J. Appl. Prob. **24**, 657–667 (1987)
Printed in Israel

SIMULATED ANNEALING METHODS WITH GENERAL ACCEPTANCE PROBABILITIES

S. ANILY,* *The University of British Columbia*
A. FEDERGRUEN,** *Columbia University*

Abstract

Heuristic solution methods for combinatorial optimization problems are often based on local neighborhood searches. These tend to get trapped in a local optimum and the final result is often heavily dependent on the starting solution. *Simulated* annealing methods attempt to avoid these problems by *randomizing* the procedure so as to allow for occasional changes that worsen the solution. In this paper we provide probabilistic analyses of different designs of these methods.

PROBABILISTIC PERFORMANCE ANALYSIS; CONVERGENCE CONDITIONS; GENERAL ACCEPTANCE PROBABILITIES

1. Introduction and summary

> 'It is a sobering thought that the only way to solve many engineering problems is still by trial and error' (from 'Problem solving: smart guess work,' *The Economist*, 28 July 1984).

Heuristic solution methods for combinatorial optimization problems are often based on local neighborhood searches. Each solution is associated with a given collection of neighbors (the *neighborhood*). At each iteration, the current solution is replaced by one of its improving neighbors provided the latter exist. Otherwise, the algorithm terminates with the current solution. These deterministic search procedures, while generating monotonically improving sequences of solutions, encounter the following probems:

(i) the final solution is heavily dependent on the starting point;

(ii) deterministic methods tend to get trapped in local optima.

Simulated annealing methods attempt to avoid these problems by *randomizing* the procedure so as to allow for occasional changes that worsen the solution: a potential switch is implemented with a prespecified *acceptance*

Received 18 June 1985; revision received 6 June 1986.

* Postal address: Faculty of Commerce and Business Administration, The University of British Columbia, Vancouver, B.C., Canada V6T 1Y8.

** Postal address: Graduate School of Business, Uris Hall, Columbia University, New York, NY 10027, USA.

probability. All acceptance probabilities depend on a control parameter c which is reduced to 0 as the algorithm progresses. As c decreases to 0, the acceptance probabilities for deteriorating (improving) switches converge to 0 (1) according to prespecified *acceptance probability functions* (a.p.f.).

It appears that the annealing concept was first developed in statistical mechanics, motivated by an analogy to the behavior of physical systems in the presence of a heat bath, see Metropolis et al. (1953). (The control parameter c plays the role of the temperature in statistical mechanics.) Recently Kirkpatrick et al. (1983) and Cerny (1985) introduced the concept as an innovative and general solution approach for discrete optimization problems. Their observations, reinforced by articles in the popular press (e.g., Wilson et al. (1982)) led to several successful applications in a variety of problem areas, e.g. the traveling salesman problem, VLSI design, code generation, speech recognition. See Aragon et al. (1985) for a list of *empirical* studies.

This paper presents a *probabilistic* analysis of various non-static implementations when applied to a general discrete optimization problem. Let s_k be the solution generated, and c_k the value of the control parameter applied at the kth iteration. We identify necessary and sufficient conditions for the following properties:

(a) *Reachability of the set of global optima*. The set of global optima is reached from every starting solution with probability 1.

(b) *Asymptotic independence of starting solution*. The dependence of the distribution of s_k with respect to the starting solution vanishes as $k \to \infty$.

(c) *Convergence in distribution*. s_k converges in distribution.

(d) *Convergence to a global optimum*. The algorithm converges to the set of global optima with probability 1.

In addition, for annealing methods satisfying the third property we identify a bound on the *rate* of convergence.

For practitioners reachability is perhaps the most important of the four properties since it is easy to keep track of the best solution encountered in the course of the algorithm. The necessary and sufficient conditions for the four properties apply to general discrete problems, search heuristics and acceptance probability functions and are shown to be tight for the most commonly used a.p.f.'s. For example, for the most popular of all investigated a.p.f.'s ('exponential' or 'Metropolis' probabilities) these conditions imply the existence of two (problem-dependent) constants K_1 and K_2 such that all four properties hold if $c_k \geqq K_1/\log k$ for k sufficiently large while they all fail to hold in *any* problem with suboptimal local minima if $c_k \leqq K_2/\log k$ (for k sufficiently large). Many of the proposed annealing algorithms (see Aragon et al. (1985) and the references therein) thus fail to exhibit *any* of these properties for *any* problem with suboptimal local optima.

Lundy and Mees (1986) and Romeo and Sangiovanni-Vincentelli (1984) analyze *static* implementations (with exponential a.p.f.'s) where the control

parameter is kept constant throughout the algorithm. Sufficient conditions for (some of) the three convergence properties (b)–(d) were independently obtained by Geman and Geman (1984), Mitra et al. (1986), Hajek (1985) and Gidas (1985), all for the special case of exponential a.p.f.'s (Gidas (1985) obtains these conditions for a number of related a.p.f.'s, see below). Again for exponential a.p.f.'s, a necessary condition for property (d) was independently obtained by Hajek (1985) under two additional conditions which seem to be verifiable for *symmetric* neighborhood structures only, see below. Mitra et al. (1986) also obtain a characterization of the convergence rate which is similar to ours (for exponential a.p.f.'s).

Section 2 introduces the notation and some preliminary results and Section 3 contains all of the main results.

2. Notation and preliminaries

A combinatorial optimization problem may be viewed as the problem of minimizing a given function $f: X \to R$ with $X = \{1, \cdots, N\}$ its finite set of feasible solutions. For example, in the traveling salesman problem (TSP) with n cities, X consists of all feasible routes ($N = (n-1)!$) and f_i denotes the length of the ith tour. Assume the elements of X are numbered in ascending order of their objective function values $\{f_i, i = 1, \cdots, N\}$.

Iterative solution methods specify a topology or neighborhood structure on X; each soution i ($1 \leqq i \leqq N$) is uniquely associated with a neighborhood $X_i \subset X$ of all the solutions that can be reached from i in a single iteration. In a reasonable neighborhood structure, one of the global optima (say solution 1) can be reached from any other solution through neighbor switches. Thus, let n be an integer such that solution 1 *can* be reached from any solution through no more than n switches. For the K-median problem, with K a given positive integer, for example, the greedy add/interchange methods are among the most powerful heuristics, see Cornuejols et al. (1977). In one such method the neighborhood of a given set of medians is given by the collection of sets which may be obtained by adding a point outside of the current set or by substituting this point for one that currently is in the set. In this case n is given by the number of points and $|X_i| = O(Kn) \ll N = O(n^K)$ ($i \in X$). Note that the neighborhood structure is *not* symmetric, i.e. if $j \in X$ is a neighbor of $i \in X$, the converse may fail to hold.

The dynamics of a simulated annealing method are as follows: assume the current solution is i ($1 \leqq i \leqq N$). A specific neighbor $j \in X_i$ is generated (as a *potential* successor) with probability g_{ij}. (For the sake of convenience we assume $i \in X_i$ for all $i \in X$, i.e. $g_{ii} > 0$, $i \in X$.) Let $G = (g_{ij})$. The switch (between i and j) is implemented according to a positive acceptance probability a_{ij}; a_{ij} depends on a control parameter c (i.e. $a_{ij} \equiv a_{ij}(c)$) which is decreased to 0 in the course of the algorithm. The functions $a_{ij}(c)$ satisfy the

property

$$
(1)\qquad
\begin{aligned}
&a_{ij}(c)\downarrow 0 && \text{for } c \text{ sufficiently small,} && \text{if } f_j > f_i\\
&a_{ij}(c)\uparrow 1 && \text{for } c \text{ sufficiently small,} && \text{if } f_i > f_j\\
&\lim_{c\downarrow 0} a_{ij}(c) && \text{exists,} && \text{if } f_i = f_j.
\end{aligned}
$$

Let c_k be the value of the control parameter in the kth iteration. Observe that the sequence $\{s_k\}_{k=1}^{\infty}$ is generated by a non-stationary Markov chain with state space X and transition probability matrices $P(k) = (p_{ij}(c_k))$, $k \geqq 1$ where

$$
(2)\qquad p_{ij}(c) = \begin{cases} g_{ij}a_{ij}(c), & j \neq i,\\ 1 - \sum\limits_{l \in X_i \setminus \{i\}} g_{il}a_{il}(c), & j = i. \end{cases}
$$

Let $\pi(k)$ denote the unique steady-state vector associated with $P(k)$ and let $P(\infty) = \lim_{k\to\infty} P(k)$.

In addition to (1), the a.p.f.'s are assumed to satisfy certain mild regularity conditions guaranteeing that $\bar{\pi}(c)$, the steady-state distribution associated with the matrix $(p_{ij}(c))$ $(c > 0)$, is a (vector) function of bounded variation (see e.g. Royden (1968), p. 98). Anily and Federgruen (1985a), theorem 2, prove that the a.p.f.'s may for example be taken from the following classes of functions.

Definition 1. A class $F \subset C^1$ of functions defined on (0, 1] is a *closed class of asymptotically monotone functions* (CAM) if

(a) $f \in F \Rightarrow f' \in F$ and $-f \in F$,

(b) $f, g \in F \Rightarrow (f+g)$ and $(f \cdot g) \in F$,

(c) all $f \in F$ change signs finitely often on [0, 1].

Definition 2. A class F of functions defined on [0, 1] is a *rationally closed class of bounded variation* (RCBV) if

(a) $f \in F \Rightarrow f$ is of bounded variation on (0, 1],

(b) $f \in F \Rightarrow -f \in F$,

(c) $f, g \in F \Rightarrow (f+g)$ and $(f \cdot g) \in F$,

(d) $f, g \in F$ with f/g bounded on $(0, 1] \Rightarrow f/g$ is of bounded variation.

For example, rational functions of polynomials and exponential functions in c or c^{-1}, or even piecewise combinations thereof (splines) all fall in one or both of these classes, see Proposition 1 in Anily and Federgruen (1985a).

The properties of annealing methods thus follow from the behavior of the chain $\{P(k)\}_{k=1}^{\infty}$. For example, asymptotic independence of the starting solution and convergence in distribution (properties (b) and (c)) are equivalent to *weak* and *strong* ergodicity respectively: (let $P^{(m,k)} \equiv P(m) \cdots P(k)$; if $m > k$, $P^{(m,k)} \equiv I$.)

Definition 3 (see Isaacson and Madsen (1976)). $\{P(k)\}$ is *weakly* ergodic if $\lim_{k\to\infty} [P_{ij}^{(m,k)} - P_{lj}^{(m,k)}] = 0$ for all $i, l, j \in X$ and all $m \geqq 1$.

Definition 4 (see Isaacson and Madsen (1976)). $\{P(k)\}$ is *strongly* ergodic if a steady-state distribution π exists with $\lim_{k\to\infty} P_{ij}^{(m,k)} = \pi_j$ for all $m \geqq 1$ and all $i, j \in X$.

The analysis in Section 3 uses the *ergodic coefficient* of the matrices $P(k)$.

Definition 5 (see Dobrushin (1956) and Isaacson and Madsen (1976)). The ergodic coefficient of a stochastic matrix P is defined by:

$$\alpha(P) = \min_{i,l} \sum_{j} \min(P_{ij}, P_{lj}).$$

Theorem 1 below exhibits some important relationships between the properties (a)–(d).

Theorem 1.

(i) (d)$\Rightarrow$(c)$\Rightarrow$(b).

(ii) Assume property (b) holds. Then property (c) holds, $\lim_{k\to\infty} \pi(k)$ exists and $\lim_{k\to\infty} \pi(k) = \pi$ (see Definition 2).

(iii) Assume property (b) holds and $\limsup_{k\to\infty} \pi(k)_1 > 0$. Then property (a) holds.

Proof. Part (i) is straightforward. Part (ii) follows from Corollary 1 in Anily and Federgruen (1985a), in view of the assumed regularity conditions on the a.p.f.'s. Part (iii): in view of (ii), $\lim_{k\to\infty} P_{i1}^{(1,k)} = \pi_1 > 0$ for all $i \in X$. Let q_i^k be the probability of reaching solution 1 for the *first* time at iteration k when starting with solution i. Note that

$$P_{i1}^{(1,k)} = \sum_{l=1}^{k} q_i^l P_{11}^{(l+1,k)} = \sum_{l=1}^{\infty} \mu_k(l) P_{11}^{(l+1,k)} \quad \text{with } \mu_k(l) = 1\{l \leqq k\} q_i^l.$$

Note the $\{\mu_k(\cdot)\}$ converges setwise on the set of positive integers. Using Royden (1968), p. 232, and letting k tend to ∞ we conclude that $\pi_1 = (\Sigma_{l=1}^{\infty} q_i^l)\pi_1$ and since $\pi_1 > 0$, $\Sigma_{l=1}^{\infty} q_i^l = 1$, thus proving (a).

The equivalence of weak and strong ergodicity ((b)$\Rightarrow$(c)) fails to hold for general non-stationary chains, see Isaacson and Madsen (1976) and Anily and Federgruen (1985a).

For any matrix A, let $\| A \| = \max_i \Sigma_j |A_{ij}|$.

3. Main results

Theorem 2 below states *necessary* and *sufficient* conditions for properties (a)–(d). Let

$$\underline{a}(c) = \min_{i \in X, j \in X_i} a_{ij}(c); \quad \bar{a}(c) = \max_{i \in X^*, j \in X_i \setminus X^*} a_{ij}(c) \tag{3}$$

where X^* is the set of recurrent states under $P(\infty)$, i.e., all local optima.

Theorem 2.

(i) If $\Sigma_{k=1}^{\infty} \underline{a}^n(c_{kn}) = \infty$, properties (a)–(c) hold and $\lim_{k\to\infty} \pi(k)$ exists; thus, if $\lim_{k\to\infty} \pi(k)_j = 0$ for all j, with $f_j > f_1$, property (d) holds as well.

(ii) If suboptimal local optima exist and $\Sigma_{k=1}^{\infty} \bar{a}(c_k) < \infty$ then none of the properties (a)–(d) hold.

Proof.

(i) Let $d = \min_{i\in X, j\in X_i} g_{ij} > 0$. Let k^* be large enough that $\underline{a}(c_k)$ is monotone in k for $k \geqq k^*$, see (1). Note that $P_{ij}(l) \geqq d\underline{a}(c_l)$ for all $i \in X$, $j \in X_i$ $l \geqq 1$. Solution 1 can be reached from any other solution in n or less iterations and since $1 \in X_1$ also in *exactly* n iterations. It follows from Definition 3 and (3) that $\alpha(P^{(kn+1,(k+1)n)}) \geqq \min_i P_{i1}^{(kn+1,(k+1)n)} \geqq d^n \underline{a}^n(c_{(k+1)n})$ for all $k \geqq \lceil k^*/n \rceil$. Let $p_i^k = \Pr\{s_{ln} \neq 1, l = 1, \cdots, k \mid s_1 = i\}$. Then,

$$p_i^k = \sum_{j\neq 1} \Pr\{s_n \neq 1, \cdots, s_{(k-2)n} \neq 1, s_{(k-1)n} = j\}(1 - P_{j1}^{((k-1)n+1,kn)})$$

$$\leqq (1 - d^n \underline{a}^n(c_{kn}))p_i^{k-1}$$

and

$$p_i^k \leqq \prod_{l=1}^{k} (1 - d^n \underline{a}^n(c_{ln})) \to 0$$

in view of Theorem I.2.5 in Isaacson and Madsen (1976). This proves (a). Also, $\Sigma_{k=1}^{\infty} \alpha(P^{(kn+1,(k+1)n)}) = \infty$, and $\{P(k)\}_{k=1}^{\infty}$ is weakly ergodic, in view of Theorem V.3.2. in Isaacson and Madsen (1976). Thus property (b) holds. The remaining assertions all follow from Theorem 1.

(ii) Let i be a local optimum with $f_i > f_1$. Solutions 1 and i are part of two distinct subchains C_1 and C_2 of $P(\infty)$, see (1) and (2). Note that for $C = C_1, C_2$ and all $j \in C$, $\Sigma_{l\in C} P(k)_{jl} = -\Sigma_{l\notin C} g_{jl} a_{jl}(c_k) \geqq 1 - \bar{a}(c_k)$. Thus the probability of eternally staying in a subchain C is bounded from below by $\Pi_{k=1}^{\infty} [1 - \bar{a}(c_k)] > 0$ since $\Sigma_{k=1}^{\infty} \bar{a}(c_k) < \infty$; see, for example, Theorem I.2.5 in Isaacson and Madsen (1976). Thus property (a) fails to hold. Also, $\liminf_{k\to\infty} P_{11}^{(1,k)} > 0$. To prove the remaining assertions, assume to the contrary that (b) holds (see Theorem 1(i)). In view of Theorem 1(ii), $\lim_{k\to\infty} \pi(k)_1 = \lim_{k\to\infty} P_{11}^{(1,k)} > 0$. Thus, in view of Theorem 1(iii), (a) holds, a contradiction.

We now discuss the implications of Theorem 2 for the most commonly used a.p.f.'s.

1. *Exponential or Metropolis a.p.f.'s*: $a_{ij}^M(c) = \min\{1, \exp((f_i - f_j)/c)\}$, there exists a symmetric matrix Q such that $g_{ij} = Q_{ij}/\Sigma_l Q_{il}$.

2. *Hastings' a.p.f.'s* (see Hastings (1970) and Gidas (1985)).

$$a_{ij}^H(c) = \frac{1 + 2[\frac{1}{2}\min\{(g_{ij}/g_{ji})\exp((f_j - f_i)/c), (g_{ji}/g_{ij})\exp((f_i - f_j)/c)\}]^{\gamma}}{1 + (g_{ij}/g_{ji})\exp((f_j - f_i)/c)},$$

$$\gamma \geqq 1.$$

(The cases $\gamma = 1$ and $\gamma = \infty$ correspond with 'generalized Metropolis' and 'heat bath' probabilities respectively.)

It is easily verified that both types of a.p.f.'s satisfy (1) and the regularity conditions stated in Section 2. In both cases, closed-form expressions for $\{\pi(k)\}$ are easily derived and $\lim_{k\to\infty} \pi(k)_j = 0$ if j is not a global minimum (see Hastings (1970), Lundy and Mees (1986), Romeo and Sangiovanni-Vincentelli (1984) and Anily and Federgruen (1985b)). Let

$$\Delta^+ = \max_{i,j} \{f_j - f_i \mid j \in X_i, f_j > f_i\}; \quad \Delta^- = \min\{f_j - f_i \mid i \in X^*, j \in X_i \setminus X^*\}.$$

Theorem 2 implies that *all* four properties (a)–(d) hold if $c_k \geqq n\Delta^+/\log k$, $k \to \infty$, while *none* holds for *any* problem with (suboptimal) local optima if $c_k \leqq \Delta^-/(\log k)$, $k \to \infty$. The necessary and sufficient conditions in Theorem 2 are thus quite tight for both types of a.p.f.'s.

We now give examples of a.p.f.'s which could be used as alternatives for the Metropolis or Hastings functions.

Example 1. For all $i \in X, j \in X_i$ let $t_{ij}(c)$ be a polynomial in c with positive coefficients. For all $i \in X$, let $g_{ij} = 1/|X_i|, j \in X_i$ and let $\gamma > 1$. Define

$$a_{ij}(c) = \min\{1; a_{ij}^M(c) + t_{ij}(c) \exp((f_i - f_j)/c^\gamma)\} \tag{4}$$

and $c_k \geqq n\Delta^+/\log k$. Note that the term $t_{ij}(c)\exp((f_i - f_j)/c^\gamma)$ dominates the 'Metropolis' term at the beginning of the algorithm when k is small, while the Metropolis term dominates towards the end. (In the beginning relatively faster decreases in the acceptance probabilities can thus be achieved.) It is again easily verified that the a.p.f.'s in (4) satisfy (1) and the regularity conditions stated in Section 2. (The a.p.f.'s belong to a CAM, see Definition 1.) We show that $\lim_{c\downarrow 0} \pi(c)_i = 0$ if i is not a global minimum. In view of Theorem 2, this establishes that all four properties (a)–(d) hold.

Let $P(c) = (p_{ij}(c))$ and $P^M(c) = (p_{ij}^M(c))$ where $p_{ij}^M(c)$ is defined by (2) with $a(\cdot) = a^M(\cdot)$. Let $\Pi^M(c)$ be a matrix with identical rows $\pi^M(c)$, the steady state probability vector associated with $P^M(c)$. As stated above, a closed-form expression for $\pi^M(c)$ is easily obtained and given by (see e.g. Lundy and Mees (1986)): (set $f_1 = 0$, without loss of generality),

$$\pi^M(c)_i = \exp(-f_i/c) \Big/ \left(1 + \sum_{l=2}^{N} \exp(-f_l/c)\right), \qquad i = 1, \cdots, N. \tag{5}$$

Also, let $Y^M(c)$ denote the deviation matrix associated with $P^M(c)$:

$$Y^M(c) \overset{\text{def}}{=} [I - P^M(c) + \Pi^M(c)]^{-1} - \Pi^M(c).$$

Finally, let $\Delta(c) = P(c) - P^M(c)$. The following perturbation result follows from Schweitzer (1968), see e.g. Meyer (1980):

$$\| \pi(c) - \pi^M(c) \| \leqq \| \pi^M(c) \| \, \| \Delta(c) \| \, \| Y^M(c) \| / (1 - \| \Delta(c) \| \, \| Y^M(c) \|).$$

Since $\| \pi^M(c) \| < 1$, it suffices to verify that $\lim_{c \downarrow 0} \| \Delta(c) \| \, \| Y^M(c) \| = 0$. Note from (2) and (5) that each entry of the matrix $[I - P^M(c) + \Pi^M(c)]$ is of the form $(\Sigma_{l \in I} q_l \exp(-\beta_l/c))/(\Sigma_{l \in I'} q'_l \exp(-\beta'_l/c))$ with I, I' finite index sets and all $\beta_l, \beta'_l \geqq 0$. Since this class of functions is closed under addition, multiplication and division, it follows from Cramer's rule that each entry in $[I - P^M(c) + \Pi^M(c)]^{-1}$ and hence each entry in $Y^M(c)$ is of this form as well. Since $\| \Delta(c) \| = o(\exp(-\beta/c))$, $c \downarrow 0$ for any $\beta > 0$ it follows that

$$\lim_{c \downarrow 0} \| \Delta(c) \| \, \| Y^M(c) \| = 0.$$

Example 2. Consider the K-median problem and the neighborhood structure described in Section 2. Consider an expanded neighborhood structure $\{\hat{X}_i : i \in X\}$ where the neighborhood of a given set of medians is given by the collection of sets which may be obtained by adding *one or two* points outside of the current set or by substituting this (these) point(s) for one (two) that currently is (are) in the set. Clearly, $X_i \subset \hat{X}_i$, $i = 1, \cdots, N$. Again, for all $i, j \in X$ let $t_{ij}(c)$ be a polynomial in c with positive coefficients. Let $g_{ij} = 1/|\hat{X}_i|$, $i \in X$, $j \in \hat{X}_i$ and choose $\gamma > 1$.

Define

$$a_{ij}(c) = \begin{cases} a_{ij}^M(c), & j \in X_i, \\ \min\{1; t_{ij}(c) \exp((f_i - f_j)/c^\gamma)\}, & j \in \hat{X}_i \setminus X_i. \end{cases}$$

In the initial phase of the algorithm (when c is relatively large), these a.p.f.'s allow for deteriorating switches in the expanded neighborhoods. As c is decreased to 0, such switches become progressively less likely than interchanges within the restricted neighborhoods $\{X_i : i \in X\}$. The proof that properties (a)–(d) hold is analogous to that in Example 1.

Theorem 2 shows, in addition, that *many popular* schemes (e.g., $c_k = \beta^k$ with $\beta < 1$, as well as the scheme proposed in Aragon et al. (1985)) *fail to exhibit any of the desired properties* when applied to any problem with (suboptimal) local optima. Schemes in which c is only decreased at a prespecified sequence of iterations $\{k_l\}_{l=1}^{\infty}$ (i.e., $c_{k_l} = c_{k_l+1} = \cdots = c_{k_{l+1}-1}$) fail if $c_{k_l} = l^{-\varepsilon}$ and $k_l = l^\beta$ for *any* choice of β, $\varepsilon > 0$. (Note that $\Sigma_{r=1}^{\infty} \bar{a}(c_r) = \Sigma_{l=1}^{\infty} l^\beta \exp(-\Delta^{-l^\varepsilon}) < \infty$, see Theorem 2(ii).)

The following example shows that for some a.p.f.'s, property (d) cannot be achieved regardless of how slowly $\{c_k\}$ decreases to 0.

Example 3. Let $a_{ij}(c) = \gamma_i(c)/(f_j - f_i)$ if $f_j > f_i$ and $a_{ij}(c) = 1$ otherwise, where $\gamma_i(c) \downarrow 0$, as $c \downarrow 0$. Note that these a.p.f.'s satisfy (1) as well as the regularity conditions provided the $\gamma_i(\cdot)$ functions do. Let $X = \{1, \cdots, 4\}$ with $f_1 = 0, f_2 = 1, f_3 = f_4 = 2$. Let $X_1 = \{1, 3, 4\}, X_2 = \{2, 3, 4\}, X_3 = \{1, 2, 3\}, X_4 = \{1, 2, 4\}$ and all $g_{ij} = 1/3$. Finally let $\gamma_1(c) = c$ and $\gamma_2(c) = c^2$. The steady-state equations for $\{\pi(k)\}$ show that

$$\tfrac{1}{3}c_k\pi(k)_1 = \tfrac{2}{3}c_k^2\pi(k)_2 \quad \text{or} \quad \pi(k)_2/\pi(k)_1 = 1/(2c_k) \to \infty$$

for *any* decreasing sequence $\{c_k\}$. Hence $\lim_{k\to\infty}\pi(k)_1 = 0$. Choose $c_k = k^{-1}$. All entries in the second column of products $P(k)P(k+1)$ are bounded from below by $(9k)^{-1}$ and in view of Theorem V.3.2 in Isaacson and Madsen (1976) property (b) holds. We conclude using Theorem 1(ii) that the method *converges* to the *local optimum* 2 with probability 1 *however slowly* $\{c_k\}$ *is decreased to* 0. Note also that in this example there is a positive probability of always generating solution 2 thus showing that the condition $\lim\sup_{k\to\infty}\pi(k)_1 > 0$ is essential in Theorem 1(iii).

We conclude with a general upper bound for the *rate* of convergence. For exponential a.p.f.'s, a similar bound was obtained independently in Mitra et al. (1986).

Theorem 3 (rate of convergence).

(i) Assume (b) (and hence (c)) hold and let $\alpha^* = \alpha(G^n)$. Then,

$$\sum_j |P_{ij}^{(1,k)} - \pi_j| \leqq 2\exp\left(\alpha^* \sum_{r=0}^{\lfloor(k-\sqrt{k-n})/n\rfloor - 1} \underline{a}^n(c_{k-rn})\right) + \sum_{r=\lfloor(k-n)^{1/2}\rfloor - n}^{\infty} \|\pi(r) - \pi(r+1)\|, \qquad k \geqq 1. \tag{6}$$

(ii) In particular, for exponential a.p.f.'s with $c_k = n\Delta^+/\log k$ there exists a constant U such that

$$\sum_j |P_{ij}^{(1,k)} - \pi_j| \leqq 2\left(\frac{1}{k-n}\right)^{\alpha^*/n} + Uk^{(-f_{N_0+1}-f_1)/2n\Delta^+}$$

where $N_0 = \max\{i \mid f_i = f_1\}$.

Proof.

(i) Let T and T_k be $N \times N$ matrices with identical rows π and $\pi(k)$ respectively, $k \geqq 1$. In view of the proof of Theorem V.4.3 in Isaacson and Madsen (1976), we get

$$\|P^{(1,k)} - T\| \leqq \|P^{(1,l)}P^{(l+1,k)} - T_{l+1}P^{(l+1,k)}\| + \sum_{r=l}^{\infty}\|\pi(r) - \pi(r+1)\|, \qquad 1 \leqq l \leqq k. \tag{7}$$

Choose $l = l^* = \max\{l \mid l \leqq \sqrt{k-n}$ and $k = l + mn$ for some $m \geqq 1\}$. It suffices to show that the first term to the right of (7) is bounded by the corresponding term in (6).

Using the proof of Theorem V.4.3 and Lemma V.2.3 in Isaacson and Madsen (1976) we get

$$
\begin{aligned}
B &\equiv \| P^{(1,l^*)}P^{(l^*+1,k)} - T_{l^*+1}P^{(l^*+1,k)} \| \\
&\leqq 2(1 - \alpha(P^{(l^*+1,k)})) \\
&\leqq 2 \prod_{r=0}^{m-1} (1 - \alpha(P^{(l^*+1+rn,l^*+(r+1)n)})) \\
&\leqq 2 \prod_{r=1}^{m} (1 - \alpha^* \underline{q}^n(c_{l^*+rn})) \\
&= 2 \prod_{r=0}^{m-1} (1 - \alpha^* \underline{q}^n(c_{k-rn})).
\end{aligned}
$$

To verify the last inequality, let $\bar{P} = P^{(l^*+1+rn,l^*+(r+1)n)}$. Note that

$$
\alpha(\bar{P}) = \min_{i,l} \left(\sum_j \min(\bar{P}_{ij}, \bar{P}_{lj}) \right) \geqq \underline{q}^n(c_{l^*+rn}) \min_{i,l} \left(\sum_j \min(G_{ij}^{(n)}, G_{lj}^{(n)}) \right).
$$

Note that $m = (k - l^*)/n \geqq (k - \sqrt{k-n})/n$; thus,

$$
B \leqq 2 \exp \left(\sum_{r=0}^{\lfloor (k-\sqrt{k-n})/n \rfloor} \log(1 - \alpha^* \underline{q}^n(c_{k-rn})) \right)
$$

and use $\log(1 - z) \leqq -z$.

(ii) Follows from (i) as in Mitra et al. (1986) and Anily and Federgruen (1985b).

Note from the definition of the ergodic coefficient that d^n (or Nd^n if *every* solution is reachable in n steps from every other solution) is a lower bound for α^* in (6).

General results for the rate of convergence of non-stationary Markov chains are only known (see Huang et al. (1976)) for the special case where $P(\infty)$ has a single subchain, i.e., no local optima exist (in which case deterministic search methods are clearly to be preferred). Incidentally, applying Huang et al. (1976) to this special case with exponential a.p.f.'s, results in a $O(k^{-\Delta/n\Delta^+})$ convergence rate bound where $\underline{\Delta} = \min\{f_j - f_i \mid f_j > f_i,\ j \in X_i\}$. This bound is in many settings inferior to the one obtained in Theorem 3. Information about the distribution of $\{f_i, i = 1, \cdots, N\}$ in specific problem settings may, of course, be exploited to refine the bounds in Theorem 3.

4. Unrestricted random searches

If *unrestricted* random searches were performed, i.e., if all $X_i = X$, $i \in X$, then $P(\infty)_{i1} \geqq d$ for all $i \in X$ and any $\{c_k\}$. By Theorem V.3.2 in Isaacson and Madsen (1976), property (b) holds and $\lim_{k\to\infty} \pi(k)_j > 0$ if and only if j is a global optimum. Thus, in view of Theorem 1 all four properties (a)–(d) hold. Following the proof of Theorem 3, one concludes that the convergence rate is

$O((1-(1/N)^k)$, $k \to \infty$ (provided $\{c_k\}$ decreases to 0 sufficiently fast). Thus convergence is extremely slow even though the rate is geometric.

Note added in proof. Recent empirical studies (I. Bohachavsky, M. Johnson and M. Stein, Generalised simulated annealing for function optimization, *Technometrics* **28** (1986), 209–217), suggest that annealing methods with generalized a.p.f.'s of the form: $a_{ij}(c) = \exp\{f_i^g(f_i - f_j)/c\}$ $(i, j \in X; g \leqq 0)$ outperform corresponding methods with standard exponential a.p.f.'s (where $g = 0$). This class of a.p.f.'s clearly satisfies the conditions in our paper.

References

ANILY, S. AND FEDERGRUEN, A. (1985a) Ergodicity in parametric non-stationary Markov chains: Application to simulated annealing methods. *Operat. Res.* To appear.

ANILY, S. AND FEDERGRUEN, A. (1985b) Probabilistic analysis of simulated annealing (Working paper; unabridged version of this paper).

ARAGON, C., JOHNSON, D. MCGEOGH, L. AND SCHEVON, C. (1985) Optimization by simulated annealing: An experimental evaluation.

BINDER, K. (1978) *Monte Carlo Methods in Statistical Physics*. Springer-Verlag, Berlin.

CERNY, V. (1985) A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Applic.* **15**, 41–51.

CORNUEJOLS, G., FISHER, M. AND NEMHAUSER, G. (1977) Location of bank accounts to optimize float: An analytic study of exact and approximate solutions. *Management Sci.* **23**, 789–811.

DOBRUSHIN, R. (1956) Central limit theorems for non-stationary Markov chains II. *Theory Prob. Appl.* **1**, 329–383.

GEMAN, S. AND GEMAN, D. (1984) Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Proc. Pattern Analysis and Machine Intelligence* **6**, 721–741.

GIDAS, B. (1985) Non-stationary Markov chains and convergence of the annealing algorithm. *J. Statist. Phys.* **39**, 73–131.

HAJEK, B. (1985) Cooling schedules for optimal annealing (working paper).

HASTINGS, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

HUANG, C., ISAACSON, D. AND VINOGRADE, B. (1976) The rate of convergence of certain nonhomogeneous Markov chains. *Z. Wahrscheinlichkeitsth.* **35**, 141–146.

ISAACSON, D. AND MADSEN, R. (1976) *Markov Chains: Theory and Applications*, Wiley and Sons, New York.

KIRKPATRICK, S., GELAT, JR. C. AND VECCHI, M. (1983) Optimization by simulated annealing. *Science* **220**, 671–680.

LUNDY, M. AND MEES, A. (1986) Convergence of the annealing algorithm. *Math. Programming* **34**, 111–124.

METROPOLIS, W., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. AND TELLER, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

MEYER, C. D. (1980) The solution of a finite Markov chain and perturbation bounds for the limiting probabilities, *SIAM J. Algebraic and Discrete Methods* **1**, 273–283.

MITRA, D., ROMEO, F. AND SANGIOVANNI-VINCENTELLI, A. (1986) Convergence and finite-time behavior of simulated annealing. *Adv. Appl. Prob.* **18**, 747–771.

ROMEO, F. AND SANGIOVANNI-VINCENTELLI, A. (1984) Probabilistic hill climbing algorithms: Properties and applications. Electronics Research Laboratory, University of California, Berkeley, California.

ROYDEN, H. (1968) *Real Analysis*, 2nd edn., MacMillan, London.

SCHWEITZER, P. J. (1968) Perturbations and finite Markov chains. *J. Appl. Prob.* **5**, 401–413.

WILSON, K., JACOBS, D. AND PRINS, J. (1982) Statistical mechanics algorithm for Monte Carlo optimization. *Physics Today*.