" to be denoted by C and defined thus

$$C \triangleq \sum_{j \in S} g(j)c(j) / \sum_{j \in S} g(j). \tag{3.6}$$

Proposition 3.3.

(i) For each $i \in S^*$,

$$\boldsymbol{\pi}_i(T_{m+1}) - \pi_i(T_m) > 0 \quad \forall m \geqslant 0.$$

(ii) For each $i \notin S^*$, there exists an unique integer $\hat{m}_i$, $0 \leqslant \hat{m}_i < \infty$, such that

$$\begin{aligned} \boldsymbol{\pi}_i(T_{m+1}) - \pi_i(T_m) &> 0 \quad 0 \leqslant m \leqslant \hat{m}_i - 1 \\ &< 0 \quad m \geqslant \hat{m}_i. \end{aligned}$$

□

An immediate corollary to Proposition 3.3 is the existence of $\tilde{m}$, $\tilde{m} < \infty$, such that for all $i \notin S^*$,

$$\boldsymbol{\pi}_i(T_{m+1}) - \boldsymbol{\pi}_i(T_m) < 0, \quad \forall m \geqslant \tilde{m}. \tag{3.7}$$

In fact

$$\tilde{m} = \max_{i \notin S^*} \hat{m}_i . \tag{3.8}$$

3.4 Uniform Monotonicity of the Quasi-Stationary Probabilities

The analysis in Section 6 on finite time behavior requires knowledge of $\tilde{m}$ which marks the onset of monotonic decrease of quasi-stationary probabilities of all but the least cost configurations. We show here how it may be identified. This is done by considering $\hat{T}_i$, for $i \notin S^*$ and $c(i) < C$, as functions of the cost associated to each state $\{c(j)\}$. The proof of the following Proposition is in [20].

Proposition 3.4. For all i such that $i \notin S^*$ and $c(i) < C$, $\hat{T}_i$ are monotonic, strictly increasing with increasing $c(i)$. □

Note that configurations with common cost have common values of $\hat{T}_i$ and $\hat{m}_i$.

To calculate $\tilde{m}$ it is helpful to identify the least-cost and the next-to-least cost of all the configurations. Let

$$c(*) \triangleq \min_{j \in S} c(j) \tag{3.9}$$

$$\text{and} \qquad \delta \triangleq \{\min_{j \notin s^*} c(j)\} - c(*) , \tag{3.10}$$

so that $c(*)$ and $\{\delta + c(*)\}$ are respectively the least-cost and the next-to-least cost. Note that δ is an important global characteristic of the cost function.

The monotonicity property in Proposition 3.4 allows (3.8) to be sharpened: $\tilde{m} = \hat{m}_{\bar{i}}$ where $\bar{i}$ is any configuration with next-to-least cost. Let $\tilde{T}$ be the unique positive solution of the following equation

$$\delta g(*) - \sum_{j : c(j) > \delta + c(*)} g(j) \{c(j) - c(*) - \delta\} e^{-\{c(j) - c(*)\}/T} = 0, \tag{3.11}$$

where $g(*)$ is given in Proposition 3.2. Then $\tilde{m}$ is the smallest integer such that $T_{\tilde{m}} \leqslant \tilde{T}$.

We conclude this section by a summary. The quasi-stationary probability distribution converges with decreasing temperature (i.e. increasing time) to the optimum vector. The quasi-stationary probabilities of least-cost configurations monotonically increase with decreasing temperature. For configurations with costs not less than the weighted mean cost, the opposite is true. Each configuration i with cost between least-cost and weighted mean cost has an associated "critical temperature" $\hat{T}_i$; while the temperature is greater than $\hat{T}_i$, the configuration's quasi-stationary probability increases with decreasing temperature, and for temperatures less than $\hat{T}_i$ the opposite is true. Furthermore, the critical temperature is an increasing function of cost. All of the above properties hold for any **update** function satisfying (3.4).

4. Time-Inhomogeneous Markov Chains

In this section a number of well known properties of time-inhomogeneous Markov chains are presented. These results will be used in Section 5 to prove the convergence properties of the Simulated Annealing algorithm and to determine the influence of the annealing schedule on the rate of convergence to the optimal solution of the combinatorial optimization problem.

All theorems and propositions are given without proof. The interested reader can find these proofs in refs. 9-11.

4.1 Notation.

For the sake of notational simplicity, from now on all vectors, matrices and functions depending on T_m will be denoted as depending on m. Let $\mathbf{P}(m,m)$ be the identity matrix, and

$$\mathbf{P}(m, n+m) \triangleq \prod_{i=0}^{n-1} \mathbf{P}(m+i), \quad m \geqslant 0,\ n \geqslant 1$$

be the n-step transition matrix. Furthermore let

$$\boldsymbol{\nu}(m) \triangleq [\boldsymbol{\nu}_1(m), \boldsymbol{\nu}_2(m), \ldots, \boldsymbol{\nu}_s(m)]$$

denote the state probability vector after m transition of the Markov chain, so that

$$\boldsymbol{\nu}(m+n) = \boldsymbol{\nu}(m)\mathbf{P}(m, m+n) .$$

We also let

$$\boldsymbol{\nu}(m,n) = \boldsymbol{\nu}(0)\mathbf{P}(m,n) .$$

4.2 Basic Results from the Theory of Time-Inhomogeneous Markov Chains.

We need the following definition

Definition 4.1. A time-inhomogeneous Markov chain is *weakly ergodic* if for all m,

$$\lim_{n \to \infty} \sup_{\boldsymbol{\nu}^1(0),\, \boldsymbol{\nu}^2(0)} \|\boldsymbol{\nu}^1(m,n) - \boldsymbol{\nu}^2(m,n)\| = 0 \tag{4.1}$$

where $\boldsymbol{\nu}^1(0)$ and $\boldsymbol{\nu}^2(0)$ are two arbitrary initial state probability vectors and

$$\boldsymbol{\nu}^1(m,n) = \boldsymbol{\nu}^1(0)\mathbf{P}(m,n)$$

$$\boldsymbol{\nu}^2(m,n) = \boldsymbol{\nu}^2(0)\mathbf{P}(m,n) .$$

□

Note that weak ergodicity does not imply the existence of limits of vectors $\boldsymbol{\nu}^1(m,n)$ and $\boldsymbol{\nu}^2(m,n)$. The investigation of conditions under which weak ergodicity holds is aided by the introduction of the following coefficient of ergodicity.

Definition 4.2. Given a stochastic matrix $\mathbf{P}$, its coefficient of ergodicity τ_1 is

$$\tau_1(\mathbf{P}) = \frac{1}{2} \max_{i,j} \sum_{k=1}^{s} |\mathbf{P}_{ik} - \mathbf{P}_{jk}| = 1 - \min_{i,j} \sum_{k=1}^{s} \min(\mathbf{P}_{ik}, \mathbf{P}_{jk}) . \tag{4.2}$$

□

With the above definition of the coefficient of ergodicity the following result can be proved [9-11].

Theorem 4.1. The time-inhomogeneous Markov chain is weakly ergodic if and only if there is a strictly increasing sequence of positive integers $\{k_i\}$, $i = 0,1,\ldots$ such that

$$\sum_{i=0}^{\infty} [1 - \tau_1(\mathbf{P}(k_i, k_{i+1}))] = \infty . \quad \square \tag{4.3}$$

Strong ergodicity is defined as follows.

Definition 4.3. The time-inhomogeneous Markov chain is strongly ergodic if there exists a vector $\mathbf{q}$, $\|\mathbf{q}\| = 1$ and $\mathbf{q}_i \geqslant 0$, $i \in S$, such that for all m

$$\lim_{n \to \infty} \sup_{\boldsymbol{\nu}(0)} \|\boldsymbol{\nu}(m,n) - \mathbf{q}\| = 0 \tag{4.4}$$

□

Strong ergodicity is obtained only with convergence in addition to loss of memory.

We will need the following result due to Madsen and Isaacson [19,10].

Theorem 4.2. If for every m there exists a $\boldsymbol{\pi}(m)$ such that $\boldsymbol{\pi}(m) = \boldsymbol{\pi}(m)\mathbf{P}(m)$, $\|\boldsymbol{\pi}(m)\| = 1$ and

$$\sum_{m=0}^{\infty} \|\boldsymbol{\pi}(m) - \boldsymbol{\pi}(m+1)\| < \infty ,$$

and the time-inhomogeneous Markov chain is weakly ergodic, then it is also strongly ergodic. Moreover if $\mathbf{e}^* = \lim_{m \to \infty} \boldsymbol{\pi}(m)$, then for all m, $\lim_{n \to \infty} \sup_{\nu(0)} \|\boldsymbol{\nu}(m,n) - \mathbf{e}^*\| = 0$. □

5. Strong Ergodicity of Simulated Annealing.

To establish weak ergodicity we use Theorem 4.1. In particular, we first determine a bound on the coefficient of ergodicity and then we determine the **update** function such that (4.3) is satisfied. Next we show that weak ergodicity together with the existence of $\boldsymbol{\pi}(T_m)$, as defined in (3.1), are sufficient conditions to ensure strong ergodicity.

5.1 Radius of G and Lipschitz Constant

We need a few definitions related to the structure of the graph underlying the Markov chain and to the slope of the cost function.

Let S_m be the set of all the points that are local maxima for the cost function, i.e.,

$$S_m \triangleq \{i \in S \mid c(j) \leqslant c(i) \quad \forall j \in N(i)\}.$$

Let

$$r \triangleq \min_{i \in (S-S_m)} \max_{j \in S} d(i,j) \tag{5.1}$$

be the *radius* of the graph, where $d(i,j)$ is the *distance* of j from i measured by the length (number of edges) of the minimum length path from i to j in G. Let $\hat{i}$, the index of a node where the minimum in (5.1) is attained, be the *center* of the graph.

We will show that at any time the radius r represent an upper bound on the number of transitions of the Markov chain that are required for the probability transition matrix to have all the elements in at least one column, namely the one indexed by $\hat{i}$, to be different from zero. Note that the radius is well defined since we assumed G is connected and, because of the symmetry of $g(i,j)$, it is also strongly connected.

A Lipschitz-like constant bounding the local slope of the cost function is given by

$$L = \max_{i \in S} \max_{j \in N(i)} |c(j) - c(i)|. \tag{5.2}$$

Finally we define a lower bound on the generation function

$$w \triangleq \min_{i \in S} \min_{j \in N(i)} \frac{g(i,j)}{g(i)}. \tag{5.3}$$

An important assumption is that $w > 0$.

5.2 Coefficient of Ergodicity.

If i and j are neighbors in G, i.e. $j \in N(i)$, then from (2.2), (5.2) and (5.3),

$$\mathbf{P}_{ij}(m) \geqslant we^{-L/T_m}, \; m = 0,1,\ldots \tag{5.4}$$

Now the diagonal elements $\mathbf{P}_{ii}(m)$, $i \in (S - S_m)$, may be quite small initially, but these terms are monotonic, increasing with increasing m. This is because the probabilities of transition from node i to neighboring nodes with lower cost are constant with respect to m, while the probabilities of transition to neighboring nodes with higher cost are monotonically decreasing with increasing m. Hence there exists some k_0, $k_0 < \infty$, such that for all $i \in S - S_m$

$$\mathbf{P}_{ii}(m) \geqslant we^{-L/T_m}, \quad m \geqslant (k_0 - 1)r, \tag{5.5}$$

since the left hand side monotonically increases and the right hand side monotonically decreases with increasing m.

We can use (5.1) and (5.5) to bound $\mathbf{P}_{i\hat{i}}(m-r,m)$ for every $i \in S$ and $m \geqslant k_0 r$

$$\mathbf{P}_{i\hat{i}}(m-r,m) \geqslant \prod_{n=m-r}^{m-1} \{we^{-L/T_n}\} \tag{5.6}$$

$$\geqslant w^r e^{-rL/T_{m-1}}.$$

Hence the coefficient of ergodicity τ_1 defined in (4.2) satisfies

$$\tau_1(\mathbf{P}(kr-r,kr)) \leqslant 1 - \min_{i,j} \{\min(\mathbf{P}_{i\hat{i}}, \mathbf{P}_{j\hat{i}})\} \tag{5.7}$$

$$\leqslant 1 - w^r \exp\{-\frac{rL}{T_{kr-1}}\}, \; k \geqslant k_0. \tag{5.8}$$

From now on, for convenience we shall abbreviate $\tau_1(\mathbf{P}(n,m))$ to $\tau_1(n,m)$.

5.3 Weak Ergodicity.

By Theorem 4.1 and (5.8), we have that the Markov chain associated with Simulated Annealing is weakly ergodic if

$$\sum_{k=k_0}^{\infty} \exp\{-\frac{rL}{T_{kr-1}}\} = \infty \tag{5.9}$$

Note that up to now, we have only assumed that the sequence of parameters $\{T_m\}$ is monotonically decreasing and $\lim_{m \to \infty} T_m = 0$; in particular, the dependency of T_m on m has not been specified. We give now an **update** function which insures that the Markov chain is weakly ergodic.

Theorem 5.1. The Markov chain associated with Simulated Annealing with the following **update** function

$$T_m = \frac{\gamma}{\log(m + m_0 + 1)}, \; m = 0, 1, 2, \ldots \tag{5.10}$$

where m_0 is any parameter satisfying $1 \leqslant m_0 < \infty$, is weakly ergodic if

$$\gamma \geqslant rL. \tag{5.11}$$

Proof. Replacing T_m in (5.8) with the formula given in (5.10) we obtain

$$\tau_1(kr-r,kr) \leqslant 1 - \frac{a}{(k + m_0/r)^{\mu}}, \; k \geqslant k_0 \tag{5.12}$$

where $\quad \mu \triangleq rL/\gamma, \;$ and $\; a \triangleq \frac{w^r}{r^{rL\gamma}}.$

It is obvious that, for any l

$$\sum_{k=l}^{\infty} \{1 - \tau_1(kr-r,kr)\} = \infty$$

if $\mu \leqslant 1$. Using Theorem 4.1, the proposition is proved. □

It is clear that weak ergodicity is preserved even if the annealing schedule in (5.10) is modified to keep the temperature unchanged at various, finitely many time steps.

5.4 Strong Ergodicity.

In Section 3 we have shown that there exists for every m, $m \geqslant 0$, a vector $\boldsymbol{\pi}(m)$ of quasi-stationary probabilities that has unit norm, satisfies (3.3) and, as shown in Proposition 3.2, converges to the optimum vector $\mathbf{e}^*$ defined in (3.5).

Hence, to prove the strong ergodicity of the Markov chain associated with Simulated Annealing using Theorem 4.2, we only have to prove the following proposition. Interestingly the proposition holds more generally than for the **update** function in (5.10).

Proposition 5.1. For **update** functions satisfying (3.4) the corresponding quasi-stationary probabilities are such that

$$\sum_{m=0}^{\infty} ||\boldsymbol{\pi}(m+1) - \boldsymbol{\pi}(m)|| \leqslant 2(\tilde{m}+1) < \infty, \tag{5.13}$$

where $\tilde{m}$ is given in (3.7) and (3.8).

Proof. From statement (i) of Proposition 3.3, and (3.7), for $m \geqslant \tilde{m}$,

$$||\boldsymbol{\pi}(m+1) - \boldsymbol{\pi}(m)|| = \sum_{i \in S} \{\boldsymbol{\pi}_i(m+1) - \boldsymbol{\pi}_i(m)\}$$

$$- \sum_{i \notin S^*} \{\boldsymbol{\pi}_i(m+1) - \boldsymbol{\pi}_i(m)\}. \tag{5.14}$$

Hence,

$$||\boldsymbol{\pi}(m+1) - \boldsymbol{\pi}(m)|| = 2\{\boldsymbol{\pi}^*(m+1) - \boldsymbol{\pi}^*(m)\}, \quad m \geqslant \tilde{m} \tag{5.15}$$

where

$$\boldsymbol{\pi}^*(m) \triangleq \sum_{i \in S^*} \boldsymbol{\pi}_i(m), \; m \geqslant 0. \tag{5.16}$$

and the result follows. □

Using Theorem 4.2 and Theorem 5.1, we can prove the fundamental result of this section.

Theorem 5.2. The time-inhomogeneous Markov chain associated with Simulated Annealing is strongly ergodic if it is weakly ergodic, and the annealing schedule satisfies (3.5). In this case, for all m

$$\lim_{n\to\infty} \sup_{\boldsymbol{\nu}(0)} ||\boldsymbol{\nu}(m,n) - \mathbf{e}^*|| = 0. \qquad (5.17)$$

In particular, the annealing schedule in (5.10) with $\gamma \geqslant rL$ gives a strongly ergodic Markov chain for which (5.17) holds. □

6. Finite-Time Behavior and Rate of Convergence.

We obtain an estimate of the departure of the state of the Markov chain at *finite* time m from the optimum vector $\mathbf{e}^*$. The results in Theorem 6.2 below give important insights at the factors affecting the rate of convergence and their implications in the design of optimum annealing schedules.

6.1 Components of Finite-Time Behavior.

The following decomposition is basic:

$$\boldsymbol{\nu}(m) - \mathbf{e}^* = \{\boldsymbol{\nu}(m) - \boldsymbol{\pi}(0)\mathbf{P}(0,m)\} + \{\boldsymbol{\pi}(0)\mathbf{P}(0,m) - \boldsymbol{\pi}(m)\}$$
$$+ \{\boldsymbol{\pi}(m) - \mathbf{e}^*\} \qquad (6.1)$$

Observe that the sum of the first two terms in braces in the right hand side measures the departure at time m of the state distribution from the quasi-stationary distribution. We have chosen to decompose this quantity further so that the first term measures the extent to which at time m the Markov chain has lost memory of the difference between $\boldsymbol{\nu}(0)$ and $\boldsymbol{\pi}(0)$.

From (6.1) we obtain

$$||\boldsymbol{\nu}(m) - \mathbf{e}^*|| \leqslant ||\boldsymbol{\nu}(m) - \boldsymbol{\pi}(0)\mathbf{P}(0,m)|| + ||\boldsymbol{\pi}(0)\mathbf{P}(0,m) - \boldsymbol{\pi}(m)||$$
$$+ ||\boldsymbol{\pi}(m) - \mathbf{e}^*||. \qquad (6.2)$$

In the next subsections, each of the three terms in the right hand side are bounded independently.

6.1.1 Bound for the First Term of (6.2).

To determine a bound for the first term in the right hand side of (6.2), we need the following fundamental result due to Dobrushin [18,9,10].

Theorem 6.1. If $\mathbf{P}$ is any stochastic matrix and $\boldsymbol{\mu}$ is any row vector with $\sum_i \mu_i = 0$, then $||\boldsymbol{\mu}\mathbf{P}|| \leqslant ||\boldsymbol{\mu}||\tau_1(\mathbf{P})$. □

In view of Theorem 6.1 for the first term of the right hand side of (6.2),

$$||\boldsymbol{\nu}(kr) - \boldsymbol{\pi}(0)\mathbf{P}(0,kr)|| = ||\{\boldsymbol{\nu}(0) - \boldsymbol{\pi}(0)\}\mathbf{P}(0,kr)|| \qquad (6.3)$$
$$\leqslant ||\boldsymbol{\nu}(0) - \boldsymbol{\pi}(0)||\tau_1(0,kr).$$

To complete the bound of the first term of (6.2) we need to bound $\tau_1(0,kr)$. To this end the following proposition [20] is necessary.

Proposition 6.1. If $\gamma \geqslant rL$ and the annealing schedule (5.10) is applied so that τ_1 satisfies (5.12), then

$$\tau_1(lr-r,kr) \leqslant \left[\frac{k_0+m_0/r}{k+m_0/r}\right]^a, \quad \text{for } l \leqslant k_0 \leqslant k \qquad (6.4.a)$$

$$\tau_1(lr-r,kr) \leqslant \left[\frac{l+m_0/r}{k+m_0/r}\right]^a, \quad \text{for } k_0 \leqslant l \leqslant k \qquad (6.4.b)$$

where a is defined by (5.12), r by (5.1), and k_0 is such that (5.5) holds and m_0 is the parameter that controls the initial value of the temperature. □

The bound in (6.4) on the coefficient of ergodicity is fundamental to the finite-time analysis of Simulated Annealing. Substituting the above bound in (6.3) yields

$$||\boldsymbol{\nu}(kr) - \boldsymbol{\pi}(0)\mathbf{P}(0,kr)|| \leqslant \frac{||\boldsymbol{\nu}(0) - \boldsymbol{\pi}(0)||(k_0+m_0/r)^a}{(k+m_0/r)^a}, \quad \forall\, k \geqslant k_0. \qquad (6.5)$$

6.1.2 Bound for the Second Term of (6.2).

Let

$$\boldsymbol{\mu}(m) \triangleq \boldsymbol{\pi}(0)\mathbf{P}(0,m) - \boldsymbol{\pi}(m), \quad m = 0,1,... \qquad (6.6)$$

Note that $\boldsymbol{\mu}(0) = \mathbf{0}$ and that $\{\boldsymbol{\mu}(i)\}$ satisfy the recursion

$$\boldsymbol{\mu}(m+r) = \boldsymbol{\mu}(m)\mathbf{P}(m,m+r)$$
$$+ \sum_{s=1}^{r} \{\boldsymbol{\pi}(m+s-1) - \boldsymbol{\pi}(m+s)\}\mathbf{P}(m+s,m+r). \qquad (6.7)$$

The recursion is solved to give

$$\boldsymbol{\mu}(kr) = \sum_{l=1}^{k} \boldsymbol{\epsilon}(lr)\mathbf{P}(lr,kr), \qquad (6.8.a)$$

where,

$$\boldsymbol{\epsilon}(lr) \triangleq \sum_{s=1}^{r} \{\boldsymbol{\pi}(lr-s) - \boldsymbol{\pi}(lr-s+1)\}\mathbf{P}(lr-s+1,lr). \qquad (6.8.b)$$

Applying Theorem 6.1 twice to obtain bounds for $||\boldsymbol{\epsilon}(lr)\mathbf{P}(lr,kr)||$ and $||\boldsymbol{\epsilon}(lr)||$ from (6.8.a) and (6.8.b) respectively, we obtain

$$||\boldsymbol{\mu}(kr)|| \leqslant \sum_{l=1}^{k} \tau_1(lr,kr) \sum_{s=1}^{r} ||\boldsymbol{\pi}(lr+1-s) - \boldsymbol{\pi}(lr-s)||, \quad k \geqslant 1. \qquad (6.9)$$

Now making use of (5.15) and (6.4) we can show [20] that for $k > l_0 \triangleq \max\{\hat{m}/r,\, k_0 - 2\}$,

$$||\boldsymbol{\mu}(kr)|| \leqslant \frac{D_\mu}{(k+m_0/r)^a} + \frac{2a}{(k+m_0/r)^a} \sum_{l=l_0+1}^{k} \frac{\hat{\boldsymbol{\pi}}^*(lr-r)}{(l+m_0/r)^{1-a}}, \qquad (6.10a)$$

where

$$D_\mu \triangleq (k_0+m_0/r)^a \sum_{l=1}^{l_0} \sum_{s=1}^{r} ||\boldsymbol{\pi}(lr+1-s) - \boldsymbol{\pi}(lr-s)||$$
$$+ 2(l_0+1+m_0/r)^a\, \hat{\pi}^*(l_0 r). \qquad (6.10.b)$$

To proceed further it is necessary to estimate $\{\hat{\boldsymbol{\pi}}^*(m)\}$ and this is undertaken in the following proposition [20]. The bound below is asymptotically (i.e. as $m \to \infty$) tight.

Proposition 6.2. For $m = 0, 1, ...$

$$\hat{\boldsymbol{\pi}}^*(m) = 1 - \boldsymbol{\pi}^*(m) = \frac{1}{2}||\boldsymbol{\pi}(m) - \mathbf{e}^*|| \leqslant \sum_{j\notin S^*} \frac{g(j)/g(*)}{(m+m_0+1)^{b(j)}}, \qquad (6.11)$$

where $\{b(j)\}$ is given by

$$b(j) \triangleq \{c(j) - c(*)\}/\gamma, \quad j \in S,$$

$c(*)$, see (3.9), is the minimum of the cost function and $g(*)$, see Proposition 3.2, is $g(*) \triangleq \sum_{j\in S^*} g(j)$.□

We can now say that

$$\hat{\boldsymbol{\pi}}^*(lr-r) \leqslant \sum_{j\notin S^*} \frac{\eta(j)}{(l-1+m_0/r)^{b(j)}}, \quad l = 1,2,... \qquad (6.12.a)$$

where

$$\eta(j) \triangleq \frac{g(j)/g(*)}{r^{b(j)}}, \quad j \in S. \qquad (6.12.b)$$

By substituting (6.12.a) in (6.10.a) and then bounding the resulting expression we obtain

$$||\boldsymbol{\mu}(kr)|| \leqslant \frac{D_\mu}{(k+m_0/r)^a}$$
$$+ \sum_{j\notin S^*} \frac{2a\,\eta(j)}{a-b(j)} \left[\frac{1}{(k+m_0/r)^{b(j)}} - \frac{E^{a-b(j)}}{(k+m_0/r)^a}\right], \qquad (6.13)$$

where $E \triangleq (l_0 - 1 + m_0/r)$. This bound in (6.13) has been obtained for $a \neq b(j)$, $j \notin S^*$; if this is not true, then for the terms corresponding to values of j for which $a = b(j)$, a related expression is obtained by a slightly different bounding procedure.

6.1.3 Bound for the Third Term of (6.2).

This bound comes directly from Proposition 6.2.

6.2 Final Results

Combining the results given in 6.1.1-6.1.3 we obtain the following final theorem.

Theorem 6.2. For every $k \geqslant l_0$, the following relation holds

$$\|\boldsymbol{\nu}(kr) - \mathbf{e}^*\| \leqslant \frac{D}{(k+m_0/r)^a} + \sum_{j \notin S^*} \frac{2a\,\eta(j)}{a-b(j)} \left[\frac{1}{(k+m_0/r)^{b(j)}} - \frac{E^{a-b(j)}}{(k+m_0/r)^a} \right] + \sum_{j \notin S^*} \frac{2\eta(j)}{(k+m_0/r)^{b(j)}}, \qquad (6.14)$$

where

$$D = D_u + \|\nu(0) - \boldsymbol{\pi}(0)\| (k_0 + m_0/r)^a .$$

Also, a, $\{b(j)\}$ and $\{\eta(j)\}$ are given in (5.12), (6.11) and (6.12.b) respectively. □

Equation (6.14) can be further simplified if we observe that the dominant term of $\frac{1}{(k+m_0/r)^{b(j)}}$, $j \notin S^*$, is given by $\frac{1}{(k+m_0/r)^b}$, where $b \triangleq \min_{j \notin S^*} b(j) = \frac{\delta}{\gamma}$, and δ, which has been defined in (3.10), is the difference between next-to-least cost and least-cost.

A simple corollary to Theorem 6.2 is

Proposition 6.3. The Simulated Annealing algorithm with the annealing schedule given by (5.10) has the following estimate for its rate of convergence

$$\|\boldsymbol{\nu}(kr) - \mathbf{e}^*\| = O(1/k^{\min(a,b)}). \; \square \qquad (6.15)$$

6.3 Discussion.

We can see from (6.15) that the bound on the asymptotic rate of convergence is limited by $\min(a,b)$. Both a and b depend on δ and L derived from the cost function, w and r from the connectivity properties of the graph underlying the Markov chain and on γ from the annealing schedule. Note that with all other parameters and time held fixed, higher γ corresponds to higher temperature and thus, in this sense, to slower cooling. Now γ has to satisfy a condition that gives weak ergodicity, i.e. $\gamma \geqslant \gamma_{WE}$ wherein by our analysis $\gamma_{WE} = rL$, but otherwise it is a free parameter. It is therefore of some interest to investigate the value of γ which maximizes $\min(a,b)$.

Recall the definition of a in (5.12) and that $b \triangleq \delta/\gamma$. Hence $a(\gamma)$ and $b(\gamma)$ are respectively increasing and decreasing with increasing γ, and it is easy to see that there exists an unique $\tilde{\gamma}$ such that $a(\tilde{\gamma}) = b(\tilde{\gamma})$. Furthermore, the problem

$$\max_{\gamma : \gamma \geqslant \gamma_{WE}} \{\min(a,b)\}$$

has the solution

$$\gamma = \max(\gamma_{WE}, \tilde{\gamma}) .$$

The above procedure for optimizing the algorithm is often feasible since for many combinatorial optimization problems, graph partitioning problems in particular, estimates of r, L and δ are available.

The above discussion has been on the effect of γ (from the annealing schedule) on the bound on the rate of convergence at finite, but large time. For behavior at smaller time, the more detailed relation (6.14) has to be considered. Observe that in the right hand side of this equation, the only factors which depend on the time kr are $1/(k+m_0/r)^a$ and $1/(k+m_0/r)^{b(j)}$, $j \notin S^*$. We may glean qualitative information on the dependence of the rate of convergence on γ by investigating the dependence of a and $\{b(j)\}$ on γ. Now, smaller γ gives larger $b(j)$ for each j and, as already noted, smaller a. Hence, reducing γ has the effect of reducing the third term and increasing the first term in the right hand side of (6.14). The dependence of the middle term is more involved since it has features of both other terms reflected in it. Roughly, it is small only when both the first and third terms are small, i.e. in the mid-range of γ.

With the benefit of analysis we can even go back to (6.2) and deduce qualitatively the effect of γ on each of the three terms there. The first term measures how effectively the difference between $\boldsymbol{\nu}(0)$ and $\boldsymbol{\pi}(0)$ is forgotten at step m of the algorithm. The bound in (6.5) corroborates our intuitive understanding that this rate of memory loss is aided by having higher γ, i.e. higher temperatures and slower cooling. The third term, for which we have the most explicit information (see Proposition 6.2) depends on the rate at which the quasi-stationary distribution approaches its asymptotic value, the optimum distribution. This term benefits from small γ. The middle term benefits from a matching of the two rates. The point in the analysis where this is most explicitly manifest is in (6.12.a). The two rates are matched and the term minimized in the mid-range of γ. In all, the above discussion illuminates the balancing of opposite mechanisms that an optimal annealing schedule must reflect.

7. Concluding Remarks

We have proven a number of results on the behavior of Simulated Annealing. In particular, we have introduced an annealing schedule which guarantees that the individual state probabilities converge either to a positive value or to zero depending upon whether the configuration corresponding to the state is globally least-cost or not. Also we have analyzed finite-time behavior in terms of a decomposition of the distance of the state probability vector from the optimum. Each of the three terms of the decomposition reflects an important component of the behavior of the algorithm. Each term has an independent bound and this allows the trade-offs in the design of the algorithm to be quantified.

We give below a selection of three directions in which the present analysis may be extended:

1. An analysis more closely attached to the evolution with time of mean cost rather than the distance of the state distribution from the optimal.
2. An analysis of schedules in which temperature is lowered at a faster rate than that allowed here by (1.1).
3. The exploitation of special properties of the cost function to design matched annealing schedules with a provable improvement in performance.

References

1. M. R. Garey, D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, 1979.
2. J. T. Schwartz, Fast Probabilistic Algorithms for Verification of Polynomial Identities, *Journal of ACM*, Vol 27, No. 4, Oct 1980.
3. S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by Simulated Annealing, *Science*, Vol. 220, N. 4598, pp. 671-680, 13 May 1983.
4. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, Equations of State Calculations by Fast Computing Machines *J. Chem. Phys.*, Vol. 21, pp. 1087-1091, 1953.
5. K. Binder, *Monte Carlo Methods in Statistical Physic*, Springer-Verlag, 1978.

6. M. P. Vecchi, S. Kirkpatrick, Global Wiring by Simulated Annealing, *IEEE Transactions on Computer-Aided Design*, Vol. CAD-2, No. 4, Oct 1983.

7. C. Sechen, A. Sangiovanni-Vincentelli, The Timber Wolf Placement and Routing Package, *Proc 1984 Custom Integrated Circuit Conference,* Rochester, May 1984.

8. D. S. Johnson, Simulated Annealing Performance Studies, presented at the Simulated Annealing Workshop, Yorktown Heights, April 1984.

9. E. Seneta, *Non-negative Matrices and Markov Chains*, Second Edition, Springer-Verlag, New York, 1980.

10. D. L. Isaacson, R. W. Madsen, *Markov Chains: Theory and Applications* John Wiley, New York, 1976.

11. M. Iosifescu *Finite Markov Processes and their Applications*, John Wiley, New York, 1980.

12. M. Lundy, A. Mees, Convergence of the Annealing Algorithm, presented at Simulated Annealing Workshop, Yorktown Heights, April 1984.

13. F. Romeo, A. Sangiovanni-Vincentelli, Probabilistic Hill Climbing Algorithms: Properties and Applications, ERL Memo, University of California, Berkeley, 1984.

14. S. Geman, D. Geman, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 6, 1984, pp. 721-741.

15. B. Gidas, Non-Stationary Markov Chains and Convergence of the Annealing Algorithm, J. of Stat. Physics, Vol. 39, pp. 73-131, 1985.

16. B. Hajek, Cooling Schedules for Optimal Annealing, preprint, 1985.

17. F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, New York, 1980.

18. R. L. Dobrushin, Central Limit Theorem for Nonstationary Markov Chains, I, II, *Theory Prob. Appl.*, 1, pp. 65-80, 329-83 (English translation), 1956.

19. R. W. Madsen, D. L. Isaacson, Strongly Ergodic Behavior for Non-stationary Markov Processes, *Ann. Prob.,* Vol. 1, No. 2, pp. 329-335, 1973.

20. D. Mitra, F. Romeo, A. Sangiovanni-Vincentelli, Convergence and Finite-Time Behavior of Simulated Annealing, University of California, Electronics Research Laboratory Memo. UCB/ERL M85/23, March 1985; to appear in Adv. App. Prob., Sept. 1986.