November 1986 LIDS-P-1623

A SURVEY OF LARGE TIME ASYMPTOTICS OF SIMULATED ANNEALING ALGORITHMS†

John N. Tsitsiklis
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Simulated annealing is a probabilistic algorithm for minimizing a general cost function which may have multiple local minima. The amount of randomness in this algorithm is controlled by the "temperature", a scalar parameter which is decreased to zero as the algorithm progresses. We consider the case where the minimization is carried out over a finite domain and we present a survey of several results and analytical tools for studying the asymptotic behavior of the simulated annealing algorithm, as time goes to infinity and temperature approaches zero.

I. Introduction.

Simulated annealing is a probabilistic algorithm for minimizing a general cost function which may have multiple local minima. It has been introduced in [1] and [2] and was motivated by the Metropolis algorithm [3] in statistical mechanics. Since then, it has been applied to a variety of problems, the main ones arising in the context of combinatorial optimization [1,4,5,6] and in the context of image restoration [7].

Let $S = \{1,...,N\}$ be a finite state space, let Z be the set of nonnegative integers and let $J : S \mapsto Z$ be a cost function to be minimized. We assume that, for each $i \in S$, we are also given a set $S(i) \subset S$, to be called the set of *neighbors* of i. Let us assume that

$$j \in S(i) \quad \text{if and only if} \quad i \in S(j). \tag{1.1}$$

The neighborhood structure of S may be also described by a graph $G = (S, E)$, where E, the set of edges is defined by $E = \{(i,j) : j \in S(i)\}$. Given a neighborhood structure, a natural method for trying to optimize J is the "descent" method: given the present state $i \in S$, one examines the neighbors of i and lets the state become some $j \in S(i)$ such that $J(j) \leq J(i)$. The descent method, in general, cannot find a global optimum; it is possible that $J(i) < J(j)$, $\forall j \in S(i)$, without i being a global minimizer of J. *Multistart* algorithms provide a popular modification of the descent method in which the above described procedure is repeated, starting from random and independently chosen initial states. This guarantees that eventually a global minimum will be reached, but depending on the structure of the problem, this may take too long.

† Research supported by the Army Research Office under contract DAAAG-29-84-K-0005.

Simulated annealing may be viewed as another probabilistic modification of the descent method in which randomness is introduced in a somewhat different way. The algorithm proceeds as follows. For each $(i,j) \in E$ we are given a positive scalar Q_{ij}. We assume that $\sum_{j \in S(i)} Q_{ij} = 1$. For notational convenience we also let $Q_{ij} = 0$, if $(i,j) \notin E$. Let i be the state of the algorithm at time t. Then, one chooses randomly a neighbor of i, with Q_{ij} being the probability that $j \in S(i)$ is selected. If $J(j) \leq J(i)$, the state moves to j. If $J(j) > J(i)$, then the state moves to j, with probability $\exp\{-(J(j) - J(i))/T(t)\}$, or stays at i, with probability $1 - \exp\{-(J(j) - J(i))/T(t)\}$. Here $T(t)$ is a time–varying parameter, called the temperature for historical reasons (see Section II), which controls the amount of randomness in the algorithm. It is clear that simulated annealing is similar to hill–climbing, except that transitions which increase the cost are sometimes allowed to occur.

Notice that when temperature is small, the probability of upward transitions is small and therefore the algorithm will take a very long time to escape from a local (but non–global) minimum. On the other hand if temperature is large, the tendency of the algorithm to move downwards is reduced and the algorithm may oscillate for a long time in the neighborhood of a global minimum before it reaches it. Thus, the choice of $T(\cdot)$ (which will be called a *temperature schedule* or simply *schedule*) becomes a very delicate issue. The common prescription is to start with a fairly large temperature and then gradually decrease it towards zero. In subsequent sections we explore the dependence of the asymptotic behavior of the simulated annealing algorithm on the rate at which temperature converges to zero. We should mention here that, as far as applications are concerned, it is also important to study the asymptotic behavior of simulated annealing as a function of the size of the problem being solved. Very few results [8] of this type are available at present.

Outline of the paper: In Section II we mention briefly the motivation of simulated annealing from statistical mechanics. In Section III we examine the behavior of the algorithm for the case where the temperature is kept at a small constant value. Then, in Section IV, we examine the case where $T(\cdot)$ is piecewise constant and motivate the logarithmic schedules of the form

$$T(t) = \frac{\delta}{\log t}, \tag{1.2}$$

where δ is a positive scalar. In Section V we present some general tools for analyzing Markov chains with rare transitions. These tools are used in Section VI to obtain a complete characterization of the asymptotics of simulated annealing for a schedule which is a piecewise constant approximation of the schedule (1.2). In Section VII we derive the smallest value of the constant δ [see equation (1.2)] for which convergence to the set of global minima is obtained.

The results in Sections III and IV are adaptations of some results in [7, 9, 10, 11, 12]. Some of the results of Sections V and VI have been proved in [14], in somewhat more general form. The main result of Section VII (Corollary 7.7) is due to Hajek [13]. The derivation here is new and is based on the results of Sections V and VI.

II. The Origins of Simulated Annealing in Statistical Mechanics.

Consider a particle which may be found in any state in a state space S, let J be a nonnegative integer valued function determining the energy of the particle, as a function of its state, and let S^* be the set of states at which J is minimized. One of the principles of statistical mechanics is that such a particle, in thermal equilibrium with a heat bath at temperature T, will reside at state i with probability

$$\pi_T(i) = \frac{1}{Z_T} \exp\left[-\frac{J(i)}{T}\right], \tag{2.1}$$

where Z_T is a normalizing constant. It is then straightforward to verify that, the limit, as $T \downarrow 0$, of the probability that the particle lies in the set S^* of global minima of J, is equal to 1. This suggests the following method for identifying elements of S^*: simulate the behavior of such a particle, with a small temperature, for a sufficiently long time to allow the particle to reach steady state; then, the particle will, with very high probability occupy a state in S^*. This is straightforward, provided that we are able to construct a Markov chain whose invariant probability distribution is given by equation (2.1). There are several alternatives for accomplishing this; the first one was suggested in [3]. For our purposes we will only consider the alternative that follows.

Let $x_T(\cdot)$ be a stationary, discrete time, Markov chain with state space $S = \{1, ..., N\}$, and let its one step transition probabilities be

$$q_T(i,j) = P(x_T(t+1) = j \,|\, x_T(t) = i) = Q_{ij} \exp\left[-\frac{1}{T}\max\{0, J(j) - J(i)\}\right], \quad i \neq j, \tag{2.2a}$$

$$q_T(i,i) = 1 - \sum_{j \neq i} q_T(i,j). \tag{2.2b}$$

Here, the scalars Q_{ij} have the properties postulated in Section I. We further assume that this Markov chain is irreducible and aperiodic and that

$$Q_{ij} = Q_{ji}, \qquad \text{(strong reversibility)}. \tag{2.3}$$

Under these assumptions we have the following.

Lemma 2.1: The (unique) vector of invariant probabilities of the Markov chain $x_T(\cdot)$ is given by (2.1).

Proof: Some straightforward algebra shows that $\sum_{i \in S} \pi_T(i) q_T(i,j) = \pi_T(j)$, and the result follows. •

We notice that the simulated annealing algorithm is no different from the above Markov chain, were it not for temperature variations. This provides us with some additional understanding of the role of the temperature parameter. Equation (2.1) shows that a small temperature is desirable, since it leads to a large probability of being at S^*. On the other hand it turns out that if T is small, then the above described Markov chain requires a long time to reach steady state (see Section III). This is just a different aspect of the tradeoff mentioned in Section I.

If strong reversibilty fails to hold but instead we assume the condition

$$Q_{ij} > 0 \quad \text{if and only if} \quad Q_{ji} > 0, \qquad \text{(structural reversibility)}, \tag{2.4}$$

then the steady state probabilities are not given by (2.1) anymore. Nevertheless, it remains true that

$$\lim_{T \downarrow 0} \lim_{t \to \infty} P(x_T(t) \in S^*) = 1. \tag{2.5}$$

Equation (2.5) shows that in order to obtain a state in S^* (with high probability) we may simulate the chain $x_T(\cdot)$, at a fixed temperature, until steady state is reached, then simulate it again with a smaller temperature and so on. However, we would like to obtain a state in S^* with a single simulation. A solution which suggests itself is that we decrease the temperature in the course of a single simulation. This argument provides some motivation for schedules with decreasing temperature, such as the one given by equation (1.2).

III. Time Evolution in the Constant Temperature Case.

The time evolution of simulated annealing is a lot simpler to understand when temperature is kept constant. We thus assume that T has been fixed to a small positive value and we consider the Markov chain $x_T(\cdot)$ with transition probabilities defined by (2.2). We first change our notation slightly by introducing a new parameter ϵ defined by $\epsilon = \exp\{-1/T\}$. (Accordingly, with a slight abuse of notation, we will write $x_\epsilon(t)$ instead of $x_T(t)$.) Notice that $0 < \epsilon < 1$ and that the one step–transition probabilities of $x_\epsilon(t)$, expressed as functions of ϵ, are given by the formula

$$q_\epsilon(i,j) = R_{ij}\epsilon^{\alpha(i,j)}, \tag{3.1}$$

where each R_{ij} is positive and where each $\alpha(i,j)$ is either a nonnegative integer or equal to infinity. (We use the convention $\epsilon^\infty = 0$.) In particular, for $j \neq i$ we have $\alpha(i,j) = \infty$, if $Q_{ij} = 0$, and $\alpha(i,j) = \max\{0, J(j) - J(i)\}$, otherwise. We are thus dealing with a Markov chain whose one step transition probabilities are powers of a small parameter. This is the situation encountered in perturbation theory of Markov chains [15,16,17,18]. In particular, the following are known. The eigenvalue of largest magnitude is equal to 1 and it is an isolated eigenvalue if $x_\epsilon(\cdot)$ is irreducible. Assuming that $x_\epsilon(\cdot)$ is also aperiodic, the eigenvalue λ_ϵ with largest modulus among the remaining eigenvalues satisfies

$$\lambda_\epsilon = 1 - A\epsilon^\Delta + o(\epsilon^\Delta), \tag{3.2}$$

where A is some constant, Δ is a nonnegative integer and $o(\epsilon^\Delta)$ represents terms which are negligible compared to ϵ^Δ, as $\epsilon \downarrow 0$ [15].

Equation (3.2) provides us with some information on the long–run behavior of $x_\epsilon(\cdot)$ because the time constant which governs the rate at which $x_\epsilon(\cdot)$ reaches steady state is equal to $1/\lambda_\epsilon$. However, the constant Δ of equation (3.2) is, in general, hard to compute and for this reason the earlier work on simulated annealing has been based on bounds for Δ.

A simple estimate for λ_ϵ, and a corresponding bound for Δ, may obtained as follows. Let τ be the smallest integer such that the number B_ϵ defined by

$$B_\epsilon = \max_j \min_i P(x_\epsilon(\tau) = i \,|\, x_\epsilon(0) = j) \tag{3.3}$$

is positive. (Such a τ is guaranteed to exist if $x_\epsilon(\cdot)$ is irreducible and aperiodic and is independent of ϵ because it only depends on the set of transitions which have positive probability and not on the exact numerical values of the transition probabilities.) We let

$$R = \min_{i,j \in S} R_{ij}, \tag{3.4}$$

$$\alpha^* = \max_i \max_{j \in S(i)} \alpha(i,j), \tag{3.5}$$

where $\alpha(i,j)$, R_{ij} have been defined in equation (3.1). Every transition probability $q_\epsilon(i,j)$ which is nonzero is bounded below by $R\epsilon^{\alpha^*}$. It follows that the constant B_ϵ defined by (3.3) satisfies $B_\epsilon \geq (R\epsilon^{\alpha^*})^\tau$.

We now use the well-known estimate $(\lambda_\epsilon)^\tau \leq 1 - B_\epsilon$. Thus,

$$\lambda_\epsilon \leq (1 - R^\tau \epsilon^{\tau\alpha^*})^{1/\tau} \leq 1 - \frac{1}{\tau} R^\tau \epsilon^{\tau\alpha^*}. \tag{3.6}$$

By comparing (3.6) with (3.2), we conclude that $\Delta \leq \tau\alpha^*$. We have thus proved the following.
Proposition 3.1: Assume that $x_\epsilon(\cdot)$ is irreducible and aperiodic, for every $\epsilon > 0$. Let Δ, τ, α^* be defined by (3.2), (3.4), (3.5), respectively. Then, $\Delta \leq \tau\alpha^*$.

IV. The Evolution in the Case of Piecewise Constant Temperatures.

Let $\{t_k\}$ be an increasing sequence, with $t_1 = 1$. Let us consider a schedule of the form $T(t) = \frac{1}{k}$, for $t_k \leq t < t_{k+1}$. Equivalently,

$$\epsilon(t) = e^{-k}, \qquad t_k \leq t < t_{k+1}. \tag{4.1}$$

Following the prescription of Section II, we shall take the difference $t_{k+1} - t_k$ to be large enough so that the stationary Markov chain (at temperature $1/k$) comes arbitrarily close to steady state. For this to occur, it is sufficient to let $t_{k+1} - t_k$ correspond to an arbitrarily large number of time constants of the Markov chain $x_T(\cdot)$, at temperature $T = 1/k$ (equivalently, with $\epsilon = e^{-k}$). Using equation (3.2), the relevant time constant is of the order of $\epsilon^{-\Delta}$, that is, of the order of $e^{k\Delta}$. We thus let

$$t_{k+1} - t_k = \exp\{k\delta\}, \tag{4.2}$$

where δ is a scalar strictly larger than Δ. Equation (4.2) implies that $t_k \approx \exp\{k\delta\}$ (within a bounded multiplicative constant). Thus, $T(t_k) = 1/k \approx \frac{\delta}{\log t_k}$. We have been thus led to a logarithmic schedule, similar to the one in equation (1.2). The schedule here is a piecewise

constant approximation of the scedule (1.2). A piecewise constant schedule is much easier to analyze because the machinery of perturbed stationary Markov chains, such as the result quoted in Section III [equation (3.2)], becomes applicable. For this reason we focus in this paper on the case of such piecewise constant schedules and refer the reader to the literature for the extension to more general cases. The following result summarizes the above discussion.

Proposition 4.1: Assume that the stationary Markov chain $x_\epsilon(\cdot)$, whose transition probabilities are determined by (2.2), is irreducible and aperiodic for every $\epsilon \in (0,1)$ and that equation (2.3) holds. With the schedule determined by equations (4.1), (4.2), and with $\delta > \Delta$, we have

$$\lim_{t\to\infty} P(x(t) \in S^* | x(0) = i) = 1, \qquad \forall i \in S, \tag{4.3}$$

where $x(\cdot)$ is the resulting non–stationary Markov chain.

The outline of the proof of Proposition 4.1 that we have provided is conceptually similar to the ones in [7] and [10], except that these references do not mention the second eigenvalue but work with conservative estimates of the second eigenvalue, similar to the estimate provided by equation (3.6). Reference [11] works with the eigenvalue exponent Δ and carries out the proof for the more general case of nonincreasing temperature schedules which are not piecewise constant and obtains an extension of Proposition 4.1. Let us also point out here that Proposition 4.1 remains valid even if the condition $\delta > \Delta$ is weakened to $\delta \geq \Delta$.

Proposition 4.1 does not furnish the best possible conditions for the validity of (4.3). To illustrate this, consider a cost function J in which every local minimum is also a global minimum. For such a cost function, pure descent (that is, zero temeperature) will satisfy (4.3) and this is also the case for any temperature schedule such that $\lim_{t\to\infty} T(t) = 0$. On the other hand if an increase in costs is necessary in order to go from one local minimum to another, then the time needed to reach equilibrium goes to infinity, as $T \downarrow 0$, and therefore Δ is nonzero; thus Proposition 2.1 places an unnecessary restriction on the admissible schedules. In the next few sections we obtain tighter conditions (necessary and sufficient) on δ, for equation (4.3) to hold.

V. Order of Magnitude Estimates for Markov Chains with Rare Transitions.

In this section we present some results concerning a family of stationary Markov chains $\{x_\epsilon(\cdot)\}$, parametrized by a small positive parameter $\epsilon \geq 0$. These results are applicable to the simulated annealing algorithm for time intervals during which the temperature is constant. We will use them, in particular, to analyze the case of the piecewise constant schedule determined by (4.1), (4.2).

Let $\bar{A}$ be the set of all functions from $S \times S$ into the set $\{0,1,2,...\} \cup \{\infty\}$. Thus any $\alpha \in \bar{A}$ corresponds to a collection $\alpha = \{\alpha(i,j) : i,j \in S\}$ of coefficients. Let C_1, C_2 be constants. We assume that, for any $\epsilon \in [0,1)$, we have a stationary Markov chain $x_\epsilon(\cdot)$ whose one–step transition probabilities satisfy the inequalities

$$C_1 \epsilon^{\alpha(i,j)} \leq P(x_\epsilon(t+1) = j \,|\, x_\epsilon(t) = j) \leq C_2 \epsilon^{\alpha(i,j)}, \quad \forall \epsilon \in [0,1). \tag{5.1}$$

(We use the conventions that $\epsilon^\infty = 0$ and $0^0 = 1$.) Notice that equation (3.1) is a special case of (5.1). Notice that in the context of Markov chains α is not completely arbitrary. In particular, the probabilities out of a fixed state have to add to 1. This implies that for every i there exists some j such that $\alpha(i,j) = 0$. We let $\mathcal{A}$ be the set of all $\alpha \in \bar{\mathcal{A}}$ that have this property.

The result that follows provides us with similar inequalities describing the long run behavior of $x_\epsilon(\cdot)$.

Proposition 5.1: We assume that for any $\epsilon \in [0,1)$ the stationary Markov chain $x_\epsilon(\cdot)$ satisfies (5.1), is irreducible and each one of the irreducible components in an ergodic decomposition of $x_0(\cdot)$ is aperiodic. Then, for any positive integer d, there exists a collection of coefficients $V^d \in \mathcal{A}$ and two positive constants C_1^d, C_2^d such that

$$C_1^d \epsilon^{V^d(i,j)} \leq P(x_\epsilon(t+\epsilon^{-d}) = i \,|\, x_\epsilon(t) = j) \leq C_2^d \epsilon^{V^d(i,j)}, \quad \forall \epsilon \in (0,1)^\dagger. \tag{5.2}$$

A proof of this result may be found in [14]. A related result, in somewhat different form, is implicit in [17]. The above result is should not be surprising and its only nontrivial feature is that the same constants C_1^d, C_2^d work for all choices of ϵ.

What is more interesting for our purposes is the actual computation of the coefficients $V^d(i,j)$ in terms of the original coefficients $\alpha(i,j)$. For this, we need to develop some auxiliary notation and terminology. We define a map $H : \mathcal{A} \mapsto \mathcal{A}$, as follows. For every $\alpha \in \mathcal{A}$, we let $(H\alpha)(i,j)$ be the shortest distance from i to j, where the length of link (k,ℓ) is taken to be equal to $\alpha(k,\ell)$. A useful property of H is that it is monotone: if $\alpha \leq \beta$, then $H\alpha \leq H\beta$. Furthermore, for every $\alpha \in \mathcal{A}$ and for any $i,j,k \in S$ we have the triangle inequality $(H\alpha)(i,k) \leq (H\alpha)(i,j) + (H\alpha)(j,k)$.

For any $\alpha \in \mathcal{A}$ we let $R(\alpha)$ be the subset of S defined by $R(\alpha) = \{i : (H\alpha)(i,j) = 0$ implies $(H\alpha)(j,i) = 0\}$. For any $\alpha \in \mathcal{A}$, $i \in R(\alpha)$, we let $R_i(\alpha) = \{j : (H\alpha)(i,j) = 0\}$. It is easy to see that if $i \in R(\alpha)$ and $j \in R_i(\alpha)$, then $j \in R(\alpha)$ and $i \in R_j(\alpha)$. It is useful to think of $R(\alpha)$ as the set of recurrent states in the fastest time scale. That is, for times of the order of 1, we neglect any transitions which have probability of the order of ϵ or smaller and we identify $R(\alpha)$ as the set of recurrent states for the resulting state transition diagram.

We define a mapping $F : \mathcal{A} \mapsto \mathcal{A}$ by

$$(F\alpha)(i,j) = \alpha(i,j) - 1, \quad \text{if } i \in R(\alpha),\ j \notin R_i(\alpha), \tag{5.3a}$$

$$(F\alpha)(i,j) = \alpha(i,j), \quad \text{otherwise.} \tag{5.3b}$$

Finally, we define a mapping $\hat{H} : \mathcal{A} \mapsto \mathcal{A}$ by

$$(\hat{H}\alpha)(i,j) = \min_{k,\ell \in R(\alpha)} [(H\alpha)(i,k) + (HF\alpha)(k,\ell) + (H\alpha)(\ell,j)]. \tag{5.4}$$

---

† Of course, ϵ^{-d} may be non–integer, but we may define $x_\epsilon(\cdot)$ to be a right–continuous step function, with jumps only at integer times; thus, $x_\epsilon(t+\epsilon^{-d})$ is well–defined.

In particular, if $i,j \in R(\alpha)$, then $\hat{H}(i,j)$ is the length of a shortest path from i to j, where the length of a path is taken to be equal to the sum of the $\alpha(k,\ell)$'s along this path, minus the number of classes $R_k(\alpha)$ which are exited in the course of this path.

The following result, shows that the coefficients V^d may be recursively computed by iteratively applying the map $\hat{H}$.

Proposition 5.2: Let $V^0 = \alpha$. Under the assumptions of Proposition 5.1, we have

$$V^{d+1}(i,j) = (\hat{H}V^d)(i,j), \qquad \forall d \geq 0, \ \forall i,j. \tag{5.5}$$

The proof of Proposition 5.2 is somewhat tedious and is therefore omitted. It may be found in [14]. It is suggested that the reader applies the iteration (5.5) on a simple example, and its content will become quite transparent.

Let us provide here an intuitive justification for the formula for $V^1(i,j)$, for the case where $i,j \in R(\alpha)$. Notice that for any path from i to j, the parameter ϵ raised to the power equal to the sum of the $\alpha(k,\ell)$'s along the path is an approximation of the probability that this path is followed, during a time interval of $O(1)$ duration. Let us now look at a time interval of the order of $1/\epsilon$. Suppose that $k \in R(\alpha)$ and that at some time the state is equal to k. Then, the state stays inside $R_k(\alpha)$ for at least $O(1/\epsilon)$ time and keeps visiting state k. Thus, a transition from a state k to a state $\ell \notin R_k(\alpha)$ has $O(1/\epsilon)$ opportunities to occur. Therefore, the probability of this transition is of the order of $\epsilon^{\alpha(k,\ell)-1}$. This should provide some insight as to why the mapping F appears in formula (5.4).

The following result will be useful later.

Proposition 5.3: Let $\alpha \in \mathcal{A}$.
(i) If $i \in R(\alpha)$ and $j \notin R_i(\alpha)$, then $\alpha(i,j) \geq 1$.
(ii) For every i, there exists some $j \in R(\alpha)$ such that $(H\alpha)(i,j) = 0$.
Proof: (i) If $j \notin R_i(\alpha)$, then $\alpha(i,j) \geq (H\alpha)(i,j) \geq 1$.
(ii) Suppose that $i \in R(\alpha)$. Let j be such that $\alpha(i,j) = 0$. Then $j \in R(\alpha)$ and $(H\alpha)(i,j) \leq \alpha(i,j) = 0$.

We now consider the case $i \notin R(\alpha)$. Consider the following algorithm. Given a current state $i_k \notin R(\alpha)$, go to a state i_{k+1} such that $\alpha(i_k, i_{k+1}) = 0$ and $(H\alpha)(i_{k+1}, i_k) \neq 0$. [Such a i_{k+1} must exist, because otherwise we would have $i_k \in R(\alpha)$.] This algorithm cannot visit twice the same state i_k, because in that case we would have $(H\alpha)(i_{k+1}, i_k) = 0$, which is a contradiction. Thus, the algorithm must eventually enter $R(\alpha)$ and it follows that there exists a zero length path (with respect to α) from $i \notin R(\alpha)$ to $R(\alpha)$. •

The following result collects a few useful properties of the coefficients V^d.

Proposition 5.4: (i) For any $d > 0$, and for any i,j,k we have $V^d(i,j) \leq V^d(i,k) + V^d(k,j)$.
(ii) If $d > 0$, $i \in R(V^d)$ and $j \in R_i(V^d)$, then $V^d(i,j) = 0$.

(iii) $R(V^{d+1}) \subset R(V^d)$.

Proof: (i) Because the map $\hat{H}$ is defined in terms of shortest path problems [see equation (5.4)], it easy to verify that $(\hat{H}\alpha)(i,j) \leq (\hat{H}\alpha)(i,k) + (\hat{H}\alpha)(k,j)$, $\forall i,j,k$, $\forall \alpha \in \mathcal{A}$, and the result follows.
(ii) If $i \in R(V^d)$, $j \in R_i(V^d)$, then $(HV^d)(i,j) = 0$. The triangle inequality (part (i)) translates to $HV^d = V^d$ and the result follows.
(iii) Suppose that $i \notin R(V^d)$. Then, $V^d(j,i) > 0$, $\forall j \in R(V^d)$, and using (5.4), (5.5), we have $V^{d+1}(j,i) > 0, \forall j \in R(V^d)$. On the other hand, there exists some $j \in R(V^d)$ such that $V^d(i,j) = 0$ (Proposition 5.3(ii)) and using (5.4), (5.5), once more, we obtain $V^{d+1}(i,j) = 0$. It follows that $i \notin R(V^{d+1})$. •

The definition of $R(V^d)$ and inequality (5.2) show that $R(V^d)$ is the set of states on which probability is concentrated for times of the order of $1/\epsilon^d$. For this reason, we may be interested in the coefficients $V^d(i,j)$ only for $i,j \in R(V^d)$. If this is the case, then the following result provides a somewhat simpler procedure for computing these coefficients.

Proposition 5.5: Given $\alpha \in \mathcal{A}$, let $U^0 = \alpha$ and

$$U^{d+1} = HFU^d. \tag{5.6}$$

Then, the following are true:
(i) $U^d(i,j) = V^d(i,j)$, $\forall i,j \in R(V^{d-1})$, $\forall d \geq 1$.
(ii) For every $d \geq 0$, $i \in S$, there exists some $j \in R(V^d)$ such that $U^d(i,j) = 0$, if $d > 0$, or $(HU^d)(i,j) = 0$, if $d = 0$;
(iii) $R(V^d) = \cap_{c=0}^{d} R(U^c)$, $\forall d \geq 0$.
Proof: From (5.6), it is easy to see that U^d satisfies the triangle inequality $U^d(i,j) \leq U^d(i,k) + U^d(k,j)$, $\forall i,j,k$, $\forall d > 0$, a fact that we will use freely. (Recall that V^d also satisfies the triangle inequality, by Proposition 5.4(i).)

The proof is by induction on d. For $d = 0$, we have $U^0 = V^0$ and part (iii) is trivially true. Part (ii) follows from Proposition 5.3(ii) and the fact $U^0 = V^0$.

We now assume that the result is true for some $d \geq 0$ and we shall prove it for $d+1$.
(i) Let $i,j \in R(V^0)$. Then, by definition (5.4), we have $V^1(i,j) = (HFV^0)(i,j) = (HFU^0)(i,j) = U^1(i,j)$, and this proves part (i) for $d = 1$.

Suppose now that the result holds for some $d > 0$. Then, V^d and U^d satisfy the triangle inequality. Let $i,j \in R(V^d)$. Using the definition (5.4) of $V^d(i,j)$ and the triangle inequality, it is easy to see that $V^{d+1}(i,j) = \sum_{k=1}^{m}[V^d(i_k,i_{k+1}) - 1]$, for some path $i_1, i_2, ..., i_m$ such that $i_1 = i$, $i_m = j$, $i_k \in R(V^d)$ and $i_{k+1} \notin R_{i_k}(V^d)$. Using part (i) of the induction hypothesis we have $V^d(i_k,i_{k+1}) - 1 = U^d(i_k,i_{k+1}) - 1$. Using part (iii) of the induction hypothesis, $R(V^d) \subset R(U^d)$ and therefore $i_k \in R(U^d)$. Furthermore, $i_{k+1} \notin R_{i_k}(V^d)$, which implies $U^d(i_k,i_{k+1}) = V^d(i_k,i_{k+1}) > 0$, which shows that $i_{k+1} \notin R_{i_k}(U^d)$. Thus $U^d(i_k,i_{k+1}) - 1 = (FU^d)(i_k,i_{k+1})$. Thus, $V^{d+1}(i,j) \geq (HFU^d)(i,j) = U^{d+1}(i,j)$.

For the reverse inequality, let $i_1, ..., i_m$ be a path from i to j which is of minimal length, with respect to FU^d. Using the triangle inequality again, we may assume, without loss of generality that $i_k \in R(U^d)$ and $i_{k+1} \notin R_{i_k}(U^d)$. Then, $U^{d+1} = \sum_{k=1}^{m}[U^d(i_k, i_{k+1}) - 1]$. For every k, let i'_k be an element of $R(V^d)$ such that $U^d(i_k, i'_k) = 0$, which exists by part (ii) of the induction hypothesis. Since $i_k \in R(U^d)$, we also have $U^d(i'_k, i_k) = 0$.

Now, using part (i) of the induction hypothesis, the triangle inequality and the above remarks, we have $V^d(i'_k, i'_{k+1}) = U^d(i'_k, i'_{k+1} \leq U^d(i'_k, i_k) + U^d(i_k, i_{k+1}) + U^d(i_{k+1}, i'_{k+1}) = U^d(i_k, i_{k+1})$. Furthermore, $i'_k \in R(V^d)$, by construction, and $V^d(i'_k, i'_{k+1}) > 0$, because otherwise the triangle inequality would yield $U^d(i_k, i_{k+1}) = 0$, which would contradict the assumption $i_{k+1} \notin R_{i_k}(U^d)$. Thus, $i'_{k+1} \notin R_{i'_k}(V^d)$. Therefore, $V^{d+1}(i,j) \leq \sum_{k=1}^{m}[V^d(i'_k, i'_{k+1}) - 1] \leq \sum_{k=1}^{m}[U^d(i_k, i_{k+1}) - 1] = U^{d+1}(i,j)$. This completes the induction step for part (i).

(ii) Let us fix some i and let $j \in R(V^d)$ be such that $U^d(i,j) = 0$. Such a j exists by part (ii) of the induction hypothesis. We also have $U^{d+1}(i,j) = 0$, because $U^{d+1} \leq U^d$. Then, let $k \in R(V^{d+1})$ be such that $V^{d+1}(j,k) = 0$, which exists by Proposition 5.3(ii). Now, k also belongs to $R(V^d)$ (Proposition 5.4(iii)) and therefore, $U^{d+1}(j,k) = V^{d+1}(j,k) = 0$. Using the triangle inequality for U^{d+1}, we obtain $U^{d+1}(i,k) = 0$, and since $k \in R(V^{d+1})$, we have completed the induction step for part (ii).

(iii) We first prove that $R(U^{d+1}) \cap R(V^d) \subset R(V^{d+1})$. Let $i \in R(U^{d+1}) \cap R(V^d)$ and let k be such that $V^{d+1}(i,k) = 0$. We need to show that $V^{d+1}(k,i) = 0$. To show this, we first find some $j \in R(V^{d+1})$ such that $V^{d+1}(k,j) = 0$ (Proposition 5.3(ii)). By the triangle inequality, $V^{d+1}(i,j) = 0$. In particular, $i, j \in R(V^d)$ and by part (i) of the induction hypothesis, we have $U^{d+1}(i,j) = 0$. Furthermore, we have assumed that $i \in R(U^{d+1})$ and this implies that $U^{d+1}(j,i) = 0$ which finally shows that $V^{d+1}(j,i) = 0$. Then the triangle inequality yields $V^{d+1}(k,i) = 0$, as desired.

We now prove the reverse inclusion. Let $i \in R(V^{d+1})$. In particular, $i \in R(V^d)$. In order to prove that $i \in R(U^{d+1})$, let j be such that $U^{d+1}(i,j) = 0$ and we need to show that $U^{d+1}(j,i) = 0$. Let $k \in R(V^{d+1})$ be such that $U^{d+1}(j,k) = 0$. (Such a k exists because part (ii) has already been proved for $d+1$.) By the triangle inequality, $U^{d+1}(i,k) = 0$, and since $i, k \in R(V^d)$, we get $V^{d+1}(i,k) = 0$, using part (i) of the induction hypothesis. Since $i \in R(V^{d+1})$, this yields $V^{d+1}(k,i) = 0$ and therefore $U^{d+1}(k,i) = 0$. Finally, $U^{d+1}(j,i) \leq U^{d+1}(j,k) + U^{d+1}(k,i) = 0$, which completes the proof. •

A further simplification of the formula for U^d is possible.

Proposition 5.6: For any $\alpha \in \mathcal{A}$ we have $HFH\alpha = F\alpha$.

Proof: (Outline) Assume that $j \in R(\alpha)$. By definition, $(HF\alpha)(i,j)$ equals the shortest distance from i to j, with respect to the cost function which is equal to the sum of the coefficients α along a path minus the number of times that the path exits from a set $R_k(\alpha)$.

Notice that $R(\alpha) = R(H\alpha)$ and $R_k(\alpha) = R_k(H\alpha)$, $\forall k \in R(\alpha)$. Thus $(HFH\alpha)(i,j)$ is equal to the shortest distance from i to j with respect to the cost function which is equal to the sum of the

coefficients $(H\alpha)(i,j)$ along a path minus the number of times that the the path exits from a set $R_k(\alpha)$. Given a shortest path for the above defined problem (that is a shortest path with respect to the coefficients $FH\alpha$), we replace each one of its arcs (k,ℓ) by a path for which the sum of the α's along that path equals $(H\alpha)(k,\ell)$. The length (with respect to $H\alpha$) of the original path is equal to the length (with respect to α) of the second path. Furthermore the number of times that a set $R_k(\alpha)$ is exited is at least as large for the second path. This shows that $(HFH\alpha)(i,j) \geq (HF\alpha)(i,j)$. An almost identical argument establishes the same conclusion if $j \notin R(\alpha)$.

On the other hand, $H\alpha \leq \alpha$ and $R(\alpha) = R(H\alpha)$ imply $FH\alpha \leq F\alpha$. The mapping H is clearly monotone, which implies that $HFH\alpha \leq HF\alpha$. This concludes the proof. •

As a corollary of Proposition 5.6 we obtain

$$U^d = HF^d\alpha, \tag{5.7}$$

where F^d is defined by $F^0 = F$ and $F^{d+1}\alpha = F(F^d\alpha)$. This formula is deceivingly simple because in order to apply F on $F^d\alpha$ we must find $R(F^d\alpha)$ and this requires the computation of $HF^d\alpha$. Nevertheless, this formula turns out to be particularly useful for analyzing the case of the simulated annealing algorithm, under a reversibility assumption, which will be done in Section VII.

VI. Necessary and Sufficient Conditions for Convergence under Piecewise Constant Schedules.

We consider a family $\{x_\epsilon(\cdot)\}$ of stationary Markov chains whose one–step transition probabilities satisfy (5.1) and an associated non–stationary Markov chain $x(\cdot)$ which is obtained by varying ϵ according to the schedule determined by equations (4.1), (4.2). For simplicity of exposition, we assume that δ is an integer.

Throughout this section α is fixed once and for all. Let V^d be the collection of coefficients defined in Proposition 5.1 and, for every d let $R^d(\alpha) = R(V^d)$ and $R_i^d(\alpha) = R_i(V^d)$. As long as α is fixed, we will employ the simpler notations R^d and R_i^d.

Proposition 6.1: Assume that the family $\{x_\epsilon(\cdot)\}$ of stationary Markov chains satisfies the assumptions of Proposition 5.1. The the following hold for the above defined non–stationary Markov chain $x(\cdot)$.

(i) $\lim_{k\to\infty} P(x(t_k) \in R^\delta \mid x(1) = i) = 1$;

(ii) If $i \in R^\delta$, then $\liminf_{k\to\infty} P(x(t_k) = i \mid x(1) = i) > 0$.

Proof: Because of equation (4.1), during the interval $[t_k, t_{k+1})$ we are dealing with a stationary Markov chain $x_\epsilon(\cdot)$, where $\epsilon = \epsilon(t_k) = e^{-k}$. We also notice that $t_{k+1} - t_k = \epsilon^{-\delta}$. Thus, Proposition 5.1 is applicable.

Let us fix i as the initial state and let $B_k = P(x(t_k) \in R^\delta)$. Propositions 5.1 and 5.3(ii) show that $P(x(t_{k+1}) \in R^\delta \mid x(t_k) = j) \geq C_1^\delta$, $\forall j \in S$. Furthermore, by Propositions 5.1, 5.3(i), $P(x(t_{k+1}) \notin R^\delta \mid x(t_k) \in R^\delta) \leq C_2^\delta \epsilon(t_k)$, $\forall j \in S$. It follows that $B_{k+1} \geq (1 - B_k)C_1^\delta + B_k(1 - C_2^\delta \epsilon(t_k))$. Using

this last inequality and the fact that $\lim_{k\to\infty} \epsilon(t_k) = 0$, we see that $\lim_{k\to\infty} B_k = 1$ and we obtain part (i).

For the proof of part (ii), let us assume that $i \in R^\delta$ and let $F_k = P(x(t_k) \in R_i^\delta \mid x(0) = i)$. Clearly, $F_1 = 1$. Using Propositions 5.1, 5.3, as before, we conclude that $F_{k+1} \geq (1 - C_2^\delta \epsilon(t_k))F_k = (1 - C_2^\delta e^{-k})F_k$. Since the sequence e^{-k} is summable it follows easily that the sequence F_k is bounded away from zero. Now given that the state at time t_k belongs to R_i^δ, it is easy to show that, for each $j \in R_i^\delta$ the probability that $x(t_k) = j$ is bounded away from zero. (This is because for the time intervals we are dealing with, all states in the same class R_i^δ "communicate" with $O(1)$ probability. This completes the proof. •

Thus, with the schedule (4.1), (4.2), R^δ is the smallest subset of S which gets all the probability, asymptotically. Suppose that our objective is to ensure that equation (4.3) holds; for this, it is necessary and sufficient that $R^\delta \subset S^*$. This, together with the fact that R^δ decreases when δ increases (Proposition 5.4(iii)), leads to the following corollary.

Corollary 6.2: A necessary and sufficient condition for (4.3) to hold, under the schedule (4.1), (4.2), and under the other assumptions of Proposition 6.1, is that $\delta \geq \delta^*$, where δ^* is the smallest integer d such that $R^d \subset S^*$.

Notice that in Proposition 6.1 and Corollary 6.2 we have not made any reversibility assumptions. On the other hand, reversibility turns out to be a useful assumption because it leads to a simple characterization of the sets R^d. We should mention here that without assuming some form of reversibility there may exist no δ such that $R^\delta \subset S^*$. In such a case, Corollary 6.2 simply states that there is no choice of δ such that equation (4.3) holds.

Proposition 6.1 and Corollary 6.2 remain valid under more general circumstances, as long as the schedule $T(\cdot)$ is nonincreasing and converges to zero, as $t \to \infty$. One generalization is the following:

Proposition 6.3: Consider the nonstationary Markov chain $x(\cdot)$ resulting from a schedule $T(\cdot)$ satisfying $\lim_{t\to\infty} T(t) = 0$ Then, (4.3) holds if and only if

$$\sum_{t=1}^{\infty} \exp\left[\frac{\delta^*}{T(t)}\right] = \infty,$$

where δ^* is the smallest value of δ such that $R^\delta \subset S^*$.

Proposition 6.3 (in a continuous time setup) was proved in [13] under a reversibility assumption, slightly more general than (2.4) and with a different definition of δ^*. Of course, that alternative definition of δ^* is equivalent to ours, under the reversibility assumption (see Section VII). The more general version stated above was proved in [14], using a more refined version of the argument in Sections V–VI.

VII. The Value of δ^* for the Reversible Case.

In this section we elaborate on the comment made in the last paragraph of the preceding section,

regarding the possibility of an alternative but equivalent characterization of δ^*, for the reversible case.

Let $\mathcal{F}$ be the set of all functions J from S into the set of nonnegative integers. Given a cost function $J \in \mathcal{F}$, the transition probabilities of the simulated annealing algorithm, at constant temperature, are of the form (3.1), where the set of coefficients α belongs to $\mathcal{A}$ and is uniquely determined by J and the set E of allowed transitions. In this section, we will be assuming that E has been fixed once and for all and that it has the following properties: if $(i,j) \in E$, then $(j,i) \in E$ and E is strongly connected, meaning that there exists a path from every i to every j. We also assume that the structural reversibility assumption (2.4) holds.

Let $G : \mathcal{F} \mapsto \mathcal{A}$ be a mapping that determines the coefficients $\alpha(i,j)$ in terms of J. More precisely, we let

$$(GJ)(i,j) = \infty, \qquad (i,j) \notin E \text{ and } i \neq j, \tag{7.1a}$$

$$(GJ)(i,j) = [J(j) - J(i)]_+, \qquad (i,j) \in E \text{ or } i = j.^\dagger \tag{7.1b}$$

In the light of Corollary 6.2, our objective is to find a characterization of the smallest δ such that $R^\delta(QJ) \subset S^*$.

Let P_{ij} be the set of all paths from i to j. For any $p \in P_{ij}$, let $h_p(i,j;J)$ be the maximum of $J(k)$, over all nodes k belonging to the path p. Let

$$h(i,j;J) = \min_{p \in P_{ij}} h_p(i,j;J). \tag{7.2}.$$

We define the depth $D(i;J)$ of a state $i \in S$ (with respect to the cost function J) to be equal to infinity if i is a global minimum of J and equal to

$$\min_{\{j:\ J(j)<J(i)\}} [h(i,j;J) - J(i)], \tag{7.3}$$

otherwise. Thus, $D(i;J)$ stands for the minimal amount by which the cost has to be temporarily increased in order to get from state i to a state of lower cost. By comparing this with the definition of $R(GJ)$, we see that

$$R(GJ) = \{i :\ D(i;J) > 0\}. \tag{7.4}$$

We now recall equation (5.7) which provides a method for determining the coefficients U^d, for any d. Once these coefficients are computed, the sets $R^d(QJ)$ are also determined by Proposition 5.5(iii). The procedure of Section V works in terms of the structure coefficients α. In the present case, it is GJ which plays the role of α. Given that GJ is determined by the cost function J, it is reasonable to try to reformulate that procedure so that it operates directly on cost functions. The following definition turns out to be appropriate. We define a mapping $T : \mathcal{F} \mapsto \mathcal{F}$ by

$$(TJ)(i) = J(i) + 1, \qquad \text{if } i \in R(GJ), \tag{7.5a}$$

---

† We employ the notation $[x]_+ = \max\{0, x\}$.

$$(TJ)(i) = J(i), \qquad \text{otherwise.} \tag{7.5b}$$

The following result shows that T is isomorphic to the mapping F of Section V.

Proposition 7.1: For any $J \in \mathcal{F}$ we have $FGJ = GTJ$.

Proof: We first prove the following.

Lemma 7.2: If $(i,j) \in E$, then

(i) $[TJ(j) - TJ(i)]_+ = [J(j) - J(i)]_+ - 1$, if $i \in R(GJ)$, $j \notin R_i(GJ)$;

(ii) $[TJ(j) - TJ(i)]_+ = [J(j) - J(i)]_+$, otherwise.

Proof of the Lemma: We consider four cases:

(a) If $i \notin R(GJ)$, $j \notin R(GJ)$, then $TJ(i) = J(i)$ and $TJ(j) = J(j)$ and the result holds.

(b) If $i \notin R(GJ)$ and $j \in R(GJ)$, then $J(i) \geq J(j) + 1$ and therefore $[TJ(j) - TJ(i)]_+ = [J(j) - J(i) - 1]_+ = 0 = [J(j) - J(i)]_+$.

(c) If $i \in R(GJ)$, $j \notin R_i(GJ)$, then $J(j) > J(i)$. (Otherwise, $(GJ)(i,j) = 0$ and $(GJ)(j,i) > 0$, contradicting the assumption $j \in R_i(G)$. Therefore, $[TJ(j) - TJ(i)]_+ = [J(j) - J(i) - 1]_+ = J(j) - J(i) - 1 = [J(j) - J(i)]_+ - 1$ and the result holds.

(d) If $i \in R(GJ)$, $j \in R_i(GJ)$, then $j \in R(GJ)$ and $TJ(i) = J(i) + 1$, $TJ(j) = J(j) + 1$, from which the result follows. This completes the proof of the Lemma. •

The proof of the proposition is completed by comparing the definition of FGJ with GTJ, where TJ is given by Lemma 7.2. and noticing that they are identical. •

Our next result shows that the coefficients U^d defined by (5.6) may be obtained directly from J, by applying the T operation d consecutive times. Let T^d be defined by $T^1 = T$ and $T^d(J) = T(T^{d-1}(J))$.

Proposition 7.3: (i) $U^d = HGT^dJ$.

(ii) $R^d(GJ) = \cap_{c=0}^{d} R(GT^cJ)$.

Proof: (i) From equation (5.7), $U^d = HF^dGJ$ which is equal to HGT^dJ, by Proposition 7.1.

(ii) This is an immediate consequence of Proposition 5.5(iii) and the fact $R(HGT^cJ) = R(GT^cJ)$.

•

Lemma 7.4: If $i \in R(GJ)$ and $j \notin R_i(GJ)$, then $h(i,j;J) = h(i,j;TJ)$.

Proof: We have $TJ \geq J$, which implies that $h(i,j;J) \leq h(i,j;TJ)$. For the converse inequality, let $p \in P_{ij}$ be such that $h(i,j;J) = h_p(i,j;J)$. Let k be a node which maximizes J, over the set of all nodes belonging to the path p. If $J(k) > J(i)$ or $J(k) > J(j)$, then $D(k;J) = 0$ and $k \notin R(GJ)$. Therefore, $(TJ)(k) = J(k)$, which shows that $h_p(i,j;TJ) = h_p(i,j;J) = h(i,j;J)$ and we are done. The case $J(k) < J(i)$ or $J(k) < J(j)$ is impossible; so we are left with the case $J(j) = J(i) = J(k)$. However, this would imply that $h(i,j;J) = J(i)$ and therefore $j \in R_i(GJ)$ which is a contradiction. •

Lemma 7.5: If d is a nonnegative integer and $D(i;J) \geq d > 0$, then $D(i;TJ) \geq d - 1$.

Proof: Suppose that the result is false. Then, there exists some i such that $D(i;J) \geq d > 0$ and there exists some j such that $(TJ)(j) < (TJ)(i)$ and $h(i,j;TJ) - (TJ)(i) \leq d-2$. Since $D(i;J) > 0$,

it follows that $i \in R(GJ)$ and $TJ(i) = J(i) + 1$. Furthermore, $j \notin R_i(GJ)$, because otherwise we would have $J(j) = J(i)$ which would imply that $(TJ)(j) = J(j) + 1 = J(i) + 1 = (TJ)(i)$. We may therefore apply Lemma 7.4 to conclude that $h(i,j;J) - J(i) = h(i,j;TJ) - TJ(i) + 1 \leq d - 1$. Therefore, $D(i;J) \leq d - 1$ which is a contradiction and proves the Lemma. •

Proposition 7.6: $R^d(GJ) = R(GT^dJ) \cap \cdots \cap R(J) = \{i : D(i;J) > d\}$.

Proof: The first equality is simply a restatement of Proposition 7.3(ii). We thus concentrate on the second equality. Suppose that $D(i;J) > d \geq 0$. Then, by Lemma 7.5, $D(i;T^kJ) > d - k \geq 0$, $\forall k \leq d$. Using equation (7.4), it follows that $i \in R(GT^kJ)$, $\forall k \leq d$. This shows that $\{i : D(i;J) > d\} \subset R(GT^dJ) \cap \cdots \cap R(J)$.

We now prove the reverse containment. Suppose that $i \in \cap_{k=0}^{d} R(GT^kJ)$. Using the definition of T, it follows that $(T^dJ)(i) = J(i) + d$. Suppose that $D(i;J) \leq d$. Then, there exists some j such that $J(j) < J(i)$ and some path $p \in P_{ij}$ such that $h_p(i,j;J) \leq J(i) + d$.

For any $k \leq d$ we have $(T^kJ)(j) \leq J(j) + k < J(i) + k = (T^kJ)(i)$. This shows that for any $k \leq d$ we have $j \notin R_i(GT^kJ)$. We then apply Lemma 7.4, d consecutive times, to conclude that $h_p(i,j;J) = h_p(i,j;T^dJ)$. Therefore, $h_p(i,j;T^dJ) - (T^dJ)(i) = h_p(i,j;J) - J(i) - d \leq 0$. Thus, $D(i;T^dJ) = 0$ which shows that $i \notin R(GT^dJ)$, which is a contradiction and concludes the proof of the proposition. •

Proposition 7.6 is the main result of this Section. An immediate corollary is the following result [13].

Corollary 7.7: Consider the simulated annealing algorithm with the schedule determined by (4.1), (4.2). Assume that structural reversibility (2.2) holds, that the graph $G = (S,E)$ is strongly connected and that the zero–temperature algorithm is an aperiodic Markov chain. Then, $P(x(t) \in S^*)$ converges to 1, for every initial state, if and only if the constant δ of equation (4.2) is greater than or equal to $\max_{i \notin S^*} D(i;J)$.

Proof: By Corollary 6.2, convergence to S^*, for every initial state, is obtained if and only if δ is such that $R^\delta(GJ) \subset S^*$. We now use Proposition 7.6 to see that $R^\delta(GJ) = \{i : D(i;J) > \delta\}$. Therefore, the condition on δ is equivalent to the requirement that if $D(i;J) > \delta$, then $i \in S^*$. Equivalently, $\delta \geq \max_{i \notin S^*} D(i;J)$. •

In a certain sense, our method of proving Corollary 7.7 is isomorphic to the proof in [13]. In particular, the mapping T corresponds to "filling the cups", in the terminology of [13]. On the other hand, our approach separates the general probabilistic issues (Sections V, VI) from the special graph–theoretic properties due to reversibility (this Section); in particular, Sections V, VI show that the handling of the probabilistic issues is independent of the reversibility assumptions.

We now have enough machinery available to characterize the constant Δ of Section III. Since $1/\lambda_\epsilon$ is the time needed for the Markov chain to exhibit some mixing, it follows that for times of the order of $\epsilon^{-\Delta}$ there should be a single "recurrent class". Thus, Δ is the smallest δ so that $R^\delta(\alpha)$ consists of a single class $R_i^\delta(\alpha)$. Equivalently, $h(i,j;J) \leq J(i) + \Delta$, $\forall i \in S$, $\forall j \in S^*$. The

above outlined argument establishes the following result, which has been established in [19] using the results of [20].

Proposition 7.8: Under the reversibility assumption (2.2), the constant Δ of equation (3.2) is equal to $\max_{i \in S} \max_{j \in S^*} [h(i,j;J) - J(i)]$.

In particular, we see that $\Delta \geq \delta^*$, as expected. Equality holds if S^* is a singleton, but the inequality may be strict if S^* is not a singleton.

VIII. Discussion.

From Corollary 7.2 we obtain a temperature schedule $T(t) = \frac{\delta^*}{\log t}$ which is has the fastest rate of cooling in the class of schedules for which convergence to the set S^* is obtained. One might be tempted to call this schedule "optimal". However, it can be shown that with this schedule and with a random initial state, the expected time until $x(t)$ first enters the set S^* is, in general, infinite. (This is easily verified with an example which has only two states with different costs.)

In all results presented in this paper we talk about the asymptotic behavior of simulated annealing for a fixed state space S and a fixed cost function J. Thus, we ignore the dependence of the parameters of interest on the size N of the state space. However, if one is to compare the algorithmic efficiency of simulated annealing with other available algorithms, it is precisely this dependence on N that has to be analyzed. A first result of this type has been obtained in [8] but more research of this nature is needed.

REFERENCES

[1] Kirkpatrick, S., Gelatt, C.D.,Jr., Vecchi, M.P., "Optimization by Simulated Annealing", *Science,* Vol. 220, 1983, pp. 671–680.

[2] Cerny, V., "A Thermodynamic Approach to the Travelling Salesman Problem: an Efficient Simulation Algorithm", Technical Report, Institute of Physics and Biophysics, Comenius Univ., Bratislava, 1982.

[3] Metropolis, N., Rosenblith, A., Rosenblith, M., Teller, A., Teller, E., " Equation of State Calculations by Fast Computing Machines", *J. Chem. Physics*, 21, 1953, pp. 1087–1092.

[4] Bonomi, E., Lutoon, J.L., "The N–City Travelling Salesman Problem: Statistical Mechanics and the Metropolis Algorithm", *SIAM Review*, 26, 4, 1984, pp. 551–568.

[5] Kirkpatrick, S., Toulouse, G., "Configuration Space Analysis of Travelling Salesman Problems", *J. Physique*, 46, 1985, pp. 1227–1292.

[6] Vecchi, M.P., Kirkpatrick, S., "Global Wiring by Simulated Annealing", *IEEE Trans. on Computer–Aided Design*, 2, 1983, pp. 215–222.

[7] Geman, S., Geman, D., "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 6, pp. 721–741.

[8] Sasaki, G.H., Hajek, B., "The Time Complexity of Maximum Matching by Simulated Annealing", Technical Report LIDS-P-1552, Laboratory for Information and Decision Systems, Mas-

sachusetts Institute of Technology, Cambridge, MA, 1986.

[9] Gelfand, S., Mitter, S.K., "Analysis of Simulated Annealing for Optimization", *Proceedings of the 24th IEEE Conference on Decision and Control*, Fort Lauderdale, Florida, 1985, pp. 779–786.

[10] Mitra, D., Romeo, F., Sangiovanni–Vincentelli, A., "Convergence and Finite–Time Behavior of Simulated Annealing", *Proceedings of the 24th IEEE Conference on Decision and Control*, Fort Lauderdale, Florida, 1985, pp. 761–767.

[11] Gidas, B., "Global Optimization via the Langevin Equation", *Proceedings of the 24th IEEE Conference on Decision and Control*, Fort Lauderdale, Florida, 1985, pp. 774–778.

[12] Gidas, B., "Non–Stationary Markov Chains and Convergence of the Annealing Algorithm", *J. Statistical Physics*, 39, 1985, pp. 73–131.

[13] Hajek, B., "Cooling Schedules for Optimal Annealing", submitted to *Mathematics of Operations Research*, 1985.

[14] Tsitsiklis, J.N., "Markov Chains with Rare Transitions and Simulated Annealing", submitted to *Mathematics of Operations Research*, 1985.

[15] Coderch, M., "Multiple Time Scale Approach to Hierarchical Aggregation of Linear Systems and Finite State Markov Processes", Ph.D. Thesis, Dept. of Electrical Engineering, M.I.T., 1982.

[16] Coderch, M., Willsky, A.S., Sastry, S.S., Castanon, D.A., Hierarchical Aggregation of Singularly Perturbed Finite State Markov Processes", *Stochastics,* 8, 1983, pp. 259–289.

[17] Rohlicek J. R., Willsky, A. S., "The Reduction of Perturbed Markov Generators: An Algorithm Exposing the Role of Transient States", Technical Report LIDS–P–1493, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[18] Delebecque, F., "A Reduction Process for Perturbed Markov Chains", *SIAM J. of Applied Mathematics*, 43, 2, 1983.

[19] Chiang, T.S., Chow, Y., "On Eigenvalues and Optimal Annealing Rate", submitted to *Mathematics of Operations Research*, 1986.

[20] Ventcel, A.D., "On the Asymptotics of Eigenvalues of Matrices with Elements of Order $\exp(-V_{ij}/(2\epsilon^2))$", *Dokl. Akad. Nauk SSR*, 202, 1972, pp. 65–68.