LIDS-P-1497

August 1985
(revised November 1985)

MARKOV CHAINS WITH RARE TRANSITIONS AND SIMULATED ANNEALING[1]

John N. Tsitsiklis[2]

ABSTRACT

We consider Markov chains in which the entries of the one–step transition probability matrix are known to be of different orders of magnitude and whose structure (that is, the orders of magnitude of the transition probabilities) does not change with time. For such Markov chains we present a method for generating order of magnitude estimates for the t–step transition probabilities, for any t. We then notice that algorithms of the simulated annealing type may be represented by a Markov chain which is approximately stationary over fairly long time intervals. Using our results we obtain a characterization of the convergent "cooling" schedules for the most general class of algorithms of the simulated annealing type.

1. Research supported by the Army Research Office under contract DAAG–29–84–K–005.

2. Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139.

I. INTRODUCTION.

The main objective of this paper is th characterization of the cooling schedules under which a simulated annealing algorithm converges to a set of desired states, such as the set where some cost function is minimized, thus generalizing the results of Hajek [9]. The added generality over the results of [9] may turn out to be useful in parallel simulated annelaing algorithms in which some of the assumptions of [9] may not hold. The method we follow is based on the observation that in simulated annealing algorithms the "temperature" remains approximately constant for sufficiently long times. For this reason, we may exploit bounds and estimates which are valid for singularly perturbed, approximately stationary Markov chains and obtain interesting conclusions for simulated annealing algorithms. In the course of developing our results on simulated annealing we derive certain results on approximately stationary singularly perturbed Markov chains which seem to be of independent interest.

The structure of the paper is the following. In Section 2 we assume that we are dealing with a discrete time Markov chain in which each of the one–step transition probabilities is roughly proportional to a certain power of ϵ, where ϵ is a small parameter. We then present an algorithm, consisting of the solution of certain shortest path problems and some graph theoretic manipulations, which provides estimates for the transition probabilities of the Markov chain for any time between 0 and $1/\epsilon$. Then, in Section 3, we indicate how the procedure of Section 2 may be applied recursively to produce similar estimates on the transition probabilities for all times. In Section 4 we use the results of Section 3 to characterize the convergence of simulated annealing algorithms. In Section 5 we discuss briefly the continuous time versions of our results.

II. MARKOV CHAINS PARAMETRIZED BY A SMALL PARAMETER.

In this Section we derive order of magnitude estimates on the transition probabilities of a non–stationary discrete time Markov chain. Our results are based on the assumption that such order of magnitude information is available on the one–step transition probabilities of the Markov chain.

We start with some notation. We use $\mathcal{N}$ and $\mathcal{N}_0$ to denote the positive and the nonnegative integers, respectively. We also let $\mathcal{U}$ denote the set of functions $f : (0,\infty) \mapsto (0,\infty)$ such that for every $n \in \mathcal{N}_0$ there exists some $c_n > 0$ such that $f(\epsilon) \leq c_n \epsilon^n$, $\forall \epsilon > 0$. Notice that $\mathcal{U}$ has the property that $f(\epsilon)/\epsilon^n \in \mathcal{U}$, $\forall f \in \mathcal{U}$, $\forall n \in \mathcal{N}$. Also notice that $c^{1/\epsilon} \in \mathcal{U}$, for any $c \in (0,1)$.

We consider a (generally non–stationary) finite state, discrete time Markov chain $X = \{x(t) : t \geq 0\}$ with state space $S = \{1,...,N\}$. For any $t \geq 0$ we let $q_{ij}(t) = P(x(t+1) = j \,|\, x(t) = i)$ and $p_{ij}(t) = P(x(t) = j \,|\, x(0) = i)$. We assume that some structural information is available on this Markov chain. More precisely, let there be given a collection $\mathcal{A} = \{\alpha_{ij} : 1 \leq i,j \leq N\}$ of elements of $\mathcal{N}_0 \cup \{\infty\}$. Let $f \in \mathcal{U}$ and let C_1, C_2 be positive constants. We assume that for some $\epsilon > 0$ we

have

$$C_1\epsilon^{\alpha_{ij}} \leq q_{ij}(t) \leq C_2\epsilon^{\alpha_{ij}}, \qquad \forall t \geq 0, \quad \text{if } \alpha_{ij} < \infty, \tag{2.1}$$

$$0 \leq q_{ij}(t) \leq f(\epsilon), \qquad \forall t \geq 0, \quad \text{if } \alpha_{ij} = \infty. \tag{2.2}$$

It is also natural to assume that for each i there exists some j such that $\alpha_{ij} = 0$, because otherwise no Markov chain could satisfy (2.1) and (2.2) for ϵ small enough. We call $\mathcal{A}$ the **structure** of the Markov chain X. We will now assume that $\mathcal{A}, C_1, C_2, f$ are fixed and we denote by $\mathcal{M}_\epsilon(\mathcal{A}, C_1, C_2, f)$ the set of all Markov chains for which (2.1) and (2.2) hold. (Occasionally we use the shorter notation $\mathcal{M}_\epsilon$, provided that no confusion may arise.)

We classify the states in the state space by considering a Markov chain in which only those transitions from i to j with $\alpha_{ij} = 0$ are allowed. Formally, we define a **path** from i to j to be a sequence $(i_1, ..., i_m)$ of (not necessarily distinct) states such that $i_1 = i$, $i_m = j$ and $m \geq 2$. The **length** of such a path is defined as $\sum_{k=1}^{m-1} \alpha_{i_k i_{k+1}}$. A state i is called **transient** if there exists a state j and a zero length path from i to j but no zero length path from j to i. Otherwise, i is called **recurrent**. This coincides with the usual definition if $\epsilon = 0$ and the Markov chain is stationary. Let TR, R denote the sets of transient and recurrent states, respectively. For any $i \in R$, we let R_i be the set of all j such that there exists a zero length path from i to j.

Lemma 2.1: $j \in R_i$ if and only if $j \in R$ and $i \in R_j$.

Proof: If $j \in R_i$ but $j \notin R$, then there exists a zero length path from i to j and a zero length path from j to some k but no zero length path from k to j. It follows that there exists a zero length path from i to k but no zero length path from k to i, which contradicts the assumption $i \in R$. Thus, $j \in R_i$. The fact that $i \in R_j$ is then obvious. •

In view of the above Lemma, the sets R_i determine a partition of R into disjoint classes which is analogous to the usual partition of recurrent states into ergodic classes for stationary Markov chains. We now introduce the following assumption on $\mathcal{A}$:

Assumption TRI: For any i, j, k, such that at least one of them belongs to R, we have

$$\alpha_{ik} \leq \alpha_{ij} + \alpha_{jk}. \tag{2.3}$$

This assumption is made for convenience because it leads to some simplification of the proofs. It will be removed in the end of this section. We now collect a few useful consequences of Assumption TRI.

Lemma 2.2: Under Assumption TRI:

(i) If $i \in R$ and $\alpha_{ij} = 0$, $i \neq j$, then $j \in R$ and $\alpha_{ji} = 0$.

(ii) If $i \in R$, then $\alpha_{ii} = 0$.

(iii) If $i \in R$ and $j \in TR$, then $\alpha_{ij} \geq 1$.

(iv) If $i \in TR$, then there exists some $j \in R$ such that $\alpha_{ij} = 0$. In particular, $P(x(t+1) \in TR \,|\, x(t) = i) \leq 1 - C_1$, $\forall i \in TR$.

Proof: (i) If $i \in R$ and $\alpha_{ij} = 0$, then there exists a zero length path from j to i, because of the

definition of R. We then apply Assumption TRI along this path to obtain $\alpha_{ji} = 0$.
(ii) Let $i \in R$. By assumption, there exists some j such that $\alpha_{ij} = 0$. By part (i) of the Lemma, we also have $\alpha_{ji} = 0$. Using Assumption TRI, we obtain $\alpha_{ii} \leq \alpha_{ij} + \alpha_{ji} = 0$.
(iii) This is an immediate consequence of part (i).
(iv) Given $i \in TR$, we define a finite sequence $(i_1, ..., i_n)$ of distinct states as follows. Let $i_1 = i$. Having chosen i_m, if $i_m \in R$ let $n = m$ and stop. Otherwise, choose i_{m+1} so that there exists a zero length path from i_m to i_{m+1} but no such path from i_{m+1} to i_m. (Such an i_{m+1} exists, by the definition of TR and must be different from $i_1, ..., i_m$ because otherwise there would be a zero length path from i_m to i_{m-1}.) Since the state space is finite, the termination condition must be met eventually. Thus there exists some $j \in R$ and a zero length path from i to j. We then use (inductively) Assumption TRI along this path, to conclude that $\alpha_{ij} = 0$. The last statement of the Lemma is then an immediate consequence of (2.1). •

We need a preliminary result which provides order of magnitude estimates on the probability that a state $j \in R$ is the first recurrent state to be visited, starting from a transient state i. We use the notation $T = \min\{t \geq 0 : x(t) \in R\}$. We also use the convention that $\epsilon^\infty = 0$.
Proposition 2.1: There exist $F > 0$ and $g \in \mathcal{U}$ such that for any $\epsilon > 0$, $X \in \mathcal{M}_\epsilon$, $i \in TR$, $j \in R$ we have

$$C_1 \epsilon^{\alpha_{ij}} \leq P(x(T) = j \,|\, x(0) = i) \leq F\epsilon^{\alpha_{ij}} + g(\epsilon). \tag{2.4}$$

Proof: Let us fix some $j \in R$. We define, for $\alpha \in \mathcal{N}_0 \cup \{\infty\}$, $S_\alpha = \{i \in TR : \alpha_{ij} = \alpha\}$ and $Q_\alpha = \{i \in TR : \alpha_{ij} \geq \alpha\}$. We then define $p_{\alpha,\epsilon} = \sup_{X \in \mathcal{M}_\epsilon} \max_{i \in Q_\alpha} P(x(T) = j \,|\, x(0) = i)$. We first prove, by induction on α, that for any $\alpha < \infty$ there exists some $F_\alpha > 0$ such that $p_{\alpha,\epsilon} \leq F_\alpha \epsilon^\alpha$, $\forall \epsilon > 0$. This is clearly true for $\alpha = 0$. Suppose it is true for all α less than some positive integer β. Let $i \in Q_\beta$ and $X \in \mathcal{M}_\epsilon$. Notice that for any state k we have $\alpha_{ik} + \alpha_{kj} \geq \alpha_{ij} \geq \beta$, because of Assumption TRI. Using (2.1) and the induction hypothesis we obtain

$$P(x(T) = j \,|\, x(0) = i) \leq \sum_{\alpha=0}^{\beta-1} \sum_{k \in S_\alpha} P(x(T) = j \,|\, x(1) = k) P(x(1) = k \,|\, x(0) = i) +$$

$$+ P(x(1) \in Q_\beta \,|\, x(0) = i) \max_{l \in Q_\beta} P(x(T) = j \,|\, x(1) = l) + P(x(1) = j \,|\, x(0) = i) \leq$$

$$\sum_{\alpha=0}^{\beta-1} \sum_{k \in S_\alpha} F_\alpha \epsilon^{\alpha_{kj}} C_2 \epsilon^{\alpha_{ik}} + (1 - C_1) p_{\beta,\epsilon} + C_2 \epsilon^\beta \leq$$

$$[N \max_{\alpha < \beta} \{F_\alpha\} C_2 + C_2] \epsilon^\beta + (1 - C_1) p_{\beta,\epsilon}.$$

Taking the supremum of the left hand side over all $i \in Q_\beta$ and all $X \in \mathcal{M}_\epsilon$, we obtain, for some constant F,

$$p_{\beta,\epsilon} \leq F\epsilon^\beta + (1 - C_1) p_{\beta,\epsilon}$$

from which it follows that the induction hypothesis is also true for β.

Finally, we assume that $i \in S_\infty$. Then,

$$P(x(T) = j \,|\, x(0) = i) \;\leq$$

$$P(x(1) \in TR,\; x(1) \notin S_\infty \,|\, x(0) = i) + P(x(1) = j \,|\, x(0) = i) + P(x(1) \in S_\infty \,|\, x(0) = i) p_{\infty,\epsilon} \;\leq$$

$$Nf(\epsilon) + (1 - C_1) p_{\infty,\epsilon}.$$

Thus, $p_{\infty,\epsilon} \leq (N/C_1) f(\epsilon)$, $\forall \epsilon > 0$. This completes the proof of the second inequality in (2.4). The first inequality is a trivial consequence of (2.1). •

Let us mention another method for proving Proposition 2.1. We could first prove it for stationary Markov chains in M_ϵ, because in this case there are explicit formulae for the absorption probabilities. (Such is a result is obtained in [12].) Then, we notice that $p_{\alpha,\epsilon}$ is bounded above by the absorption probabilities which would result if an adversary was allowed to choose $q_{ij}(t)$ at each time t after observing the current state, subject to the constraints (2.1) and (2.2). It follows from standard results in Markovian decision theory that the optimal policy for the adversary is a stationary one and therefore the bounds obtained for stationary Markov chains also apply to the nonstationary ones. Unfortunately, this method does not seem to work for our subsequent results because they correspond to a maximization over a finite horizon for which stationary policies are not in general optimal.

Let us also point out that Proposition 2.1 is false if Assumption TRI is removed.

The main result of this section is based on the following algorithm which provides important structural information on the long run behavior of Markov chains in M_ϵ.

Algorithm I: (Input: $A = \{\alpha_{ij} : 1 \leq i, j \leq N\}$; Output: $V = \{V(i,j) : 1 \leq i, j \leq N\}$, sets $R \subset S$, $TR \subset S$ and for each $i \in R$ a set $R_i \subset R$.)

1. Given A, determine R, TR and the classes R_i using the definition given earlier.

2. Let $c_{ij} = \alpha_{ij} - 1$, if $i \in R$, $j \in R$, $j \notin R_i$ and $c_{ij} = \alpha_{ij}$, otherwise. (Notice that $c_{ij} \geq 0$ always holds.)

3. Solve the shortest path problem from any origin $i \in R$ to any destination $j \in R$, with respect to the link lengths c_{ij} and subject to the constraint that any intermediate state on a path must be an element of R. Let $V(i,j)$ be the length of such a shortest path. Notice that the $V(i,j)$'s satisfy $V(i,i) = 0$, $\forall i \in R$ and

$$V(i,j) = \min_{k \in R} \{V(i,k) + c_{kj}\}, \qquad i, j \in R. \tag{2.5}$$

4. If $i \in R$, $j \in TR$, let

$$V(i,j) = \min_{k \in R} \{V(i,k) + c_{kj}\} = \min_{k \in R} \{V(i,k) + \alpha_{kj}\}. \tag{2.6}$$

5. If $i \in TR$, let

$$V(i,j) = \min_{k \in R}\{c_{ik} + V(k,j)\} = \min_{k \in R}\{\alpha_{ik} + V(k,j)\}. \tag{2.7}$$

Notice that the output $V(i,j)$ of the above algorithm is equal to the length (with respect to the c_{ij}'s) of a shortest path from i to j subject to the constraint that all states on the path belong to R, except possibly for the first and the last one. We continue with a few elementary observations on this algorithm:

Proposition 2.2: (i) $V(i,j) \geq 0$, $\forall i,j$.

(ii) $V(i,j) \geq 1$, $\forall i$, $\forall j \in TR$.

(iii) $V(i,j) \leq V(i,k) + V(k,j)$, $\forall i,j,k$.

(iv) If $j \in R$ and $j' \in R_j$, then $V(i,j) = V(i,j')$, $\forall i$. Also, If $i \in R$ and $i' \in R_i$, then $V(i,j) = V(i',j)$, $\forall j$.

Proof: Part (i) follows from the shortest path interpretation and the nonnegativity of the c_{ij}'s. Part (ii) follows from (2.6) and the fact (Lemma 2.2(iii)) that $\alpha_{kj} \geq 1$, whenever $k \in R$ and $j \in TR$. Part (iii) is clearly true for $k \in R$, due to the shortest path interpretation. So, assume that $k \in TR$. Let us take shortest paths from i to k (of length $V(i,k)$) and from k to j (of length $V(k,j)$) and concatenate them. This produces a path from i to j, of length $V(i,k) + V(k,j)$, such that all intermediate states, except from k, belong to R. If k_1 and k_2 are the predecessor and the successor, respectively, of k in this path, we have $k_1 \in R$, $k_2 \in R$ and we can use (2.3) to conclude that $c_{k_1 k} + c_{k k_2} \geq c_{k_1 k_2}$ which shows that k may be eliminated from this path, to produce a path from i to j, with all intermediate states belonging to R, and with length less than or equal to $V(i,k) + V(k,j)$, as desired. Finally, part (iv) follows from the shortest path interpretation and the fact that $c_{kl} = \alpha_{kl} = 0$, whenever $k \in R$ and $l \in R_k$, which is a straightforward consequence of Assumption TRI. •

We notice that, as a consequence of part (iv) of the proposition, the algorithm need not be carried out for all states. It suffices to consider transient states and one representative from each class R_i.

The following proposition establishes the relevance of the $V(i,j)$'s to the Markov chains under study.

Proposition 2.3: For any $C_3 > 0$, there exist positive constants G_1, G_2, G_3, G_4, with $G_4 < 1$, and some $g \in \mathcal{U}$ such that, for any $\epsilon > 0$, for any Markov chain in $\mathcal{M}_\epsilon$ and any states i, j we have

$$G_1(\epsilon(t-N))^N \epsilon^{V(i,j)} \leq p_{ij}(t) \leq G_2 \epsilon^{V(i,j)} + \chi_i G_3 G_4^t \epsilon^{\alpha_{ij}} + g(\epsilon), \quad \forall t \in [N, C_3/\epsilon], \tag{2.8}$$

where $\chi_i = 0$, if $i \in R$, and $\chi_i = 1$, otherwise. (The upper bound in (2.8) is also true for $t \in [1,N]$.) In particular, there exist $G_1 > 0$, $G_2 > 0$, $g \in \mathcal{U}$ such that

$$G_1 \epsilon^{V(i,j)} \leq p_{ij}(\frac{1}{\epsilon}) \leq G_2 \epsilon^{V(i,j)} + g(\epsilon). \tag{2.9}$$

Proof: Notice that for any $i \in R$, $j \notin R_i$ we have $q_{ij}(t) \leq C_2\epsilon$, $\forall t$. It follows that $P(x(t+1) \notin R_i \,|\, x(t) \in R_i) \leq NC_2\epsilon$, from which we easily conclude that there exists some $F_1 > 0$ such that

$$P(x(t) \in R_i \,|\, x(s) \in R_i) \;\geq\; F_1, \quad 0 \leq s \leq t \leq C_3/\epsilon,\ \forall \epsilon > 0,\ \forall X \in \mathcal{M}_\epsilon,\ \forall i \in R. \tag{2.10}$$

We now start the proof of the lower bound in (2.8). If $V(i,j) = \infty$, there is nothing to prove, so we will be assuming that $V(i,j) < \infty$. We first assume that $i \in R$ and $j \in R$. Then, there exists a sequence $i = i_1, i_2, ..., i_n = j$ of elements of R, (with $n \leq N$) such that $\sum_{k=1}^{n-1} c_{i_k i_{k+1}} = V(i,j)$. Let $k \in \{1, ..., n\}$ and suppose that there exists some $F_k > 0$ such that, for all $\epsilon > 0$ and for all $X \in \mathcal{M}_\epsilon$,

$$P(x(t) \in R_{i_k} \,|\, x(0) = i) \;\geq\; F_k(\epsilon(t-k+1))^{k-1}\epsilon^{\sum_{l=1}^{k-1} c_{i_l i_{l+1}}}, \qquad \forall t \in [k-1, C_3/\epsilon]. \tag{2.11}$$

We then have

$$P(x(t) \in R_{i_{k+1}} \,|\, x(0) = i) \;\geq$$

$$\sum_{s=0}^{t-1} P(x(t) \in R_{i_{k+1}} \,|\, x(s+1) \in R_{i_{k+1}})\, P(x(s+1) \in R_{i_{k+1}} \,|\, x(s) \in R_{i_k})\, P(x(s) \in R_{i_k} \,|\, x(0) = i) \;\geq$$

$$\sum_{s=k}^{t-1} F_1 C_1 \epsilon^{\alpha_{i_k i_{k+1}}} \left(F_k(\epsilon(s-k+1))^{k-1}\epsilon^{\sum_{l=1}^{k-1} c_{i_l i_{l+1}}} \right) \;\geq$$

$$(F_k F_1 C_1)\epsilon^{\sum_{l=1}^{k} c_{i_l i_{l+1}}} \epsilon \sum_{s=k}^{t-1} (\epsilon(s-k+1))^{k-1}. \tag{2.12}$$

Clearly, there exists a constant F_k' such that

$$\sum_{s=k}^{t-1} (s-k+1)^{k-1} \;\geq\; F_k'(t-k)^k, \quad \forall t.$$

Inequality (2.10) shows that (2.11) holds for $k = 1$. We have thus proved by induction on k that (2.11) holds for all k. Notice that

$$P(x(t) = j \,|\, x(0) = i) \;\geq\; P(x(t) = j \,|\, x(t-1) \in R_j)\; P(x(t-1) \in R_j \,|\, x(0) = i) \;\geq$$

$$C_1 P(x(t-1) \in R_j \,|\, x(0) = i),$$

which completes the proof of the left hand side of (2.8), for the case where $i \in R$ and $j \in R$.

Suppose now that $i \in R$, $j \in TR$ and let $k \in R$ be such that $V(i,j) = V(i,k) + \alpha_{kj}$. If $\alpha_{kj} = \infty$, then $V(i,j) = \infty$ and there is nothing to prove. So, assume that $\alpha_{kj} < \infty$. Then,

$$P(x(t) = j \,|\, x(0) = i) \;\geq\; P(x(t) = j \,|\, x(t-1) = k)\, P(x(t-1) = k \,|\, x(0) = i) \;\geq$$

$$C_1\epsilon^{\alpha_{kj}} P(x(t-1)=k \,|\, x(0)=i).$$

Given that we have already proved the lower bound for $p_{ik}(t)$, the desired result for $p_{ij}(t)$ follows.

Finally, let $i \in TR$. The result follows similarly by choosing $k \in R$ so that $\alpha_{ik}+V(k,j)=V(i,j)$ and using the inequality

$$P(x(t)=j \,|\, x(0)=i) \;\geq\; P(x(1)=k \,|\, x(0)=i)\, P(x(t)=j \,|\, x(1)=k).$$

We now turn to the proof of the upper bound in (2.8). Let $i \in R$ be fixed. We define $E_\alpha = \{j \in R:\ V(i,j)=\alpha\}$, $T_\alpha = \{j \in TR:\ V(i,j)=\alpha\}$, $E_{\leq\alpha} = \cup_{\beta\leq\alpha} E_\beta$. We also define similarly $E_{\geq\alpha}$, $T_{\leq\alpha}$, $T_{\geq\alpha}$. We will prove by induction that for any $\alpha < \infty$ the following statements hold:
(SE_α) : There exists some G_α such that $\forall\epsilon > 0$, $\forall X \in \mathcal{M}_\epsilon$, $\forall j \in E_{\geq\alpha}$ and $\forall t \leq C_3/\epsilon$ we have $p_{ij}(t) \leq G_\alpha\epsilon^\alpha$.
(ST_α) : There exists some G'_α such that $\forall\epsilon > 0$, $\forall X \in \mathcal{M}_\epsilon$, $\forall j \in T_{\geq\alpha}$ and $\forall t \leq C_3/\epsilon$ we have $p_{ij}(t) \leq G'_\alpha\epsilon^\alpha$.

Statements SE_0 and ST_0 are trivially true, with $G_0 = G'_0 = 1$. We now prove ST_1. (Notice that $T_{\geq 1}$ =TR.) Now,

$$P(x(t+1) \in TR \,|\, x(0)=i) \;\leq$$

$$P(x(t+1) \in TR \,|\, x(t) \in TR)\, P(x(t) \in TR \,|\, x(0)=i) + P(x(t+1) \in TR \,|\, x(t) \in R) \;\leq$$

$$(1-C_1) P(x(t) \in TR \,|\, x(0)=i) + NC_2\epsilon. \tag{2.13}$$

Since $i \in R$, $P(x(0) \in TR \,|\, x(0)=i) = 0$ and (2.13) implies $P(x(t) \in TR \,|\, x(0)=i) \leq (NC_2\epsilon)/C_1$, $\forall t \geq 0$, which proves ST_1.

Now let α be some positive integer and assume that statements $SE_{\beta-1}$ and ST_β are true, for all $\beta \leq \alpha$. We will prove that SE_α and $ST_{\alpha+1}$ are also true. We first need the following Lemma.
Lemma 2.3: If $j \in J = E_{\leq(\alpha-1)} \cup T_{\leq\alpha}$ and $k \in K = E_{\geq\alpha} \cup T_{\geq(\alpha+1)}$, then $V(i,j)+\alpha_{jk} \geq \alpha+1$.
Proof: (i) If $j \in E_{\leq(\alpha-1)}$, $k \in E_{\geq\alpha}$, then $V(i,j)+\alpha_{jk} = V(i,j)+c_{jk}+1 \geq V(i,k)+1 \geq \alpha+1$.
(ii) If $j \in E_{\leq(\alpha-1)}$, $k \in T_{\geq(\alpha+1)}$, then $V(i,j)+\alpha_{jk} = V(i,j)+c_{jk} \geq V(i,k) \geq \alpha+1$.
(iii) If $j \in T_{\leq\alpha}$, $k \in E_{\geq\alpha}$, let $l \in R$ be such that $V(i,l)+\alpha_{lj} = V(i,j)$. Suppose that $l \in R_k$. Then, $V(i,l) = V(i,k) \geq \alpha$ and $V(i,j) = V(i,l)+\alpha_{lj} \geq \alpha+1$, which contradicts the assumption $j \in T_{\leq\alpha}$ We thus assume that $l \notin R_k$. Then, $V(i,j)+\alpha_{jk} = V(i,l)+\alpha_{lj}+\alpha_{jk} \geq V(i,l)+\alpha_{lk} = V(i,l)+c_{lk}+1 \geq V(i,k)+1 \geq \alpha+1$.
(iv) If $j \in T_{\leq\alpha}$, $k \in T_{\geq(\alpha+1)}$, let $l \in R$ be such that $V(i,l)+\alpha_{lj} = V(i,j)$. Then, $V(i,j)+\alpha_{jk} = V(i,l)+\alpha_{lj}+\alpha_{jk} \geq V(i,l)+\alpha_{lk} \geq V(i,k) \geq \alpha+1$. •

We now use the induction hypothesis and Lemma 2.3 to obtain

$$P(x(t+1) \in K \,|\, x(t) \in J)\, P(x(t) \in J \,|\, x(0)=i) \;\leq$$

$$\sum_{k\in K,\ j\in J} P(x(t+1)=k \,|\, x(t)=j) P(x(t)=j \,|\, x(0)=i) \;\leq$$

$$\sum_{k\in K,\ j\in J} C_2\epsilon^{\alpha_{jk}}G\epsilon^{V(i,j)} \ \leq (N^2C_2G)\epsilon^{\alpha+1},$$

where $G = \max\{G_{\beta-1}, G'_\beta;\ \beta \leq \alpha\}$. It follows that

$$P(x(t)\in K \mid x(0)=i) \ \leq \ (N^2C_2G)\epsilon^{\alpha+1}C_3/\epsilon, \quad \forall t\in[1,C_3/\epsilon],$$

which proves SE_α. Finally,

$$P(x(t+1)\in T_{\geq\alpha+1} \mid x(0)=i) \ \leq \ (1-C_1)P(x(t)\in T_{\geq\alpha+1} \mid x(0)=i) + NG_\alpha\epsilon^\alpha C_2\epsilon + N^2C_2G\epsilon^{\alpha+1}$$

which shows that

$$P(x(t)\in T_{\geq\alpha+1} \mid x(0)=i) \leq (1/C_1)(NG_\alpha C_2 + N^2C_2G)\epsilon^{\alpha+1}, \quad \forall t\in[1,C_3/\epsilon].$$

This proves ST_α and completes the induction.

We have thus completed the proof of the upper bound in (2.9) for the case where $i\in R$ and $V(i,j)<\infty$. The proof for the case $i\in R$ and $V(i,j)=\infty$ is very simple and is omitted. We now assume that $i\in TR$. Let T be the random time of Proposition 2.1. Then, for some $F>0$, $G>0$, $g,g',g''\in\mathcal{U}$, we have

$$p_{ij}(t) \leq$$

$$P(T>t) + \sum_{k\in R} P(x(t)=j \mid x(T)=k, T\leq t)\ P(x(T)=k, T\leq t \mid x(0)=i) \ \leq$$

$$(1-C_1)^t + \sum_{k\in R}[G\epsilon^{V(k,j)}+g(\epsilon)][F\epsilon^{\alpha_{ik}}+g'(\epsilon)] \leq$$

$$(1-C_1)^t + NGF\epsilon^{V(i,j)} + g''(\epsilon), \qquad \forall t\in[1,C_3/\epsilon].$$

This completes the proof of the proposition. •

Notice that the upper and lower bounds are tight, within a multiplicative constant independent of ϵ, when $t=1/\epsilon$. For smaller times the bounds are much further apart. It is not hard to close this gap, although we do not need to do this for our purposes. In particular, the exponent in the term $(\epsilon(t-N))^N$ in the lower bound may be reduced. This may be accomplished with a minor modification of the induction hypothesis in the proof of the lower bound. The upper bound may be also improved in a similar manner.

The remainder of this section is devoted to showing that Assumption TRI is not an essential restriction. Roughly speaking, we will establish that our results are applicable to any Markov chain which is aperiodic in the fastest time scale.

Let there be given a set $\mathcal{A} = \{\alpha_{ij} :\ 1\leq i,j\leq N\}$ of elements of $\mathcal{N}_0\cup\{\infty\}$, not necessarily satisfying (2.3). For any $i,j\in S$, $n\in\mathcal{N}$, we define $\beta^1_{ij}=\alpha_{ij}$ and $\beta^{n+1}_{ij}=\min_k\{\beta^n_{ik}+\alpha_{kj}\}$. That is, β^n_{ij} equals the distance of a shortest path (with respect to the link lengths α_{ij}) from i to j

which has exactly n hops. Also, let $\beta_{ij} = \min_{n \in \mathcal{N}} \beta_{ij}^n$. Notice that for every i, j, k, m, n we have $\beta_{ik}^m + \beta_{kj}^n \geq \beta_{ij}^{m+n}$. We introduce the following assumption on $\mathcal{A}$.

Assumption AP: There exists some positive integer M with the following property: if $i \in R$ or $j \in R$, then $\beta_{ij}^n = \beta_{ij}$, $\forall n \geq M$.

For any Markov chain whose structure is described by $\mathcal{A}$, meaning that the estimates (2.1), (2.2) are valid, assumption AP amounts to the following: if we substitute 0 for ϵ, and decompose the resulting Markov chain into ergodic classes, in the usual manner, then each of the non–communicating classes of recurrent states is aperiodic.

It can be shown that if $\mathcal{A}$ satisfies assumption AP, then M can be chosen to be smaller than N^2. (This is related to the fact that the "index of primitivity" of any primitive nonnegative matrix is bounded above by $N^2 - 2N + 2$; for more details, see Chapter 2 of [13].)

Given $\mathcal{A}$, some positive constants C_1, C_2, some $f \in \mathcal{U}$ and some $\epsilon > 0$, consider the set $\mathcal{M}_\epsilon(\mathcal{A}, C_1, C_2, f)$. Let Q be some positive integer. For any $X \in \mathcal{M}_\epsilon(\mathcal{A}, C_1, C_2, f)$, let us define X^Q to be the discrete time Markov chain obtained by sampling X every Q time units. Finally, let $\mathcal{B}^Q = \{\beta_{ij}^Q : 1 \leq i, j \leq N\}$. The following Proposition establishes that the coefficients β_{ij}^Q describe the structure of the sampled Markov chain X^Q.

Proposition 2.5: For any $\mathcal{A}$, C_1, C_2, $f \in \mathcal{U}$, and for any $Q \in \mathcal{N}$, there exist some positive C_1', C_2' and some $f' \in \mathcal{U}$ such that $\{X^Q : X \in \mathcal{M}_\epsilon(\mathcal{A}, C_1, C_2, f)\}$ is a subset of $\mathcal{M}_\epsilon(\mathcal{B}^Q, C_1', C_2', f')$.

Proof: The result is immediate from the fact that $P(x^Q(t+1) = j \,|\, x^Q(t) = i)$ equals to the sum, over all paths from i to j with exactly Q hops, of the probability that $x(t)$ follows any such path. The desired conclusion follows with $C_1' = C_1^Q$, $C_2' = C_2^Q N^Q$ and $f'(\epsilon) = N^Q f(\epsilon)$. (The factor N^Q arises because there are less than N^Q such paths.) •

The main reason, however, for introducing X^Q is the following.

Proposition 2.6: Suppose that $\mathcal{A}$ satisfies Assumption AP and that $Q \geq M + N$, where M is the constant in that assumption. Then $\mathcal{B}^Q$ satisfies Assumption TRI.

Proof: Let us fix some i, j, k and suppose that $i \in R$. Let n be such that $\beta_{jk}^n = \beta_{jk}$. Without loss of generality we may assume that $n \leq N$. Furthermore, $\beta_{jk}^Q \geq \beta_{jk} = \beta_{jk}^n$. Therefore, using Assumption AP and the inequalities $Q \geq M$, $Q - n \geq M$, we have $\beta_{ij}^Q + \beta_{jk}^Q \geq \beta_{ij}^Q + \beta_{jk}^n = \beta_{ij}^{Q-n} + \beta_{jk}^n \geq \beta_{ik}^Q$, which is the desired inequality. The proof for the cases $j \in R$ or $k \in R$ is identical and is omitted. •

As a consequence of Propositions 2.5 and 2.6, Proposition 2.3 becomes applicable to an appropriately sampled version of a given Markov chain, assuming condition AP. We notice that Proposition 2.3 will provide us with estimates of the transition probabilities only for those times which are integer multiples of Q. However, it is easy to show that the same estimates are also valid for intermediate times as well.

With a more elaborate choice of Q, the conclusions of Proposition 2.6 are valid even if Assumption AP fails to hold. However, in this case, the corresponding conclusions of Proposition 2.3 will only

be valid for the sampled chain X^Q and not, in general, for the original Markov chain. Rather, the order of magnitude of $p_{ij}(t)$ will vary periodically with t.

We close this section by pointing out that there is nothing special about the coefficients α_{ij} being integer. For example, if the α_{ij} are rationals we could introduce another small parameter δ (to replace ϵ) and another set of integer coefficients β_{ij}, so that $\delta^{\beta_{ij}} = \epsilon^{\alpha_{ij}}$. Even if the α_{ij}'s are not rational, neither are their ratios rational, the proof of Proposition 2.3 remains valid, as long as $\min_{i,j}\{\alpha_{ij}\} \geq 1$. This can be always achieved by redefining the small parameter ϵ.

III. DETERMINING THE STRUCTURE AT SUCCESSIVELY SLOWER TIME SCALES.

Proposition 2.3 allows us to determine the structure of a Markov chain $X \in M_\epsilon$ in the first of the slow time scales, that is for times of the order of $1/\epsilon$. We now notice that the transition probabilities $P(x(1/\epsilon) = j \,|\, x(0) = i)$ satisfy (2.1), (2.2), (with a new choice of C_1, C_2, f) provided that we replace α_{ij} by $V(i,j)$. Moreover, due to part (iii) of Proposition 2.2, the coefficients $V(i,j)$ satisfy the triangle inequality (2.3) and, therefore, Proposition 2.3 becomes applicable once more. This yields estimates for the transition probabilities $P(x(1/\epsilon^2) = j | x(0) = i)$. This procedure may be repeated to yield estimates for $P(x(1/\epsilon^d) = j | x(0) = i)$, for any positive integer d. To summarize, we have the following algorithm:

Algorithm II: (Input: $\mathcal{A} = \{\alpha_{ij} : 1 \le i,j \le N\}$, satisfying Assumption TRI; Output: for each $d \in \mathcal{N}_0$, a collection $V^d = \{V^d(i,j) : 1 \le i,j \le N\}$, a subset R^d of the state space and for each $i \in R^d$ a set $R_i^d \subset R^d$.)

1. Let $V^0(i,j) = \alpha_{ij}$, $\forall i,j$.
2. Let V^d be the input to Algorithm I; let V^{d+1}, R^d, TR^d, R_i^d be the outputs returned by Algorithm I.

Notice that R^d is the set of all states such that $V^d(i,j) = 0$ implies $V^d(j,i) = 0$. Also, for any $i \in R^d$, $R_i^d = \{j \in R^d : V^d(i,j) = 0\}$. The remarks preceding Algorithm II establish the the next proposition. (Notice that when we use Proposition 2.3 to obtain estimates for $t \approx 1/\epsilon^d$, the unit of time becomes $1/\epsilon^{d-1}$. For this reason, the variable t in Proposition 2.3 must be replaced by $t\epsilon^{d-1}$.)

Proposition 3.1: Given some $\mathcal{A}$ satisfying Assumption TRI and some $d \in \mathcal{N}$, let $V^d(i,j)$, R^d, be the collection of integers and the subset returned by Algorithm II. Then, for any positive constants C_1, C_2, C_3 and for any $f \in \mathcal{U}$, there exist positive constants D_1, D_2, D_3, $D_4 < 1$ and $g \in \mathcal{U}$, such that, for any $\epsilon > 0$ and for any Markov chain $X \in M_\epsilon(\mathcal{A}, C_1, C_2, f)$ we have

$$D_1(\epsilon(\epsilon^{d-1}t - N))^N \epsilon^{V^d(i,j)} \le P(x(t) = j \,|\, x(0) = i) \le D_2\epsilon^{V^d(i,j)} + \chi_i D_3 D_4^{t\epsilon^{d-1}} \epsilon^{V^{d-1}(i,j)} + g(\epsilon),$$

$$\forall t \in [N/\epsilon^{d-1}, C_3/\epsilon^d], \tag{3.1}$$

where $\chi_i = 0$, if $i \in R^{d-1}$ and $\chi_i = 1$, otherwise. (The upper bound in (3.1) is also valid for $t \in [1/\epsilon^{d-1}, N/\epsilon^{d-1}]$.) In particular, there exist $D_1, D_2 > 0$, $g \in \mathcal{U}$ such that

$$D_1\epsilon^{V^d(i,j)} \le p_{ij}(\frac{1}{\epsilon^d}) \le D_2\epsilon^{V^d(i,j)} + g(\epsilon). \tag{3.2}$$

We continue with a few remarks on the quantities computed by Algorithm II.

Proposition 3.2: (i) For any d, i, j, k, we have $V^d(i,j) \le V^d(i,k) + +V^d(k,j)$.

(ii) For any d, we have $R^{d+1} \subset R^d$.

(iii) $V^d(i,j) + V^c(j,k) \ge V^{\max\{c,d\}}(i,k)$, $\forall i,j,k,c,d$.

Proof: (i) This is an immediate consequence of part (iii) of Proposition 2.2.

(ii) Suppose that $i \in R^{d+1}$. Then, $V^{d+1}(i,i) = 0$. Using part (ii) of Proposition 2.2, we conclude that $i \notin TR^d$, or, equivalently, $i \in R^d$.
(iii) Using Proposition 3.1 twice, there exist constants D_1, D_2 such that

$$D_1\epsilon^{V^d(i,j)+V^c(j,k)} \leq P\left(x(\frac{1}{\epsilon^d}+\frac{1}{\epsilon^c}) = k \mid x(0) = i\right) \leq D_2\epsilon^{V^{\max\{c,d\}}}(i,k).$$

Moreover, this inequality is true for all $X \in \mathcal{M}_\epsilon$ and for all $\epsilon > 0$. Letting ϵ be arbitrarily small, we conclude that the claimed result holds. •

As a corollary of Proposition 3.2 we conclude that some of the upper bounds of Proposition 3.1 are true even for times smaller than $1/\epsilon^{d-1}$.
Corollary 3.1: If $i \in R^d$, or if $j \in R^d$, or if $V^d(i,j) \leq V^c(i,j)$, $\forall c \leq d$, then there exists some $C > 0$ such that

$$p_{ij}(t) \leq C\epsilon^{V^d(i,j)}, \quad \forall t \in [0, 1/\epsilon^d],\ \forall X \in \mathcal{M}_\epsilon, \forall \epsilon > 0. \tag{3.3}$$

Proof: If $i \in R^d$, then $V^d(i,i) = 0$. For any $c \leq d$, and for any j, we may apply part (iii) of Proposition 3.2 to obtain $V^d(i,j) \leq V^d(i,i) + V^c(i,j) = V^c(i,j)$. A similar argument leads to the same conclusion if $j \in R^d$. Now, given some $t \leq 1/\epsilon^d$, find some integer c such that $t \in [1/\epsilon^{c-1}, 1/\epsilon^c]$. We then use Proposition 3.1 to obtain $p_{ij}(t) \leq D\epsilon^{V^c(i,j)} \leq De^{V^d(i,j)}$. •

Inequality (3.3) is in general false if its assumption fails to hold. We continue with a few remarks on the applicability and usefulness of Algorithms I and II.

Looking back at Algorithm I, we see that in order to determine $V(i,j)$ for $i \in R$ and $j \in R$, we only need to know the coefficients α_{ij} for i and j belonging to R. This has the following implication for Algorithm II: in order to compute the coefficients $\{V^{d+1}(i,j) : i,j \in R^d\}$, we only need to know the coefficients $\{V^d(i,j) : i,j \in R^d\}$. Since $R^{d+1} \subset R^d$, it follows that the coefficients $\{V^{d+1}(i,j) : i,j \in R^{d+1}\}$ may be computed from the coefficients $\{V^d(i,j) : i,j \in R^d\}$. Thus, if we are only interested in determining which states are recurrent for each time scale (as well as in determining the corresponding ergodic decomposition) we may eliminate, at each stage of Algorithm II, the states which have been found to be transient, that is the elements of TR^d. This observation, together with the fact that we only need to carry out the algorithm for just one representative from each class R_i^d, should result in a substantial amount of savings, were the algorithm to be implemented.

Naturally, Algorithm II is applicable to the appropriately sampled versions of Markov chains satisfying Assumption AP. Then, inequalities (3.1) and (3.2) are valid for times which are integer multiples of the sampling period Q. Moreover, a simple argument shows that these inequalities are valid for intermediate times as well. Also notice that sampling needs be carried out only once. Even if $\mathcal{A}$ satisfies AP but not TRI, still the coefficients V^d, $d > 0$, will automatically satisfy Assumption TRI and no sampling is required at subsequent stages of Algorithm II.

We compare Algorithm II and Proposition 3.1 to the results available in the literature. There has been a substantial amount of research on singularly perturbed stationary Markov chains [1,2,3,4,12].

Typical results obtain exact asymptotic expressions for the transition probabilities, as a small parameter ϵ converges to zero. These asymptotic expressions are obtained recursively, by proceeding from one time scale to the next, similarly with Algorithm II. Each step in this recursion involves the solution of systems of linear equations and, possibly, the evaluation of the pseudoinverse of some matrices [1], which may be computationally demanding, especially if we are dealing with large scale systems. However, we may conceive of situations in which we are not so much interested in knowing the values of the transition probabilities, but rather we want to know which events are likely to occur (over a certain time interval) and which events have asymptotically negligible probability (as ϵ goes to zero). For the latter case, a non–numerical, graph–theoretic, method is more natural. Such a method (for stationary Markov chains) is implicit and easy to extract from the results of [12]. Algorithm II also accomplishes the same.

On the more technical side, it does not follow from the literature, neither is it a priori obvious, that there exist integer coefficients $V^d(i,j)$ such that inequalities of the type (3.1) hold. The existing results provide approximations for those transition probabilities which do not vanish as ϵ approaches zero [1,2,3,4,12], but much less is known about the asymptotic behavior of the vanishing transition probabilities. Furthermore, the techniques which are usually employed are tailored to stationary Markov chains (e.g. perturbation theory of linear operators) and do not seem applicable to the analysis of non–stationary chains. The discussion following Proposition 2.1 suggests one method for applying results for stationary chains to non–stationary ones but it does not seem to be universally applicable. Let us also point out that Proposition 3.1 is fairly easy to derive for "nearly decomposable" Markov chains [3]. This is not the case for more general Markov chains; in particular, the existence of transient states which feed into different ergodic classes are the main source of difficulty [12].

IV. COOLING SCHEDULES FOR SIMULATED ANNEALING.

In simulated annealing [6,10] we are given a set $S = \{1,...,N\}$ of states together with a cost function $J : S \mapsto \mathcal{N}$ to be minimized. (Our restriction that J takes integer values can be relaxed.) The algorithm jumps randomly from one state to another and forms a Markov chain with the following transition probabilities:

$$P(x(t+1) = j \,|\, x(t) = i) = Q(i,j) \exp\left[\min\{0, -(J(j) - J(i))/T(t)\}\right], \qquad \text{if } j \neq i, \tag{4.1}$$

$$P(x(t+1) = i \,|\, x(t) = i) = 1 - \sum_{j \neq i} P(x(t+1) = j \,|\, x(t) = i), \tag{4.2}$$

where the kernel $Q(i,j)$ is nonnegative and satisfies $\sum_j Q(i,j) = 1$ and $T(t) > 0$ is the "temperature" at time t. It is known that if $T(t)$ decreases to zero slowly enough, then $x(t)$ converges (in probability) to the set at which J is minimized [5–9,11]. We are interested in determining how slowly $T(t)$ must converge to zero, so that convergence to the minimizing states is obtained. This issue has been resolved by Hajek [9] under some restrictions on the structure of $Q(i,j)$. We shall derive shortly the answer to this question in a more general setting. Moreover our method establishes a connection between simulated annealing and the structure of singularly perturbed stationary Markov chains.

We formulate the problem to be studied in a slightly more general manner, as follows. Suppose that we are given, a stochastic matrix P^ϵ, (whose ij-th entry is denoted by p^ϵ_{ij}) parameterized by a positive parameter ϵ and assume that there exist positive constants C_1, C_2 and a collection $\mathcal{A} = \{\alpha_{ij} : 1 \leq i,j \leq N\}$ such that $\alpha_{ij} \in \mathcal{N}_0 \cup \{\infty\}$, $\forall i,j$ and such that $p^\epsilon_{ij} = 0$, whenever $\alpha_{ij} = \infty$, and $C_1\epsilon^{\alpha_{ij}} \leq p^\epsilon_{ij} \leq C_2\epsilon^{\alpha_{ij}}$, $\forall \epsilon \in (0,1]$, whenever $\alpha_{ij} < \infty$. Finally, we are given a monotonically nonincreasing function (cooling schedule) $\epsilon : \mathcal{N}_0 \mapsto (0,1)$. We are interested in the Markov chain $x(t)$ with transition probabilities given by $P(x(t+1) = j \,|\, x(t) = i) = p^{\epsilon(t)}_{ij}$.

Clearly, the simulated annealing algorithm is of the type described in the preceding paragraph, provided that we identify $\epsilon(t)$ with $e^{-1/T(t)}$ and provided that we define $\alpha_{ij} = \infty$, if $Q(i,j) = 0$, $i \neq j$, and $\alpha_{ij} = \max\{0, J(j) - J(i)\}$, if $Q(i,j) \neq 0$, $i \neq j$. Also, α_{ii} has to be accordingly defined.

We now return to our general formulation. We thus assume that $\mathcal{A}$, C_1, C_2 are given, together with the schedule $\{\epsilon(t)\}$. We assume that $\mathcal{A}$ satisfies Assumption TRI and we define, for any $d \in \mathcal{N}_0$, the quantities $V^d(i,j)$ and the sets R^d by means of Algorithm II of Section III. Our main result is the following.

Proposition 4.1: Assume that for some integer $d \geq 0$,

$$\sum_{t=0}^{\infty} \epsilon^d(t) = \infty, \tag{4.3}$$

$$\sum_{t=0}^{\infty} \epsilon^{d+1}(t) < \infty. \tag{4.4}$$

Then,
(i) $\lim_{t\to\infty} P(x(t) \in R^d \,|\, x(0) = i) = 1, \quad \forall i.$
(ii) For any $i \in R^d$, $\limsup_{t\to\infty} P(x(t) = i \,|\, x(0) = i) > 0.$
Proof: The main idea of the proof is to partition $[0,\infty)$ into a set of disjoint time intervals $[t_k, t_{k+1})$ such that $x(t)$ is approximately stationary during each such interval, in the sense of Section II, and then use the estimates available for such Markov chains. To simplify notation, for any function defined on the integers, we extend it on the real line in a piecewise constant and right–continuous fashion. Thus, for example, $\epsilon(t) = \epsilon(n)$, $\forall t \in [n, n+1)$, $n \in \mathcal{N}$.

The proof for the case $d = 0$ is rather easy and is omitted. We present the comparatively harder proof for the case $d \geq 1$.

We start with the proof of part (i) of the proposition. We define $t_0 = 0$ and

$$t_{k+1} = t_k + \frac{1}{\epsilon^{d-1}(t_k)}, \quad \text{if } \epsilon\left(t_k + \frac{1}{\epsilon^{d-1}(t_k)}\right) \geq \frac{1}{2}\epsilon(t_k), \tag{4.5}$$

$$t_{k+1} = \min\{t : \ \epsilon(t) \leq \frac{1}{2}\epsilon(t_k)\}, \quad \text{otherwise.} \tag{4.6}$$

We define A_L (respectively, A_S) as the set of all k's such that t_{k+1} is defined by (4.5) (respectively, (4.6)). We will need the following properties of the sequence $\{\epsilon(t_k)\}$.
Lemma 4.1:

$$\frac{1}{2}\epsilon(t_k) \ \leq \ \epsilon(t) \ \leq \ \epsilon(t_k), \quad \forall t \in [t_k, t_{k+1}) \tag{4.7},$$

$$\sum_{k\in A_L} \epsilon(t_k) = \infty, \tag{4.8}$$

$$\sum_{k=0}^{\infty} \epsilon^2(t_k) < \infty \tag{4.9}$$

Let $f(k,l)$ be the cardinality of $A_L \cap \{l, ..., k-1\}$, for $k \geq l$. Then, for any $C \in (0,1)$,

$$\sum_{k=0}^{\infty}\sum_{l=0}^{k}(1-C)^{f(k,l)}\epsilon(t_k)\epsilon(t_l) \ < \infty \tag{4.10}$$

$$\lim_{k\to\infty}\sum_{l=0}^{k}(1-C)^{f(k,l)}\epsilon(t_l) = 0, \quad \forall c \in (0,1). \tag{4.11}$$

Proof: Inequalities (4.7) are an immediate consequence of (4.5), (4.6).

We notice that for any $k \in A_S$, $k' \in A_S$, with $k' > k$, we have $\epsilon(t_{k'}) \leq (1/2)\epsilon(t_k)$. Hence,

$$\sum_{k\in A_S} \epsilon(t_k) \leq \epsilon(0) \sum_{k=0}^{\infty} 2^{-k} < \infty. \tag{4.12}$$

Finally,

$$\sum_{k\in A_L}\epsilon(t_k) \;=\; \sum_{k\in A_L}\epsilon^d(t_k)[t_{k+1}-t_k] \;=\; \sum_{k=0}^{\infty}\epsilon^d(t_k)[t_{k+1}-t_k] - \sum_{k\in A_S}\epsilon^d(t_k)[t_{k+1}-t_k] \geq$$

$$\sum_{t=0}^{\infty}\epsilon^d(t) - \sum_{k\in A_S}\epsilon(t_k) = \infty,$$

which proves (4.8).

From (4.12) we conclude that $\sum_{k\in A_S}\epsilon^2(t_k) < \infty$. Also,

$$\sum_{k\in A_L}\epsilon^2(t_k) \;=\; \sum_{k=0}^{\infty}\epsilon^{d+1}(t_k)[t_{k+1}-t_k] \;\leq 2^{d+1}\sum_{t=0}^{\infty}\epsilon^{d+1}(t) < \infty,$$

which proves (4.9).

Given any $C \in (0,1)$, we define a constant a by $[2(1-C)]^a = 3/2$, if $2(1-C) \geq 1$; otherwise, we let $a = 1$. Let $B = \{(k,l) :\ k \geq l \text{ and } f(k,l) \geq a(k-l)\}$. Then,

$$\sum_{(k,l)\in B}(1-C)^{f(k,l)}\epsilon(t_k)\epsilon(t_l) \;\leq \sum_{k=0}^{\infty}\sum_{l=0}^{k}[(1-C)^a]^{k-l}\epsilon(t_k)\epsilon(t_l) < \infty,$$

because $(1-C)^a < 1$ and $\epsilon(t_k)$ is square summable by (4.9). Now notice that $\epsilon(t_k) \leq 2^{-(k-l)+f(k,l)}\epsilon(t_l)$, if $k \geq l$. Hence,

$$\sum_{(k,l)\notin B,\ k\geq l}(1-C)^{f(k,l)}\epsilon(t_k)\epsilon(t_l) \;\leq \sum_{(k,l)\notin B,\ k\geq l}[2(1-C)]^{f(k,l)}2^{-(k-l)}\epsilon^2(t_l) \;\leq$$

$$\sum_{k\geq l}(3/2)^{k-l}(1/2)^{k-l}\epsilon^2(t_l) \;< \infty,$$

which proves (4.10). The proof of (4.11) is similar and is omitted. •

We now define subsets $S_0, S_1, \ldots$ of the state space inductively as follows.

$$S_0 = R^d = \{i :\ \text{if } V^d(i,j) = 0 \text{ then } V^d(j,i) = 0\},$$

$$S_{n+1} = \{i \in R^{d-1} :\ i \notin S_0 \cup \ldots \cup S_n \text{ and } \exists j \in S_n \text{ such that } V^{d-1}(i,j) = 1\},$$

Also let

$$T_0 = \{i \in TR^{d-1} :\ \exists j \in S_0 \text{ such that } V^{d-1}(j,i) = 1\}$$

and let T_1 be the complement of T_0 in TR^{d-1}. Notice that $(\cup_{n\geq 0}S_n)\cup T_0 \cup T_1 = \{1,\ldots,N\}$. Also, if $i \in S_n$, $n \neq 0$ and $V^{d-1}(i,j) = 0$, then $j \in R_i^{d-1}$ and $j \in S_n$. (For a proof of this fact, if $i \in S_n$, then $i \in R^{d-1}$; so, if $V^{d-1}(i,j) = 0$, then $V^{d-1}(j,i) = 0$ and therefore $j \in R_i^{d-1}$. Let $l \in S_{n-1}$ be such that $V^{d-1}(i,l) = 1$. Then, $V^{d-1}(j,l) = 1$. So, either $j \in S_n$ and we are done, or

$j \in S_0 \cup ... \cup S_{n-1}$. In the second case, the same argument shows that $i \in S_0 \cup ... \cup S_{n-1}$ which is a contradiction.)

We let $y(k) = x(t_k)$. We need estimates on the transition probabilities of the $y(k)$ process. These are obtained by noting that, for any k, the Markov chain $\{x(t) : t \in [t_k, t_{k+1}]\}$ belongs to $M_{\epsilon(t_k)}(\mathcal{A}, 2^{-K}C_1, C_2, 0)$, where $K = \max\{\alpha_{ij} : \alpha_{ij} < \infty\}$. Since $t_{k+1} - t_k \leq 1/(\epsilon^{d-1}(t_k))$, Corollary 3.1 may be used to obtain upper bounds. Also, for $k \in A_L$, $t_{k+1} - t_k = 1/(\epsilon^{d-1}(t_k))$ and therefore Proposition 3.1 may be used to obtain lower bounds. In more detail, we have:

Lemma 4.2: There are constants $F > 0$, $G > 0$, such that, for every $k \in \mathcal{N}_0$ we have

(i) If $k \in A_L$, then $P(y(k+1) \in S_n \,|\, y(k) \in S_{n+1}) \geq F\epsilon(t_k)$, $\forall n$. (4.13)

(ii) $P(y(k+1) \notin S_n \,|\, y(k) \in S_n) \leq G\epsilon(t_k)$, $\forall n$. (4.14)

(iii) $P(y(k+1) \notin S_0 \cup T_0 \,|\, y(k) \in S_0) \leq G\epsilon^2(t_k)$. (4.15)

(iv) $P(y(k+1) \notin S_0 \cup T_0 \,|\, y(k) \in T_0) \leq G\epsilon(t_k)$. (4.16)

(v) $P(y(k+1) \in T_0 \,|\, y(k) \in S_0) \leq G\epsilon(t_k)$. (4.17)

(vi) If $k \in A_L$, then $P(y(k+1) \in S_0 \,|\, y(k) \in T_0) \geq F$. (4.18)

(vii) If $k \in A_L$, then, for all i, $P(y(k+1) \in TR^{d-1} \,|\, y(k) = i) \leq 1 - F$. (4.19)

Proof: (i) If $i \in S_{n+1}$, then (by definition) there is some $j \in S_n$ such that $V^{d-1}(i,j) = 1$. The result follows from the lower bound in (3.2).

(ii) Let $i \in S_n$, $j \notin S_n$. We have shown earlier that we must have $V^{d-1}(i,j) \geq 1$ and the result follows from (3.3).

(iii) Let $i \in S_0$ and $j \notin S_0 \cup T_0$. If $j \in S_n$, $n \neq 0$, then $j \notin R^d$; hence $V^d(i,j) \geq 1$. Therefore, using the definition of V^d, we have $1 \leq V^d(i,j) \leq V^d(i,i) + V^{d-1}(i,j) - 1 = V^{d-1}(i,j) - 1$. Hence $V^{d-1}(i,j) \geq 2$. Finally, if $j \in T_1$, then $V^{d-1}(i,j) \geq 2$, because otherwise we would have $j \in T_0$. The result follows from (3.3).

(iv) Let $i \in T_0$ and choose some $l \in S_0$ such that $V^{d-1}(l,i) = 1$ (which exists by the definition of T_0). Suppose that $j \notin S_0 \cup T_0$. If $j \in S_n$, $n \neq 0$, then $V^{d-1}(i,j) \geq 1$, because otherwise $V^{d-1}(l,j) = 1$, which contradicts the discussion in the proof of part (iii). So, for this case the result follows from (3.3). Suppose now that $j \in T_1$. For any $c \leq d-1$ we must have $V^c(i,j) \geq 1$ because otherwise (using Proposition 3.2) $V^{d-1}(l,j) \leq V^{d-1}(l,i) + V^c(i,j) = 1$, which contradicts the assumption $j \in T_1$. The result follows again from (3.3).

(v) This is immediate from $V^{d-1}(i,j) \geq 1$, $\forall i \in R^{d-1}$, $\forall j \in TR^{d-1}$ (Proposition 2.2, part (ii)).

(vi) Let $i \in T_0$. Since $i \in TR^{d-1}$, there exists some $j \in R^{d-1}$ such that $V^{d-1}(i,j) = 0$. By the

previous discussion, such a j cannot belong to S_n, for $n \geq 1$. The result follows from (3.2).
(vii) Similarly, for any i there exists some $j \in R^{d-1}$ such that $V^{d-1}(i,j) = 0$ and the result follows from (3.2). •

Let

$$H_k = P(y(n) \in S_0 \cup T_0,\ \forall n \in [0,k] \,|\, y(0) \in S_0),$$

$$Q_k = P(y(k) \in T_0 \,|\, y(n) \in S_0 \cup T_0,\ \forall n \in (0,k-1),\ y(0) \in S_0).$$

Using (4.17), (4.18), we obtain

$$Q_{k+1} \leq G\epsilon(t_k) + (1 - \chi_k F)Q_k,$$

where $\chi_k = 1$ if $k \in A_L$ and $\chi_k = 0$, otherwise. So,

$$Q_k \ \leq \ G\sum_{l=0}^{k} \epsilon(t_l)(1-F)^{f(k,l)}.$$

Using (4.15), (4.16),

$$H_{k+1} \ \geq \ [1 - G\epsilon(t_k)Q_k - G\epsilon^2(t_k)]H_k \tag{4.20}$$

Now, $\epsilon(t_k)Q_k$ is summable, by (4.10); also, $\epsilon^2(t_k)$ is summable, by (4.9). Hence $\inf_{k\to\infty} H_k > 0$. More intuitively, once the state enters S_0, there is positive probability that it never leaves $S_0 \cup T_0$. Consequently, the total flow of probability into S_0 from S_1 must be finite. Hence, using (4.13), we have

$$\sum_{k=0}^{\infty} \epsilon(t_k)P(y(k) \in S_1) < \infty.$$

We will prove by induction that for all $n \geq 1$,

$$\sum_{k=0}^{\infty} \epsilon(t_k)P(y(k) \in S_n) < \infty. \tag{4.21}$$

Using (4.13), (4.14), we have

$$P(y(k+1) \in S_n) \ \geq \ P(y(k) \in S_n) - G\epsilon(t_k)P(y(k) \in S_n) + \chi_k F\epsilon(t_k)P(y(k) \in S_{n+1}). \tag{4.22}$$

By telescoping the inequality (4.22) and using the induction hypothesis (4.21), we see that $\sum_{k=0}^{\infty} \chi_k\epsilon(t_k)P(y(k) \in S_{n+1}) < \infty$. Also, $\sum_{k\in A_S} \epsilon(t_k)P(y(k) \in S_n + 1) \leq \sum_{k\in A_S} \epsilon(t_k) < \infty$ (because of (4.12)) which completes the induction step. Using (4.21) and the fact that $\epsilon(t_k)$ sums to infinity we conclude that $\limsup_{k\to\infty} P(y(k) \in S_0 \cup TR^{d-1}) = 1$. We show next that the probability of transient states goes to zero. Inequalities (4.14) and (4.19) imply

$$P(y(k+1) \in TR^{d-1}) \leq G\epsilon(t_k) + (1 - \chi_k F)P(y(k) \in TR^{d-1}).$$

Thus,

$$P(y(k) \in TR^{d-1}) \leq (1-F)^{f(k,0)} + G\sum_{l=0}^{k}(1-F)^{f(k,l)}\epsilon(t_l),$$

which converges to 0, as k tends to infinity, due to (4.11). We may thus conclude that $\limsup_{k\to\infty} P(y(k) \in S_0) = 1$. By repeating the argument that led to (4.20) we can see that the probability that y ever exits $S_0 \cup T_0$, given that $y(k) \in S_0$, converges to zero, as $k \to \infty$. (This is a consequence of the square summability of $\epsilon(t_k)$.) It follows that $\lim_{k\to\infty} P(y(k) \in S_0) = 1$. Finally, for any $t \in [t_k, t_{k+1}]$ we have $P(x(t) \in S_0) \geq P(y(k) \in S_0) - G\epsilon(t_k)$, which converges to 1, as $k \to \infty$. This completes the proof of part (i) of the proposition.

For part (ii) of the proposition, in order to avoid introducing new notation, we prove the equivalent statement that if $\sum_{t=0}^{\infty}\epsilon^d(t) < \infty$, then $\limsup_{t\to\infty} P(x(t) = i \,|\, x(0) = i) > 0$, $\forall i \in R^{d-1}$. So, let $i \in R^{d-1}$ and consider the set R_i^{d-1}. For any $j \notin R_i^{d-1}$, we have $V^{d-1}(i,j) \geq 1$ and, therefore, (using Corollary 3.1), there exists some $G > 0$ such that

$$P(y(k+1) \notin R_i^{d-1} \,|\, y(k) \in R_i^{d-1}) \;\leq\; G\epsilon(t_k), \quad \forall k.$$

Since we are assuming that $\sum_{t=0}^{\infty}\epsilon^d(t) < \infty$, it follows (as in the proof of (4.9)), that $\sum_{k=0}^{\infty}\epsilon(t_k) < \infty$. Consequently,

$$\inf_k P(y(k) \in R_i^{d-1} \,|\, y(0) = i) > 0. \tag{4.23}$$

Finally, for any $j \in R_i^{d-1}$ we have $V^{d-1}(j,i) = 0$. Hence, using Proposition 3.1, there exists some $F > 0$ such that

$$P(y(k+1) = i \,|\, y(k) \in R_i^{d-1}) \;\geq\; F. \tag{4.24}$$

By combining (4.23), (4.24), we obtain the desired result. •

Remarks: 1. With a little more effort along the lines of (4.23), (4.24) it can be shown that, under (4.4), we have $\inf_t P(x(t) = i \,|\, x(0) = i) > 0$, $\forall i \in R^d$.

2. Proposition 4.2 may be extended to the case where $\mathcal{A}$ satisfies Assumption AP in a straightforward manner, using the method discussed in Section 2.

Corollary 4.1: Let the transition probabilities for the simulated annealing algorithm be given by (4.1), (4.2). Consider cooling schedules of the form $T(t) = c/\log t$. For any initial state, the algorithm converges (in probability) to the set of global minima of J if and only if there exists some d such that the set of global minima contains R^d and c is larger than or equal to the smallest such d, to be denoted by d^*.

Proof: Having identified $\exp[-1/T(t)]$ with $\epsilon(t)$, we see that $\sum_{t=1}^{\infty}\epsilon^c(t) = \sum_{t=1}^{\infty}\frac{1}{t} = \infty$ and $\sum_{t=1}^{\infty}\epsilon^{c+1}(t) = \sum_{t=1}^{\infty}(1/t^{(c+1)/c}) < \infty$. Thus, by Proposition 4.1, R^c is the smallest set to which the algorithm converges (in probability). Thus, for convergence to the set of global minima, we need that set to contain R^c, which establishes the desired result. •

A possibility for generalizing Proposition 4.1 arises if we allow the schedule $\epsilon(t)$ to be non-monotonic. In fact the proof goes through (with a minor modification in the definition of the sequence $\{t_k\}$) if we assume that there exists some $C > 0$ such that $\epsilon(t) \leq C\epsilon(s)$, $\forall t \geq s$, which allows for mild non-monotinicity. On the other hand, if $\epsilon(t)$ is allowed to have more substantial variations, then the conclusions of Proposition 4.1 are no more true. For a simple example consider the Markov chain of Figure 1, together with the schedule $\epsilon(t) = t^{-1/2}$, if t is even, and $\epsilon(t) = 1/t$, if t is odd. For this schedule, the largest integer for which $\sum_{t=0}^{\infty} \epsilon^d(t) = \infty$ is equal to 2. Also, $R^2 = \{3\}$. On the other hand, $P(x(t) = 3 \,|\, x(0) = 1)$ does not converge to 1.

We have claimed that our result generalizes the results of [9] and we conclude this section by supporting this claim. Hajek's result characterized d^* in an explicit manner, as the maximum depth[1] of local minima which are not global minima, under a "weak reversibility" assumption, which is equivalent to imposing certain restrictions on the structure $\mathcal{A}$. Our characterization is less explicit because instead of describing d^* we give an algorithm for computing it in terms of $\mathcal{A}$. Nevertheless, for the class of structures $\mathcal{A}$ considered in [9], we can use our Algorithm II to show that R^d is the set of all local minima of the cost function J, of depth $d+1$, or more. Hence, the d^* produced by our approach is the smallest d such that all local (but not global) minima have depth d or less, which agrees with the result of [9]. We do not present the details of this argument since it would amount to rederiving a known result.

1. The depth of a state i is defined as the minimum over all j, such that $J(j) < J(i)$, of the minimum over all paths leading from i to j, of the maximum of $J(k) - J(i)$, over all k's belonging to that path; the depth of i is infinite if no such j exists.

V. THE CONTINUOUS TIME CASE.

Algorithm II and the results of Section III are also applicable to continuous time Markov chains. For example, let there be given a stationary (for simplicity) Markov chain whose generator is a polynomial in ϵ and where ϵ is an unspecified positive parameter. Then, the transition probabilities, over a time interval of unit duration, satisfy inequalities (2.1), (2.2) for a suitable choice of α_{ij}. (In fact, the α_{ij}'s may be read off the Taylor series expansion of e^{A_ϵ}, or, equivalently, by solving a shortest path problem; the details are omitted.) Moreover, it can be shown that these coefficients automatically satisfy inequality (2.3), for all i, j, k. Therefore, Propositions 2.3 and 3.1 may be applied to the discrete time Markov chain obtained by sampling the continuous time Markov chain at integer times. Then, an elementary argument shows that the estimates obtained are valid for non–integer times as well.

Suppose now that the continuous time Markov chain is non–stationary and its generator is given by $A_{\epsilon(t)}$, where A_ϵ is as above and where $\epsilon(t)$ is some positive function of time. If $\epsilon(t)$ does not vary by more than a constant factor during time intervals of unit duration, then the unit time transition probabilities will again satisfy estimates of the form $C_1\epsilon^{\alpha_{ij}}(t) \leq P(x(t+1) = j \mid x(t) = i) \leq C_2\epsilon^{\alpha_{ij}}(t)$, with the same coefficients α_{ij} as in the previous paragraph. Then Proposition 4.1 may be applied to the sampled Markov chain to characterize optimal cooling schedules for continuous time simulated annealing algorithms.

VI. REFERENCES

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

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

3. Courtois, P.J., *Decomposability: Queuing and Computer System Applications,* Academic Press, New York, 1977.

4. Delebecque, F., "A Reduction Process for Perturbed Markov Chains", *SIAM J. Applied Math.,* 43, 2, 1983.

5. Gelfand, S.B., Mitter, S.K., "Analysis of Simulated Annealing for Optimization", preprint, Dept. of Electrical Engineering, M.I.T., 1985.

6. 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.

7. Geman, S., Hwang, C.-R., "Difusions for Global Optimization," preprint, Division of Applied Mathematics, Brown University, 1984.

8. Gidas, B., "Non–Stationary Markov Chains and Convergence of the Simulated Annealing Algorithm", preprint, Dept. of Mathematics, Rutgers University, 1984.

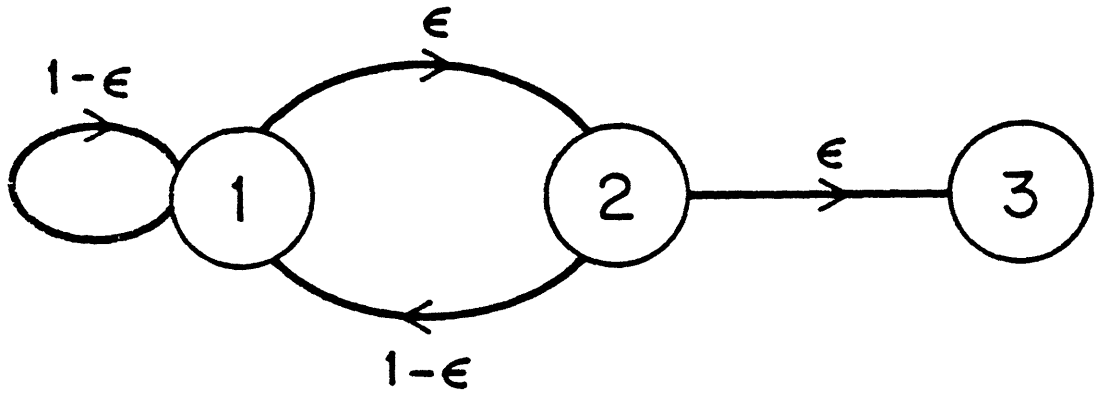
9. Hajek, B., "Cooling Schedules for Optimal Annealing", preprint, Dept. of Electrical Engineering, University of Illinois at Urbana–Champaign, 1985.

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

11. Mitra, D., Romeo, F., Sangiovanni–Vincentelli, A., "Convergence and Finite–Time Behavior of Simulated Annealing", preprint, Dept. of Electrical Engineering, University of California at Berkeley, 1985.

12. Rohlicek, J.R., and Willsky, A.S., "The Reduction of Perturbed Markov Generators: An Algorithm Exposing the Role of Transient States", Report LIDS–P–1493, LIDS, MIT, Cambridge, MA, 1985.

13. Varga, R.S., *Matrix Iterative Analysis,* Prentice Hall, Englewood Cliffs, NJ, 1962.


1-ε
ε
1
2
ε
3
1-ε