SIMULATED ANNEALING

by

Saul B. Gelfand[1]
Sanjoy K. Mitter[2]

[1] Department of Electrical Engineering, Purdue University, Lafayette, Indiana

[2] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139

This research has been supported by the Air Force Office of Scientific Research under grant AFOSR-85-0227 and the Army Research Office under grants DAAG29-84-K-0005 and DAAL03-86-K-1071.

CHAPTER I
INTRODUCTION

Algorithms for finding a global extremum of a real-valued function may be classified into two groups: deterministic and random. The distinction here is of course that the random or Monte-Carlo algorithms make use of pseudo random variates whereas the deterministic algorithms do not. The earliest global optimization algorithms were of the deterministic type and were associated with evaluating the cost function at points on a grid. One drawback of these methods is that they typically require certain prior information about the cost function such as a Lipshitz constant. Most global optimization algorithms are of the random type and are related to the so-called multistart algorithm. In this approach, a local optimization algorithm is run from different starting points which are selected at random, usually from a uniform distribution on the domain of the cost function. See [5], [29] for a discussion of global optimization algorithms.

Recently, motivated by hard combinatorial optimization problems such as arise in computer design and operations research, Kirkpatrick et. al. [19] and independently Cerny [3] have proposed a different kind of random algorithm called *simulated annealing*. The annealing algorithm is based on an analogy between large scale optimization problems and statistical mechanics. For our purposes this analogy consists simply of viewing the cost function as an energy function defined on a finite state space of an imaginary physical system. The annealing algorithm is then seen as a variation on a Monte-Carlo algorithm developed by Metropolis et. al. [25] for making statistical mechanics calculations, which we now describe. It is well-known that the states of a physical system in thermal equilibrium obey a Gibbs distribution $\propto \exp[-U(\cdot)/T]$, where $U(\cdot)$ is an energy function and T is the temperature. The Metropolis algorithm was developed for obtaining samples from such a Gibbs distribution and for computing estimates of functionals averaged over the Gibbs distribution. The Metropolis algorithm proceeds as follows:

> Given a state i of the system, select a candidate state j in a random manner corresponding to a small perturbation of the system, and compute the change in energy $\Delta U = U(j) - U(i)$. If $\Delta U \leq 0$ accept state j as the new state for the next iteration of the algorithm. If $\Delta U > 0$ accept state j with probability $\exp[-\Delta U/T]$; otherwise the algorithm starts at state i for the next iteration.

The annealing algorithm consists of identifying the cost function to be minimized with the energy function $U(\cdot)$ and taking the temperature T as a function of time and slowly lowering it to zero. Suppose that the distribution of a candidate state is independent of past states given the current state. Then it is clear that the Metropolis algorithm simulates the sample paths of a Markov chain, and it can be shown that if the candidate states are selected in a suitable manner then this chain infact has a Gibbs distribution $\propto \exp[-U(i)/T]$ as its (unique) equilibrium distribution (see Chapter 2 for details). Furthermore as the temperature T is decreased to zero the Gibbs distribution concentrates more and more on the lower energy states. The motivation behind the annealing algorithm is that if $T \rightarrow 0$ slowly enough such that the system is never far away from equilibrium, then presumably there is convergence (in some probabilistic sense) to the global minima of $U(\cdot)$.

The annealing algorithm stands in contrast to heuristic methods for combinatorial optimization which are based on iterative improvement, allowing only decreases in the cost function at each iteration. Iterative improvement algorithms in statistical mechanics terms correspond to rapidly quenching a system from a high to a very low temperature. Such quenching can result in the system getting trapped in a so-called metastable state, and analogously the iterative improvement algorithm getting trapped in a strictly local minimum of the cost function. On the other hand, the annealing algorithm corresponds to slowly cooling a system. Such cooling should result in the system spending most of its time among low energy states and analogously the annealing algorithm finding a global or nearly global minimum of the cost function.

The annealing algorithm as described above is suitable for combinatorial optimization. Motivated by optimization problems with continuous variables which arise in image processing problems, Geman and independently Grenander [13] have proposed a diffusion-type algorithm called the *Langevin algorithm* (as coined by Gidas [11]). Consider the diffusion solution of the Langevin equation

$$dx(t) = -\nabla U(x(t))dt + \sqrt{2T}\, dw(t)$$

where $U(\cdot)$ is now a smooth function on r-dimensional Euclidean space (again called energy), T is a positive constant (again called temperature), and $w(\cdot)$ is a standard r-dimensional Wiener process. The Langevin equation describes the motion of a particle in a viscous fluid. The Langevin algorithm consists of identifying the cost function to be minimized with the energy function $U(\cdot)$ and taking the temperature T as a function of time and slowly lowering it to zero. Now it is well known that under suitable conditions on $U(\cdot)$ the diffusion solution of the Langevin equation has a Gibbs density $\propto \exp[-U(\cdot)/T]$ as its (unique) equilibrium density, and as the temperature T is decreased to zero this density becomes more and more concentrated on the lower energy states. Like the annealing algorithm, the motivation behind the Langevin algorithm is that if $T \to 0$ slowly enough such that the system is never far away from equilibrium, then presumably there is convergence (in some probabilistic sense) to the global minima of $U(\cdot)$.

The annealing algorithm has been applied with varying success to a wide range of problems including circuit placement and wire routing for VLSI chip design [19], image reconstruction [8], and assorted hard combinatorial problems which arise in operations research [3], [12], [18], [19]. There has also been intense theoretical interest in both the annealing algorithm [8], [10], [11], [14], [15], [26], [31] and the Langevin algorithm [4], [9], [11], [15], [21].

CHAPTER II
FINITE STATE ANNEALING TYPE ALGORITHMS

2.1 Introduction to the Annealing Algorithm

In Chapter 1 we briefly described the annealing algorithm and discussed the heuristic motivation based on the connection that Kirkpatrick [19] has suggested between statistical mechanics and large-scale optimization problems. Mathematically, the annealing algorithm consists of simulating a nonstationary finite-state Markov chain whose state space is the domain of the cost function (called energy) to be minimized. In this Section we shall discuss in detail the annealing algorithm and describe some of the considerable literature which has been devoted to its analysis.

We first give some standard finite state space Markov chain notation (c.f. [6], [7]). Let Σ be a finite set. $P = [p_{ij}]_{i,j\in\Sigma}$ is a stochastic matrix on Σ if $p_{ij} \geq 0$ for all $i,j\in\Sigma$ and

$$\sum_{j\in\Sigma} p_{ij} = 1 \qquad \forall\, i\in\Sigma\,.$$

$\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ are the 1-step transition matrices for a Markov chain $\{\xi_k\}$ with state space Σ if for every $k\in\mathbb{N}$ $P^{(k,k+1)}$ is a stochastic matrix on Σ and

$$P\{\xi_{k+1} = j\,|\,\xi_k = i\} = p_{ij}^{(k,k+1)} \quad (\text{if } P\{\xi_k = i\} > 0) \tag{2.1}$$

for all $i,j\in\Sigma$. Conversely, given a sequence $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+d)}]\}$ of stochastic matrices on Σ we can construct on a suitable probability space $(\Lambda,\mathcal{F},P)$ a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.1). For each $d\in\mathbb{N}$ let

$$P^{(k,k+d)} = P^{(k,k+1)} \cdot \ldots \cdot P^{(k+d-1,k+d)}\,.$$

$P^{(k,k+d)} = [p_{ij}^{(k,k+d)}]$ is a stochastic matrix on Σ and

$$P\{\xi_{k+d} = j\,|\,\xi_k = i\} = p_{ij}^{(k,k+d)} \quad (\text{if } P\{\xi_k = i\} > 0)$$

for all $i,j\in\Sigma$. It will be convenient to have a fixed version of the conditional probability of ξ_{k+d} given ξ_k which we define by

$$P\{\xi_{k+d}\in A|\xi_k = i\} = \sum_{j\in A} p_{ij}^{(k,k+d)}$$

for all $i\in\Sigma$ and $A\subset\Sigma$.

We now define the annealing algorithm. Let $U(\cdot)$ be a nonnegative function on Σ, called the *energy function.* The goal is to find a point in Σ which minimizes or nearly minimizes $U(\cdot)$. Let $\{T_k\}$ be a sequence of positive numbers, called the *temperature schedule*. Let $Q = [q_{ij}]$ be a stochastic matrix on Σ. Now let $\{\xi_k\}$ be the Markov chain with state space Σ and 1-step transition matrices $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ given by

$$p_{ij}^{(k,k+1)} = \begin{cases} q_{ij}\exp\left[-\dfrac{U(j)-U(i)}{T_k}\right] & \text{if } U(j) > U(i) \\ q_{ij} & \text{if } U(j) \le U(i)\ ,\ j\neq i \\ 1-\sum\limits_{\ell\neq i} p_{i\ell}^{(k,k+1)} & \text{if } j = i \end{cases} \qquad (2.2)$$

for all $i,j\in\Sigma$. $\{\xi_k\}$ shall be called the *annealing chain.* For each $d\in\mathbb{N}$ let $Q^d = [q_{ij}^{(d)}]$. Recall that Q is *irreducible* if for every $i,j\in\Sigma$ there exists a $d\in\mathbb{N}$ such that $q_{ij}^{(d)} > 0$. Also, Q is *symmetric* if $q_{ij} = q_{ji}$ for all $i,j\in\Sigma$. In the special case where Q is irreducible and symmetric and $T_k = T$, a positive constant, $\{\xi_k\}$ is the stationary Markov chain introduced by Metropolis et. al. [25] for computing statistics of a physical system in thermal equilibrium at temperature T. It was Kirkpatrick et. al. [19] and Cerny [3] who suggested that the Metropolis scheme could be used for minimizing $U(\cdot)$ by letting $T = T_k \to 0$. We shall call the algorithm which simulates the sample paths of $\{\xi_k\}$ with $T_k\to 0$ the *annealing algorithm.*

The heuristic motivation behind the annealing algorithm was discussed (briefly) in Chapter 1. Here we give the motivation in more mathematical terms. Suppose that Q is irreducible and symmetric, and let $\{\xi_k^T\}$ be the stationary chain with 1-step (stationary) transition matrix $P^T = [p_{ij}^T]$ given by the r.h.s of (2.2) with $T_k = T$, a positive constant. Then it can be shown that P^T has an invariant Gibbs vector $\Pi^T = [\pi_i^T]$ (a row vector), i.e.,

$$\Pi^T = \Pi^T P^T$$

where

$$\pi_i^T = \frac{\exp[-U(i)/T]}{\sum\limits_{j\in\Sigma}\exp[-U(j)/T]} \qquad \forall\ i\in\Sigma\ .$$

This follows from the detailed reversibility

$$\pi_i^T p_{ij}^T = \pi_j^T p_{ji}^T \qquad \forall\ i,j\in\Sigma\ .$$

Furthermore, Q irreducible and symmetric implies that $\{\xi_k^T\}$ is an irreducible† (and aperiodic) chain and by the Markov Convergence Theorem [6, p. 177]

$$\lim_{k\to\infty} P\{\xi_k^T = i\} = \pi_i^T \qquad \forall\ i\in\Sigma\ . \qquad (2.3)$$

Let S be the set of global minima of $U(\cdot)$, i.e.

$$S = \{i\in\Sigma:\ U(i) \le U(j)\ \ \forall\ j\in\Sigma\}\ .$$

Now

$$\lim_{T\to 0} \pi_i^T = \pi_i^* \qquad \forall\ i\in\Sigma \qquad (2.4)$$

where $\Pi^* = [\pi_i^*]$ is a probability vector with support in S. In view of (2.3) and (2.4) the idea behind the annealing algorithm is that by choosing $T = T_k \to 0$ slowly enough hopefully

$$P\{\xi_k = i\} \approx \pi_i^{T_k} \qquad (k\ \ \text{large}) \qquad (2.5)$$

and then perhaps

$$\lim_{k\to\infty} P\{\xi_k = i\} = \pi_i^* \qquad \forall\ i\in\Sigma \qquad (2.6)$$

and consequently ξ_k converges in probability to S.

In Chapter 1 we roughly described the procedure by which the sample paths of the annealing chain are simulated. It is seen that the Q matrix governs the small perturbations in the system configurations which are then accepted or rejected probabilistically depending on the corresponding energy changes and the temperature. More precisely, the annealing chain may be simulated as follows. Suppose $\xi_k = i$. Then generate a Σ-valued random variable η with $P\{\eta = j\} = q_{ij}$. Suppose $\eta = j$. Then set

†A stationary chain is irreducible if its 1-step (stationary) transition matrix is irreducible.

$$\xi_{k+1} = \begin{cases} j & \text{if } U(j) \leq U(i) \\ j & \text{if } U(j) > U(i) \text{ with probability } \exp\left[-\dfrac{U(j)-U(i)}{T_k}\right] \\ i & \text{else} \end{cases}$$

There are two in depth numerical studies of simulated annealing of which we are aware. Johnson et. al. [18] applied the annealing algorithm to four well-studied problems in combinational optimization: graph partitioning, number partitioning, graph coloring, and the travelling salesman problem. They compare the annealing algorithm with the best of the traditional algorithms for each problem. They found that although annealing is able to produce quite good solutions on three of the four problems, only on one of the four (graph partitioning) does it outperform the best of its rivals. Golden and Skiscim [12] have tested the annealing algorithm on routing and location problems, specifically the travelling salesman problem and the p-median problem. They conclude that there are more efficient and effective heuristics for these problems.

We shall now outline the convergence results on the annealing algorithm which are known to us. We refer the reader to the specific papers for full details.

Geman and Geman [8] were the first to obtain a convergence result for the annealing algorithm. The consider a version of the annealing algorithm which they call the *Gibbs sampler*. They show that for temperature schedules of the form

$$T_k = \frac{c}{\log k} \qquad (k \text{ large})$$

that if c is sufficiently large then (2.6) is obtained.

Gidas [10] also considers the convergence of the annealing algorithm and similar algorithms based on Markov chain sampling methods related to the Metropolis method.

We next discuss the work of Mitra et. al. [26]. The idea behind their work is similar to that of Geman and Geman and also Gidas in that they show that for temperature schedules which vary slowly enough the annealing chain reaches "quasiequilibrium", i.e., something like (2.5) holds. In order to state Mitra et. al.'s result we will need the following notation. Let

$$N(i) = \{j \in \Sigma : \ q_{ij} > 0\} \qquad \forall\, i \in \Sigma\,.$$

Let S_M be the set of states that are local maxima of $U(\cdot)$, i.e.,

$$S_M = \{i \in \Sigma : \ U(i) \geq U(j) \quad \forall\, j \in N(i)\}\,.$$

Let

$$r = \min_{i \in \Sigma \backslash S_M} \max_{j \in \Sigma} d(i,j)$$

where d(i,j) is the minimum number of steps to get from state i to state j. Finally, let

$$L = \max_{i \in \Sigma} \max_{j \in N(i)} |U(j) - U(i)|\,.$$

Here is Mitra et. al.'s result:

Theorem 2.1 (Mitra et. al. [26]) Assume Q is irreducible and symmetric†. Let $T_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{r\,L}{T_{kr-1}}\right) = \infty\,. \tag{2.7}$$

Then

$$\lim_{k \to \infty} P\{\xi_k = i\} = \pi_i^* \qquad \forall\, i \in \Sigma\,. \tag{2.8}$$

Remarks

(1) If $T_k = c/\log k$ then (2.7) holds iff $c \geq r\,L$.

(2) An estimate of the rate of convergence in (2.8) is obtained for annealing schedules of the form $T_k = c/\log k$ for $c \geq r\,L$. Let

$$w = \min_{i \in \Sigma} \min_{j \in N(i)} q_{ij}\ ,$$

$$\gamma = \min_{i \in \Sigma \backslash S} U(i) - \min_{j \in S} U(j)\,.$$

It is shown that

$$P\{\xi_k = i\} = \pi_i^* + O\left(\frac{1}{k^{\min\{\alpha,\beta\}}}\right) \qquad \text{as } k \to \infty \tag{2.9}$$

where

† for just $q_{ij} > 0$ iff $q_{ji} > 0$ for all $i,j \in \Sigma$

$$\alpha = \frac{w^r}{r^{r\,L/c}} \quad , \quad \beta = \frac{\gamma}{c} \; .$$

Since α and β are increasing and decreasing respectively with increasing c, it is suggested that $c \geq r\,L$ be chosen to maximize $\min\{\alpha,\beta\}$.

We next discuss the work of Hajek [14]. The idea behind his work is that for temperature schedules which vary slowly enough, the annealing chain escapes from local minima of U(·) at essentially the same rate as for a constant temperature. In order to state Hajek's result we will need the following notation. We shall say that given states i and j, i can *reach* j if there exists a sequence of states $i = i_o,\ldots,i_p = j$ such that $q_{i_n i_{n+1}} \geq 0$ for all $n = 0,\ldots,p-1$; if $U(i_n) \leq E$ (a nonnegative number) for all $n = 0,\ldots,p$ then we shall say that i can *reach* j *at height* E. We shall say that the annealing chain is *strongly irreducible* if i can reach j for all $i,j \in \Sigma$. Clearly, strong irreducibility is equivalent to Q irreducible, but we introduce strong irreducibility to conform with Hajek's notation. We shall also say that the annealing chain is *weakly reversible* if for every $E > 0$, i can reach j at energy E iff j can reach i at energy E, for all $i,j \in \Sigma$. Let S_m be the states that are local minima of U(·), i.e.,

$$S_m = \{i \in \Sigma : \; U(i) \leq U(j) \quad \forall\, j \in N(i)\} \; .$$

For each $i \in S_m \backslash S$ let $\Delta(i)$ be the smallest number E such that i can reach some $j \in \Sigma$ with $U(j) < U(i)$ at height $U(i) + E$. $\Delta(i)$ is the "depth" of the local (but not global) minimum i. Let

$$\Delta^* = \max_{i \in S_m \backslash S} \Delta(i) \; . \qquad (2.10)$$

Here is Hajek's result:

Theorem 2.2 (Hajek [14]) Assume that the annealing chain is strongly irreducible and weakly reversible. Let $T_k \downarrow 0$. Then

$$\lim_{k \to \infty} P\{\xi_k \in S\} = 1 \qquad (2.11)$$

iff

$$\sum_{k=1}^{\infty} \exp\left(-\frac{\Delta^*}{T_k}\right) = \infty \; . \qquad (2.12)$$

Remark If $T_k = c/\log k$ then (2.12) and hence (2.11) holds iff $c \geq \Delta^*$. For this reason Δ^* has been called the *optimal constant* and $T_k = \Delta^*/\log k$ the *optimal schedule*.

We should also mention that Tsitsiklis [30] has proved of generalization of Theorem 2.2 which does not assume weak reversibility, using (and extending) the theory of singularly perturbed Markov chains.

In view of Theorem 2.2 and the refinement in [30], the analysis of the convergence in probability of the annealing algorithm is essentially complete, with the exception that it does not appear that anyone has determined the rate of convergence for optimal or nearly optimal temperature schedules. Recall that Mitra et. al. have shown that (2.9) holds if

$$T_k = \frac{c}{\log k} \quad , \quad c \geq r\,L \; ,$$

but r L is in general much larger than Δ^*. In 2.2, 2.3 we shall analyze the rate of convergence in probability of the annealing algorithm for a special case with two local minima. We will obtain results on the convergence rate for nonparametric temperature schedules (schedules *not* of the form $T_k = c/\log k$) and also for temperature schedules $T_k = c/\log k$ for $c \geq \Delta^*$. We remark that in the latter case with $c = \Delta^*$ there is apparently some interesting and unexpected behavior. Our results are different although consistent with (2.9).

Also in 2.2, 2.3 we shall explore the sample path behavior (as opposed to the ensemble behavior) of the annealing algorithm. We shall give a number of results, the most important of which is conditions such that the annealing chain visit the set S (infinitely often) with probability one. Suppose we let

$$\varsigma_1 = \xi_1$$

$$\varsigma_{k+1} = \begin{cases} \xi_{k+1} & \text{if } U(\xi_{k+1}) < U(\varsigma_k) \\ \varsigma_k & \text{else .} \end{cases}$$

Note that if $\{\xi_k\}$ visits S with probability one then $\{\varsigma_k\}$ traps in S with probability one, and furthermore no additional evaluations of U(·) are required to compute $\{\varsigma_k\}$ over what are required to simulate $\{\xi_k\}$. Hence by just doubling the memory requirements and keeping track of $\{\varsigma_k\}$, it seems sufficient to show that $\{\xi_k\}$ visit S with probability one rather than converge to S in probability. Now it might be imagined that the conditions on the temperature schedule under which $\{\xi_k\}$ visits S with probability one are weaker than those under which $\{\xi_k\}$ converges to S in probability. However, the proof of Theorem 2.2 shows that (assuming strong irreducibility and weak reversibility) $\{\xi_k\}$ visits S with probability one iff (2.12) holds. From this point of view our result does not offer anything new; infact the temperature schedules we consider are not even optimal. However, we believe our result is important in the following sense. In Chapter 3 we extend the annealing algorithm to general state spaces. It turns out that our result on the finite state annealing chain visiting S infinitely often with probability one can also be extended, essentially under the condition that the state space be a compact metric space and the energy function be continuous. It is not clear whether convergence to S in probability can be shown in such a general setting; the methods used to analyze the finite state case (quasiequilibrium distributions, large deviations and perturbation theory) do not seem directly applicable.

Finally, in 2.4 we give a modification of the annealing algorithm which allows for noisy measurements of the energy function and examine its convergence.

2.2 Asymptotic Analysis of a Class of Nonstationary Markov Chains

In this Section we analyze the asymptotic properties of a certain class of nonstationary (finite state) Markov chains. These chains will have the property that their 1-step transition probabilities will satisfy bounds similar to those satisfied by the d-step transition probabilities of the annealing chain. The results of this Section will be used in 2.3 to deduce corresponding asymptotic properties of the annealing chain.

We shall consider the following class of Markov chains. Let Σ be a finite set. Let $\alpha_{ij}, \beta_{ij} \in [0,\infty]$ for $i,j \in \Sigma$, and $\{\theta_k\}$ a sequence of real numbers with $0 < \theta_k \leq 1$. Let $\{\xi_k\}$ be a Markov chain with state space Σ and 1-step transition matrices $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ with the following property: there exists positive numbers A, B such that

$$p_{ij}^{(k,k+1)} \geq A\ \theta_k^{\alpha_{ij}} \tag{2.13}$$

$$p_{ij}^{(k,k+1)} \leq B\ \theta_k^{\beta_{ij}} \tag{2.14}$$

for all $i,j \in \Sigma$. Actually, we shall assume that (2.13) and/or (2.14) hold depending on the result we wish to prove.

2.2.1 Convergence in Probability and Rate of Convergence for a Three State System

We now establish the convergence in probability and rate of convergence of a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.13) and (2.14) for a special case with $|\Sigma| = 3$. In 2.3.2 we shall apply this result to the annealing chain with an energy function which has two local minima. It will be useful here to consider the more detailed bounds

$$A_{ij}\theta_k^{\alpha_{ij}} \leq p_{ij}^{(k,k+1)} \leq B_{ij}\theta_k^{\beta_{ij}} \qquad \forall\ i,j \in \Sigma\ , \tag{2.15}$$

where A_{ij}, B_{ij} are positive constants. Here is our theorem.

Theorem 2.3 Let $\Sigma = \{1,2,3\}$ and assume that (2.15) holds. Let

$$\begin{aligned} a &= \max\{\alpha_{21}, \alpha_{31}\} < \infty\ , \\ b &= \min\{\beta_{12}, \beta_{13}\} > a\ , \\ \gamma &= b - a\ , \\ \delta &= \begin{cases} \min\{A_{21}, A_{31}\} & \text{if } \alpha_{21} = \alpha_{31} \\ A_{21} & \text{if } \alpha_{21} > \alpha_{31} \\ A_{31} & \text{if } \alpha_{21} < \alpha_{31}\ . \end{cases} \end{aligned}$$

(a) Suppose that $\theta_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^a = \infty\ . \tag{2.16}$$

Then

$$\lim_{k\to\infty} P\{\xi_k = 1\} = 1\ .$$

(b) Suppose (more strongly) that $\theta_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that

$$\sum_{n = k\cdot\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \to \infty \quad \text{as } k\to\infty\ , \tag{2.17}$$

$$\sup_k \frac{\theta_{k\cdot\epsilon_k}}{\theta_k} < \infty\ . \tag{2.18}$$

Then

$$P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma) \quad \text{as} \quad k \to \infty .$$

The proof of Theorem 2.3 will require the following lemmas.

Lemma 2.1 Let $\{s_k\}$ be a sequence of positive numbers with $s_k \to 0$ and

$$\sum_{k=1}^{\infty} s_k = \infty .$$

Then

$$\sum_{k=1}^{\infty} s_k \prod_{n=1}^{k-1} (1 - s_n) < \infty .$$

Proof Let

$$S_k = \sum_{n=1}^{k} s_k .$$

Now since $s_k \to 0$ and $S_k \to \infty$ we have

$$\exp(-S_{k-1}) = \exp(s_k) \exp(-S_k) \le \frac{c}{S_k^2}$$

for some constant c. Hence

$$\begin{aligned}\sum_{k=1}^{\infty} s_k \prod_{n=1}^{k-1} (1 - s_n) &\le \sum_{k=1}^{\infty} s_k \exp(-S_{k-1}) \\ &\le c \cdot \sum_{k=1}^{\infty} \frac{s_k}{S_k^2} \\ &< \infty\end{aligned}$$

where the convergence of the last series follows from the Abel-Dini Theorem [20, p. 290]. □

Lemma 2.2 Let $b > a > 0$ and assume that $\theta_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^a = \infty . \tag{2.19}$$

Then

$$\lim_{k \to \infty} \sum_{m=1}^{k} \theta_m^b \prod_{n=m+1}^{k} (1 - \theta_n^a) = 0 .$$

Proof Let

$$p_k = \sum_{m=1}^{k} \theta_m^b \prod_{n=m+1}^{k} (1 - \theta_n^a) .$$

Let $s_k = \theta_k^a$. Then for $K \in \mathbb{N}$

$$\begin{aligned}p_k &= \sum_{m=1}^{k} s_m^{b/a} \prod_{n=m+1}^{k} (1 - s_n) \\ &\le K \cdot s_1^{b/a} \prod_{n=K}^{k} (1 - s_n) + \theta_{K+1}^\gamma \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1 - s_n) \quad \forall\, k \ge K ,\end{aligned}$$

where $\gamma = b - a > 0$. Hence

$$\limsup_{k \to \infty} p_k \le \theta_{K+1}^\gamma \sup_k \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1 - s_n) \tag{2.20}$$

since

$$\prod_{n=K}^{\infty} (1 - s_n) = \prod_{n=K}^{\infty} (1 - \theta_n^a) = 0$$

which follows from (2.19). Now

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1 - s_n) = \sum_{m=1}^{k} s_m \prod_{n=1}^{m-1} (1 - s_n)$$

which is established by induction on k. Hence by Lemma 2.1

$$\sup_k \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1 - s_n) < \infty . \tag{2.21}$$

Combining (2.20), (2.21) and letting $K \to \infty$ (so that $\theta_{K+1}^\gamma \to 0$) gives $p_k \to 0$ as required. □

Lemma 2.3 Let $b > a > 0$, $\gamma = b - a$, and assume that $\theta_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 0$ such that

$$\sup_k \frac{\theta_{k \cdot \epsilon_k}}{\theta_k} < \infty . \tag{2.22}$$

Then

$$\sum_{m=k\cdot\epsilon_k}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_m^a) = O(\theta_k^\gamma) \quad \text{as } k\to\infty$$

Proof Let

$$p_k = \sum_{m=k\cdot\epsilon_k}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_n^a)\,.$$

Let $s_k = \theta_k^a$. Then

$$\begin{aligned} p_k &= \sum_{m=k\cdot\epsilon_k}^{k} s_m^{b/a} \prod_{n=m+1}^{k} (1-s_n) \\ &\le \theta^\gamma_{k\cdot\epsilon_k} \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n)\,. \end{aligned} \tag{2.23}$$

Now

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) = \sum_{m=1}^{k} s_m \prod_{n=1}^{m-1} (1-s_n)$$

which is established by induction on k. Hence by Lemma 2.1 there exists a constant c_1 such that

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) \le c_1\,. \tag{2.24}$$

Also from (2.22) there exists a constant c_2 such that

$$\theta^\gamma_{k\cdot\epsilon_k} \le c_2 \cdot \theta_k^\gamma\,. \tag{2.25}$$

Combining (2.23)-(2.25) gives $p_k = O(\theta_k^\gamma)$ as required. □

Proof of Theorem 2.3

(a) Define the events

$$C_{m,n} = \bigcap_{k=m}^{n} \{\xi_k \in \{2,3\}\} \qquad \forall\, n \ge m\ , \tag{2.26}$$

$$D_{m,n} = \{\xi_m = 1\} \cap C_{m+1,n} \qquad \forall\, n > m\,. \tag{2.27}$$

Then

$$\{\xi_k \in \{2,3\}\} = C_{1,k} \cup \bigcup_{m=1}^{k-1} D_{m,k}$$

and

$$P\{\xi_k \in \{2,3\}\} = PC_{1,k} + \sum_{m=1}^{k-1} PD_{m,k}\,. \tag{2.28}$$

Now using the lower bound in (2.15) and the Markov property, for $i \in \{2,3\}$

$$\begin{aligned} P\{C_{m,k} | \xi_m = i\} &\le P\{\xi_m = i\} \cdot \prod_{n=m}^{k-1} \max_{j=2,3} P\{\xi_{n+1} = 1 | \xi_n = j\} \\ &\le \prod_{n=m}^{k-1} \left(1 - \min_{j=2,3} P\{\xi_{n+1} \in \{2,3\} | \xi_n = j\}\right) \\ &\le \prod_{n=m}^{k-1} \left(1 - \min_{j=2,3} A_{j1} \theta_n^{\alpha_{j1}}\right) \\ &\le c_1 \cdot \prod_{n=m}^{k-1} \left(1 - \delta\, \theta_n^a\right) \qquad \forall\, k > m\,, \end{aligned} \tag{2.29}$$

for some constant c_1. Also, using the upper bound in (2.15), the Markov property, and (2.29)

$$\begin{aligned} PD_{m,k} &= \sum_{i=2,3} P\{\xi_m = 1\}\, p_{1i}^{(m,m+1)}\, P\{C_{m+1,k} | \xi_{m+1} = i\} \\ &\le 2 \cdot \max_{i=2,3} B_{1i}\, \theta_m^{\beta_{1i}} \cdot c_1 \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &\le c_2 \cdot \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \qquad \forall\, k > m\,, \end{aligned} \tag{2.30}$$

for some constant c_2. Hence from (2.29) and (2.16)

$$\begin{aligned} \lim_{k\to\infty} PC_{1,k} &\le c_1 \cdot \prod_{n=1}^{\infty} (1 - \delta\, \theta_n^a) \\ &= 0\,, \end{aligned} \tag{2.31}$$

and from (2.30) and Lemma 2.2

$$\begin{aligned} \lim_{k\to\infty} \sum_{m=1}^{k-1} PD_{m,k} &\le \lim_{k\to\infty} c_2 \cdot \sum_{m=1}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &= 0\,. \end{aligned} \tag{2.32}$$

Combining (2.28), (2.31), and (2.32) gives $P\{\xi_k = 1\} \to 1$ as required.

(b) Define $C_{m,n}$, $D_{m,n}$ as in (2.26), (2.27). Then

$$\{\xi_k \in \{2,3\}\} = C_{k^\bullet\epsilon_k, k} \cup \bigcup_{m=k^\bullet\epsilon_k}^{k-1} D_{m,k}$$

and

$$P\{\xi_k \in \{2,3\}\} = PC_{k^\bullet\epsilon_k, k} + \sum_{m=k^\bullet\epsilon_k}^{k-1} PD_{m,k} \; . \tag{2.33}$$

From (2.29) we have

$$\begin{aligned} PC_{k^\bullet\epsilon_k, k} &\le c_1 \prod_{n=k^\bullet\epsilon_k}^{k-1} (1 - \delta\, \theta_n^a) \\ &\le c_1 \exp\left(- \sum_{n=k^\bullet\epsilon_k}^{k-1} \delta\, \theta_n^a\right) \\ &= c_1 \exp(\delta\theta_k^a) \exp\left(-\delta\left(\sum_{n=k^\bullet\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k\right)\right) \theta_k^\gamma \\ &= o(\theta_k^\gamma) \qquad \text{as } k \to \infty \, , \end{aligned} \tag{2.34}$$

where the last equality follows from $\theta_k^a \to 0$ and (2.17). From (2.30) and Lemma 2.2

$$\begin{aligned} \sum_{m=k^\bullet\epsilon_k}^{k-1} PD_{m,k} &\le c_2 \sum_{m=k^\bullet\epsilon_k}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &= O(\theta_k^\gamma) \qquad \text{as } k \to \infty \, . \end{aligned} \tag{2.35}$$

Combining (2.33)-(2.35) gives $P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma)$ as required. □

The following corollary considers a choice of $\{\theta_k\}$ which will be seen to correspond to a temperature schedule $T_k = c/\log k$ for the annealing algorithm.

Corollary 2.1 Let Σ, a, b, γ, and δ be given as in Theorem 2.3. Assume that

$$\theta_k = \frac{1}{k^{1/c}}$$

where c is a positive constant.

(a) If $c \ge a$ then

$$\lim_{k \to \infty} P\{\xi_k = 1\} = 1 \; .$$

(b) If $c > a$ then

$$P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma) \qquad \text{as } k \to \infty \, .$$

(c) If $c = a$ then

$$P\{\xi_k = 1\} = \begin{cases} 1 + O(\theta_k^\gamma) & \text{if } \gamma < \delta \\ 1 + O(\theta_k^\gamma \log k) & \text{if } \gamma = \delta \\ 1 + O(\theta_k^\delta) & \text{if } \gamma > \delta \, , \end{cases} \qquad \text{as } k \to \infty \, .$$

Proof We shall assume that $c = 1$; the general case follows easily.

(a) If $a \le 1$ then

$$\sum_{k=1}^{\infty} \theta_k^a = \sum_{k=1}^{\infty} \frac{1}{k^a} = \infty$$

and Theorem 2.3(a) applies.

(b) Suppose $a < 1$. To apply Theorem 2.3(b) we must construct a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that conditions (2.17), (2.18) are satisfied. Fix $0 < \eta < 1 - a$ and let

$$\epsilon_k = 1 - \frac{1}{k^\eta} \qquad \text{(k large)} \; .$$

Then for sufficiently large k

$$\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a = \sum_{n=k(1-k^{-\eta})}^{k} \frac{1}{n^a}$$

$$\geq \int_{k(1-k^{-\eta})}^{k} \frac{1}{x^a}\,dx$$

$$\geq \eta k^{1-a-\eta}$$

after evaluating the integrel and applying the Mean Value Theorem. Hence

$$\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta}\log\theta_k \geq \eta k^{1-a-\eta} - \frac{\gamma}{\delta}\log k \to \infty \quad \text{as} \quad k\to\infty\,,$$

and consequently (2.17) is satisfied. (2.18) is also satisfied. Hence Theorem 2.3 (b) applies.

(c) Suppose $a = 1$. It is not apparent in this case how to construct the $\{\epsilon_k\}$ sequence which is necessary to apply Theorem 2.3 (b). However, we can directly use (2.28)-(2.30) to get the desired estimate of $P\{\xi_k = 1\}$. So, from (2.28)

$$P\{\xi_k \in \{2,3,\}\} = PC_{1,k} + \sum_{m=1}^{k-1} PD_{m,k}\,. \tag{2.36}$$

Now from (2.29)

$$PC_{1,k} \leq c_1 \prod_{n=1}^{k-1} (1 - \delta\,\theta_n^a)$$

$$\leq c_1 \exp\left(-\delta \sum_{n=1}^{k-1} \frac{1}{n}\right)$$

$$\leq c_1 \exp\left(-\delta \int_1^k \frac{1}{x}\,dx\right)$$

$$= \frac{c_1}{k^\delta}\,. \tag{2.37}$$

Also, from (2.30)

$$\sum_{m=1}^{k-1} PD_{m,k} \leq c_2 \sum_{m=1}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\,\theta_n^a)$$

$$\leq c_2 \sum_{m=1}^{k-1} \frac{1}{m^b} \exp\left(-\delta \sum_{n=m+1}^{k-1} \frac{1}{n}\right)$$

$$\leq c_2 \sum_{m=1}^{k-1} \frac{1}{m^b} \exp\left(-\delta \int_{m+1}^{k} \frac{1}{x}\,dx\right)$$

$$= \frac{c_2}{k^\delta} \sum_{m=1}^{k-1} \frac{1}{m^b} \cdot (m+1)^\delta$$

$$\leq \frac{2c_2}{k^\delta} \sum_{m=1}^{k-1} \frac{1}{m^{b-\delta}}$$

since $(p+q)^r \leq p^r + q^r$ for $p,q \geq 0$, $0 \leq r \leq 1$. Since $\delta \leq 1$ (use $\theta_1 = 1$ in (2.15)) and $b > a = 1$ we have $b - \delta > 0$. Hence

$$\sum_{m=1}^{k-1} PD_{m,k} \leq \frac{2c_2}{k^\delta}\left(1 + \int_1^k \frac{1}{x^{b-\delta}}\,dx\right)$$

$$= \begin{cases} c_3 \cdot \dfrac{1}{k^\delta} + c_4 \cdot \dfrac{1}{k^\gamma} & \text{if } \gamma \neq \delta \\[2ex] 2c_2\,(1 + \log k) \cdot \dfrac{1}{k^\gamma}\,. & \text{if } \gamma = \delta \end{cases} \tag{2.38}$$

where c_3, c_4 are suitable constants. Combining (2.36)-(2.38) completes the proof of part (c). □

2.2.2 Sample Path Analysis

We now analyze the sample path behavior of a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.13) and/or (2.14). We shall give (different) conditions such that

- $\{\xi_k\}$ visits a subset of Σ (infinitely often) with probability one
- $\{\xi_k\}$ visits a subset of Σ with probability less then one
- $\{\xi_k\}$ converges to (i.e. eventually stays in) a subset of Σ with probability one

It will be convenient to use the following notation. For J a subset of Σ define the events

$$\{\xi_k \in J \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k>n} \{\xi_k \in J\} ,$$

$$\{\xi_k \in J \text{ a.a.}\} = \bigcup_{n=1}^{\infty} \bigcap_{k>n} \{\xi_k \in J\}$$

(i.o. and a.a. stand for *infinitely often* and *almost always,* respectively).

Our first theorem gives sufficient conditions under which $\{\xi_k\}$ visits a subset of Σ infinitely often with probability one.

Theorem 2.4 Assume that (2.13) holds. Let J be a subset of Σ and

$$a = \max_{i \in \Sigma \setminus J} \min_{j \in J} \alpha_{ij} < \infty . \tag{2.39}$$

Suppose

$$\sum_{k=1}^{\infty} \theta_k^a = \infty . \tag{2.40}$$

Then $P\{\xi_k \in J \text{ i.o.}\} = 1$.

Proof Let $I = \Sigma \setminus J$. Using (2.13) and the Markov property

$$\begin{aligned} P \bigcap_{k=m}^{n} \{\xi_k \in I\} &\leq P\{\xi_m \in I\} \prod_{k=m}^{n-1} \max_{i \in I} P\{\xi_{k+1} \in I | \xi_k = i\} \\ &\leq \prod_{k=m}^{n-1} \left(1 - \min_{i \in I} P\{\xi_{k+1} \in J | \xi_k = i\}\right) \\ &\leq \prod_{k=m}^{n-1} \left(1 - \min_{i \in I} \sum_{j \in J} A\, \theta_k^{\alpha_{ij}}\right) \\ &\leq \prod_{k=m}^{n-1} (1 - A\, \theta_k^a) \qquad \forall\, n > m . \end{aligned}$$

Hence

$$P \bigcap_{k=m}^{\infty} \{\xi_k \in I\} \leq \prod_{k=m}^{\infty} (1 - A\, \theta_k^a) = 0 \qquad \forall\, m ,$$

where the divergence of the infinite product follows from the divergence of the infinite sum (2.40), and the Theorem follows. □

The next theorem gives sufficient conditions that $\{\xi_k\}$ visits a subset of Σ with probability strictly less than one, at least starting from certain initial states. Let $P_i(\cdot) = P\{\cdot | \xi_1 = i\}$ for all $i \in \Sigma$.

Theorem 2.5 Assume that (2.14) holds. Let J be a subset of Σ and

$$b = \max_{K \supset J} \min_{i \in \Sigma \setminus K} \min_{j \in K} \beta_{ij} > 0 . \tag{2.41}$$

Suppose that $\theta_k \longrightarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^b < \infty . \tag{2.42}$$

Then there exists an $i \in \Sigma$ such that

$$P_i \bigcup_{k=1}^{\infty} \{\xi_k \in J\} < 1 .$$

Proof Let J^* be a subset of Σ containing J which obtains the maximum in (2.41) and let $I^* = \Sigma \setminus J^*$. Let $\ell \in I^*$. Using (2.14) and the Markov property

$$\begin{aligned} P_\ell \bigcap_{k=1}^{n} \{\xi_k \in I^*\} &\geq \prod_{k=1}^{n-1} \min_{i \in I^*} P\{\xi_{k+1} \in I^* | \xi_k = i\} \\ &= \prod_{k=1}^{n-1} \left(1 - \max_{i \in I^*} P\{\xi_{k+1} \in J^* | \xi_k = i\}\right) \\ &\geq \prod_{k=1}^{n-1} \left(1 - \max_{i \in I^*} \sum_{j \in J^*} B\, \theta_k^{\beta_{ij}}\right) \\ &\geq \prod_{k=1}^{n-1} (1 - B|J^*|\theta_k^b) \qquad \forall n . \end{aligned}$$

Hence

$$P_\ell \bigcap_{k=1}^{\infty} \{\xi_k \in I^*\} \geq \prod_{k=1}^{\infty} (1 - B|J^*|\theta_k^b) > 0$$

where the convergence of the infinite product follows from the convergence of the infinite series (2.42), and the Theorem follows. □

Finally, we give a theorem which gives conditions such that $\{\xi_k\}$ converges to a subset of Σ with probability one, provided it visits that subset infinitely often with probability one.

Theorem 2.6 Assume (2.14) holds. Let J be a subset of Σ and

$$c = \min_{j \in J} \min_{i \in \Sigma \setminus J} \beta_{ji} . \tag{2.43}$$

Suppose $\theta_k \longrightarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^c < \infty . \tag{2.44}$$

Under these conditions, if $P\{\xi_k \in J \text{ i.o.}\} = 1$ then $P\{\xi_k \in J \text{ a.a.}\} = 1$.

Proof Let $I = \Sigma \setminus J$. Using (2.14) and the Markov property

$$\begin{aligned} P\{\xi_k \in J,\ \xi_{k+1} \in I\} &\le P\{\xi_k \in J\} \max_{j \in J} P\{\xi_{k+1} \in I | \xi_k = j\} \\ &\le \max_{j \in J} \sum_{i \in I} B\ \theta_k^{\beta_{ji}} \\ &\le B|I|\ \theta_k^c . \end{aligned}$$

Hence

$$\sum_{k=1}^{\infty} P\{\xi_k \in J,\ \xi_{k+1} \in I\} \le \sum_{k=1}^{\infty} B|I|\ \theta_k^c < \infty$$

by (2.44). Hence by the "first" Borel-Cantelli Lemma we must have $P\{\xi_k \in J,\ \xi_{k+1} \in I \text{ i.o.}\} = 0$, and it follows that $P\{\xi_k \in J \text{ a.a}\} = 1$ whenever $P\{\xi_k \in J \text{ i.o.}\} = 1$. □

2.3 Convergence of the Annealing Algorithm

In this Section we apply the results of 2.2 to obtain asymptotic properties of the annealing algorithm. Throughout this Section (2.3) we use the notation introduced in 2.1.

2.3.1 Bounds on the Transition Probabilities of the Annealing Chain

In order to apply the results in 2.2 we need to obtain bounds on the d-step transition probabilities $p_{ij}^{(k,k+d)}$ of the annealing chain $\{\xi_k\}$. Toward this end we make the following definitions. For every $i,j \in \Sigma$ and $d \in \mathbb{N}$ let $\Lambda_d(i,j)$ be the subset of Σ^{d+1} such that $(i = i_o,\dots,i_d = j) \in \Lambda_d(i,j)$ if

$$p_{i_n i_{n+1}}^{(k,k+1)} > 0 \qquad \forall\, n = 0,\dots,d-1 ,$$

for any $k \in \mathbb{N}$ (this definition is valid since $\{T_k\}$ is a positive sequence and so $p_{ij}^{(k,k+1)} > 0$ for all k whenever $p_{ij}^{(k,k+1)} > 0$ for some k). Hence $\Lambda_d(i,j)$ is just

the set of possible d-step transitions from state i to state j for the annealing chain. An alternate characterization of $\Lambda_d(i,j)$ is as follows: $(i = i_o,\dots,i_d = j) \in \Lambda_d(i,j)$ iff for every $n = 0,\dots,d-1$ *either*

(i) $q_{i_n i_{n+1}} > 0$ *or*

(ii) $i_{n+1} = i_n$ and $q(i_n,\ell) > 0$ for some $\ell \in \Sigma$ with $U(\ell) > U(i_n)$.

This characterization follows easily from (2.2).

For each $d \in \mathbb{N}$ let

$$U_d(i_o,\dots,i_d) = \sum_{n=0}^{d-1} \max\{0, U(i_{n+1}) - U(i_n)\} ,$$

for all $i_o,\dots,i_d \in \Sigma$, and

$$V_d(i,j) = \inf_{\lambda \in \Lambda_d(i,j)} U_d(\lambda) , \tag{2.45}$$

$$V(i,j) = \inf_d V_d(i,j) \tag{2.46}$$

for all $i,j \in \Sigma$. Note that the infinum in (2.46) is obtained for $d \le |\Sigma|$. Also note that

$$V(i,j) \le V(i,\ell) + V(\ell,j) \qquad \forall\, i,j,\ell \in \Sigma . \tag{2.47}$$

We shall call $V_d(i,j)$ the *d-step transition energy from* i *to* j, and V(i,j) the *transition energy from* i *to* j.

The following theorem gives upper and lower bounds on the d-step transition probabilities of the annealing chain in terms of the d-step transition energy.

Theorem 2.7 Let $\{T_k\}$ be monotone nonincreasing and $d \in \mathbb{N}$ Then there exists positive numbers A, B such that

$$A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \le p_{ij}^{(k,k+d)} \le B \exp\left[-\frac{V_d(i,j)}{T_k}\right] \qquad \forall\, i,j \in \Sigma . \tag{2.48}$$

Proof We prove the lower bound in (2.48); the upper bound is similar. Let

$$r_k(i,j) = \begin{cases} q_{ij} & \text{if } j \ne i \\ p_{ii}^{(k,k+1)} & \text{if } j = i \end{cases} \tag{2.49}$$

for all $i,j \in \Sigma$, and

$$\tilde{r}_k(i_o,\ldots,i_d) = \prod_{n=0}^{d-1} r_k(i_n,i_{n+1}) \,, \tag{2.50}$$

$$\tilde{r}(i_o,\ldots,i_d) = \inf_k \tilde{r}_k(i_o,\ldots,i_d) \,, \tag{2.51}$$

for all $i_o,\ldots,i_d \in \Sigma$. If $\lambda \in \Sigma^{d+1}$ then since $\{T_k\}$ is nonincreasing $\{\tilde{r}_k(\lambda)\}$ is nondecreasing and so $\tilde{r}(\lambda) = \tilde{r}_1(\lambda)$ obtains the infimum. Note that $\tilde{r}(\lambda) > 0$ for all $\lambda \in \Lambda_d(i,j)$, $i,j \in \Sigma$.

Now from (2.2) and (2.49)-(2.51) we have that

$$p_{ij}^{(k,k+d)} \geq \sum_{\lambda \in \Lambda_d(i,j)} \tilde{r}(\lambda) \exp\left[-\frac{U_d(\lambda)}{T_{k+d-1}}\right] \qquad \forall\ i,j \in \Sigma \,. \tag{2.52}$$

For each $i,j \in \Sigma$ if $V_d(i,j) < \infty$ let

$$N(i,j) = \{\lambda \in \Lambda_d(i,j) :\ U_d(\lambda) = V_d(i,j)\} \neq \emptyset$$

and set

$$a_{ij} = \sum_{\lambda \in N(i,j)} \tilde{r}(\lambda) > 0 \,;$$

if $V_d(i,j) = \infty$ set $a_{ij} = 1$. Then from (2.52)

$$p_{ij}^{(k,k+d)} \geq A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \qquad \forall\ i,j \in \Sigma$$

where $A = \min_{i,j \in \Sigma} a_{ij} > 0$. □

Remark We note that the proof of Theorem 2.7 is quite trivial, and we would like to point out that our reason for presenting it in detail is for comparison with the (more difficult) proof of the general state analog (Theorem 3.3) to come.

2.3.2 Convergence in Probability and Rate of Convergence for Two Local Minima

We now apply the results of 2.2.1 to establish the convergence in probability and rate of convergence for an annealing chain $\{\xi_k\}$ with an energy function $U(\cdot)$ with two local minima. We shall consider the following example in detail:

$$\text{(H)} \qquad \begin{aligned} &\Sigma = \{1,2,3\} \\ &U(1) < U(3) < U(2) \\ &q_{12},\ q_{21},\ q_{23},\ q_{32} > 0 \\ &q_{ij} = 0 \qquad \text{otherwise}\,. \end{aligned}$$

The annealing chain corresponding to (H) is illustrated by the transition diagram in Figure 2.1. Let

$$\begin{aligned} a &= U(2) - U(3) \,, \\ b &= U(2) - U(1) \,, \\ \gamma &= U(3) - U(1) \,, \\ \delta &= q_{32} \cdot q_{21} \,. \end{aligned}$$

Here is our theorem.

Theorem 2.8 Assume the conditions in (H).

(a) Suppose $T_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{a}{T_k}\right) = \infty \,. \tag{2.53}$$

Then

$$\lim_{k \to \infty} P\{\xi_k = 1\} = 1 \,.$$

(b) Suppose (more strongly) that $T_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that

$$\sum_{n=k \cdot \epsilon_k}^{k} \exp\left(-\frac{a}{T_{2k}}\right) - \frac{\gamma}{\delta} \cdot \frac{1}{T_{2k}} \to \infty \qquad \text{as } k \to \infty \,, \tag{2.54}$$

$$\sup_k \left(\frac{1}{T_{2k}} - \frac{1}{T_{2k \cdot \epsilon_k}}\right) < \infty \,. \tag{2.55}$$

Then

$$P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{\gamma}{T_k}\right)\right) \qquad \text{as } k \to \infty \,. \tag{2.56}$$

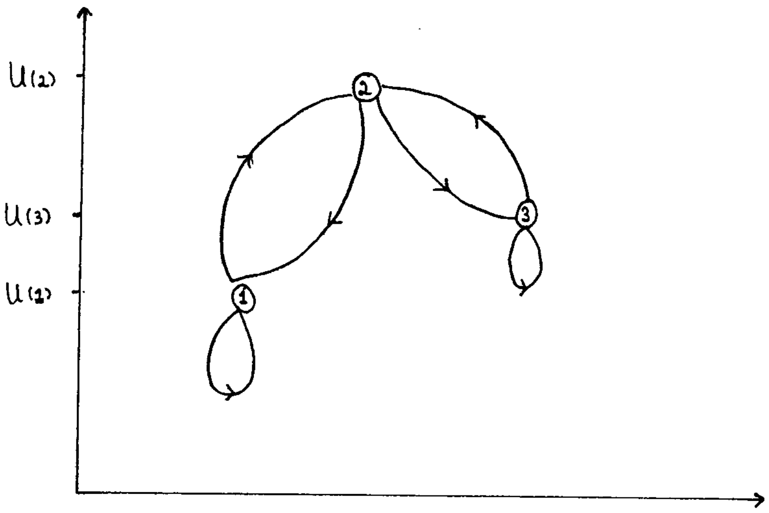

Figure 2.1. Transition Diagram for Annealing Chain with Two Local Minima

Proof Let

$$\eta_k = \xi_{2k} \,, \;\; \varsigma_k = \xi_{2k+1} \,,$$

$$\theta_k = \exp\left(-\frac{1}{T_{2k}}\right), \;\; \tau_k = \exp\left(-\frac{1}{T_{2k+1}}\right).$$

Then $\{\eta_k\}$, $\{\varsigma_k\}$ are Markov chains with 1-step transition matrices $\{R^{(k,k+1)}\} = \{[r_{ij}^{(k,k+1)}]\}$, $\{S^{(k,k+1)}\} = \{[s_{ij}^{(k,k+1)}]\}$, respectively, which satisfy

$$A_{ij}\theta_k^{\alpha_{ij}} < r_{ij}^{(k,k+1)} \le B_{ij}\theta_k^{\beta_{ij}} \,,$$

$$A_{ij}\tau_k^{\alpha_{ij}} \le s_{ij}^{(k,k+1)} \le B_{ij}\tau_k^{\beta_{ij}} \,,$$

for appropriate α_{ij}, β_{ij}, A_{ij}, B_{ij}, and it is clear that these constants may be chosen such that

$$\begin{aligned} a &= \max\{\alpha_{21}, \alpha_{31}\} < \infty \,, \\ b &= \min\{\beta_{12}, \beta_{13}\} > a \,, \\ \gamma &= b - a \,, \\ \delta &= A_{31} \,. \end{aligned}$$

Hence we are (almost) in a position to apply Theorem 2.3 to $\{\eta_k\}$ and $\{\varsigma_k\}$.

Suppose that $T_k \downarrow 0$ and (2.53) holds. Since $\{T_k\}$ is nonincreasing, the divergence of the series in (2.53) implies that

$$\sum_{k=1}^{\infty} \theta_k^a = \infty \,, \;\; \sum_{k=1}^{\infty} \tau_k^a = \infty \,.$$

Hence we may apply Theorem 2.3 (a) to $\{\eta_k\}$, $\{\varsigma_k\}$ to get

$$\lim_{k\to\infty} P\{\xi_{2k} = 1\} = \lim_{k\to\infty} P\{\eta_k = 1\} = 1 \,,$$

$$\lim_{k\to\infty} P\{\xi_{2k+1} = 1\} = \lim_{k\to\infty} P\{\varsigma_k = 1\} = 1 \,,$$

and hence

$$\lim_{k\to\infty} P\{\xi_k = 1\} = 1$$

which proves (a).

Suppose that $T_k \downarrow 0$ and (2.54), (2.55) hold. Now (2.54), (2.55) are equivalent to, respectively,

$$\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \longrightarrow \infty \qquad \text{as } k\to\infty , \tag{2.57}$$

$$\sup_k \frac{\theta_{k\cdot\epsilon_k}}{\theta_k} < \infty . \tag{2.58}$$

Hence we may apply Theorem 2.3 (b) to $\{\eta_k\}$ to get

$$P\{\xi_{2k}=1\} = P\{\eta_k = 1\} = 1 + O\left(\exp\left(-\frac{a}{T_{2k}}\right)\right) \qquad \text{as } k\to\infty . \tag{2.59}$$

We make the following

Claim

$$\sum_{n=k\cdot\epsilon_k}^{k} \tau_n^a + \frac{\gamma}{\delta} \log \tau_k \longrightarrow \infty \qquad \text{as } k\to\infty , \tag{2.60}$$

$$\sup_k \frac{\tau_{k\cdot\epsilon_k}}{\tau_k} < \infty . \tag{2.61}$$

Suppose the Claim is true. Then we may apply Theorem 2.3 (b) to $\{\varsigma_k\}$ to get

$$P\{\xi_{2k+1}=1\} = P\{\varsigma_k = 1\} = 1 + O\left(\exp\left(\frac{a}{T_{2k+1}}\right)\right) \qquad \text{as } k\to\infty , \tag{2.62}$$

and it would follow from (2.59) and (2.62) that

$$P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{a}{T_k}\right)\right) \qquad \text{as } k\to\infty ,$$

which would prove (b). It remains to prove the Claim.

Proof of Claim We first show that

$$\sup_k \left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}}\right) < \infty . \tag{2.63}$$

Now

$$\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}} < \frac{1}{T_{2(k+1)}} - \frac{1}{T_{2(k+1)\cdot\epsilon_k}} + \frac{1}{T_{2(k+1)\cdot\epsilon_k}} - \frac{1}{T_{2k}} .$$

In view of (2.55) it is enough to show

$$\sup_k \left(\frac{1}{T_{2(k+1)\epsilon_k}} - \frac{1}{T_{2k}}\right) < \infty ,$$

or since $\{T_k\}$ is nonincreasing,

$$2(k+1)\cdot\epsilon_k \le 2k \qquad (k \text{ large}) .$$

Suppose this last inequality is not satisfied. Then there exists a sequence $\{k_n\}$ of positive integers with $k_n \uparrow \infty$ and

$$k_n\epsilon_{k_n} > k_n - \epsilon_{k_n} > k_n - 1 .$$

Hence

$$\liminf_{k\to\infty} \left(\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k\right)$$

$$\le \lim_{n\to\infty} \left(\theta_{k_n}^a + \frac{\gamma}{\delta} \log \theta_{k_n}\right)$$

$$= -\infty$$

which contradicts (2.57). Hence (2.63) must be true. Now using (2.63) we obtain

$$\sup_k \left(\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a - \sum_{n=k\cdot\epsilon_k}^{k} \tau_n^a\right) < \sup_k \left(\theta_{k\cdot\epsilon_k} - \theta_{k+1}\right) < \infty ,$$

$$\sup_k (\log \theta_k - \log \tau_k) = \sup_k \left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}}\right) < \infty ,$$

$$\sup_k \left(\frac{\tau_{k\cdot\epsilon_k}}{\tau_k} - \frac{\theta_{k\cdot\epsilon_k}}{\theta_k}\right) < \sup_k \exp\left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}}\right) < \infty ,$$

and (2.60), (2.61) now follow from (2.57), (2.58). This completes the proof of the Claim and hence the Theorem. □

Corollary 2.2 Assume the conditions in (H). Let

$$T_k = \frac{c}{\log k} \qquad (k \text{ large})$$

where c is a positive constant.

(a) If $c \geq a$ then

$$\lim_{k \to \infty} P\{\xi_k = 1\} = 1 .$$

(b) If $c > a$ then

$$P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{\gamma}{T_k}\right)\right) \qquad \text{as } k \to \infty . \tag{2.64}$$

(c) If $c = a$ then

$$P\{\xi_k = 1\} = \begin{cases} 1 + O\left(\exp\left(-\frac{\gamma}{T_k}\right)\right) & \text{if } \gamma < \bar{\delta} \\ 1 + O\left(\exp\left(-\frac{\gamma}{T_k} + \log\log k\right)\right) & \text{if } \gamma = \bar{\delta} \\ 1 + O\left(\exp\left(-\frac{\bar{\delta}}{T_k}\right)\right) & \text{if } \gamma > \bar{\delta}, \text{ as } k \to \infty , \end{cases} \tag{2.65}$$

where $\bar{\delta} = \delta/2$.

Proof We may use Corollary 2.1 by appropriately identifying variables. Let

$$\eta_k = \xi_{2k} , \quad \varsigma_k = \xi_{2k+1} ,$$

and

$$\theta_k = \frac{1}{k^{1/c}} .$$

Then $\{\eta_k\}$, $\{\varsigma_k\}$ are Markov chains with one step transition matrices $\{R^{(k,k\neg1)}\} = \{[r_{ij}^{(k,k+1)}]\}$, $\{S^{(k,k+1)}\} = \{[s_{ij}^{(k,k+1)}]\}$, respectively, which satisfy

$$A_{ij}\, \theta_k^{\alpha_{ij}} \leq r_{ij}^{(k,k+1)},\ s_{ij}^{(k,k+1)} \leq B_{ij}\, \theta_k^{\beta_{ij}} \qquad (k \text{ large})$$

for approopriate α_{ij}, β_{ij}, A_{ij}, B_{ij}, and these constants may be chosen such that

$$a = \max\{\alpha_{21},\alpha_{31}\} < \infty ,$$
$$b = \min\{\beta_{12},\beta_{13}\} > a ,$$
$$\gamma = b - a ,$$
$$\bar{\delta} = A_{31} .$$

Hence we may apply Corollary 2.1 (a)-(c) to $\{\eta_k\}$, $\{\varsigma_k\}$ to get the corresponding (a)-(c) here. □

Remarks on Theorem 2.8 and Corollary 2.2

(1) Theorem 2.8 (a) is a simple case of Theorem 2.2 (Hajek's Theorem) since $a = \Delta(2) = \Delta^*$, the optimal constant (see (2.10)).

(2) We compare our results with the rate of convergence (2.9) given by Mitra et. al. First, Theorem 2.8 (b) gives the rate of convergence of $P\{\xi_k = 1\}$ to 1 for nonparametric temperature schedules, in particular schedules *not* of the form $T_k = c/\log k$. This is possible essentially due to the application of the Abel-Dini Theorem on infinite series in the proof of Lemma 2.1. (2.9) is valid only for temperature schedules of the form $T_k = c/\log k$. Second, Corollary 2.2 (b), (c) gives the rate of convergence for temperature schedules of the form $T_k = c/\log k$ for $c \geq a$, whereas (2.9) only holds for $c \geq r\,L = 2\cdot[U(2) - U(1)] > U(2) - U(3) = a$. Furthermore, for $c \geq r\,L$ where (2.9) does hold, (2.64) is in general tighter:

$$\exp\left(-\frac{\gamma}{T_k}\right) = \frac{1}{k^{\beta}} \leq \frac{1}{k^{\min\{\alpha,\beta\}}} .$$

Recall that Mitra et. al. suggest choosing $c \geq r\,L$ such that $\min\{\alpha,\beta\}$ is maximized (see (2.9)). Our results suggest choosing

$$c = \begin{cases} a & \text{if } \gamma \leq \bar{\delta} \\ a + \epsilon & \text{if } \gamma > \bar{\delta} \end{cases}$$

where $0 < \epsilon < a\ [(\gamma/\bar{\delta}) - 1]$ (see (2.64) and (2.65)). We want to stress that (2.9) holds for general $U(\cdot)$ whereas we have not been able to extend Theorem 2.8 and Corollary 2.2 to a $U(\cdot)$ with more than two local minma.

(3) The proof of Theorem 2.8 and Corollary 2.2 (which rely on Theorem 2.3 and Corollary 2.1) show that there are two factors which limit the rate at which $P\{\xi_k = 1\}$ converges to 1. One factor corresponds to the rate at which the annealing chain makes transitions from state 1 to state 3 and back. For temperature schedules of the form $T_k = c/\log k$ this factor dominates for $c > a$ and has a characteristic time scale $1/\gamma$. Note that $\gamma = U(3) - U(1)$

depends only on the energy function $U(\cdot)$. The other factor corresponds to the rate at which the annealing chain makes it first transition from state 3 to state 1. For temperature schedules of the form $T_k = c/\log k$ this factor is only important for $c = a$ and has characteristic time scale $1/\delta$. Note that $\delta = q_{32}q_{21}/2$ does not depend on the energy function $U(\cdot)$. We wonder whether there is some physical significance to all of this.

2.3.3 Sample Path Analysis

We now apply the results of 2.2.2 to analyze the sample path behavior of the annealing chain $\{\xi_k\}$. To avoid trivialities we will need the following assumptions:

(P1) Every $i \in \Sigma \backslash S$ can reach some $j \in S$

(P2) There exists an $i \in \Sigma \backslash S$ such that for every $j \in S$, i can only reach j at height greater than $U(i)$.

The following theorem gives conditions under which the annealing chain $\{\xi_k\}$ visits S infinitely often with probability one. Let

$$V^* = \max_{i \in \Sigma \backslash S} \min_{j \in S} V(i,j) \tag{2.66}$$

Note that (P1) holds iff $V^* < \infty$.

Theorem 2.9 Assume (P1). Let $\{T_k\}$ be monotone nonincreasing and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V^*}{T_k}\right) = \infty . \tag{2.67}$$

Then $P\{\xi_k \in S \text{ i.o.}\} = 1$.

Proof We first show there exists a $d \in \mathbb{N}$ such that

$$V^* = \max_{i \in \Sigma \backslash S} \min_{j \in S} V_d(i,j) . \tag{2.68}$$

For every $i \in \Sigma \backslash S$ there exists a $d_i \in \mathbb{N}$ such that

$$\min_{j \in S} V_{d_i}(i,j) = \min_{j \in S} V(i,j) \le V^* .$$

Let $d^* = \max_{i \in \Sigma \backslash S} d_i$. Now it is easy to see that for every $i \in \Sigma$

$$\min_{j \in S} V_n(i,j) \le \min_{j \in S} V_m(i,j) \qquad \forall\, n \ge m .$$

Hence for every $i \in \Sigma \backslash S$

$$\min_{j \in S} V_{d^*}(i,j) = \min_{n \le d^*} \min_{j \in S} V_n(i,j) \le V^*$$

and (2.68) follows by setting $d = d^*$.

Next, from Theorem 2.7 there exists a positive number A such that

$$p_{ij}^{(k,k+d)} \ge A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \qquad \forall\, i,j \in \Sigma .$$

Let

$$\tilde{\xi}_k = \xi_{kd}, \;\; \theta_k = \exp\left(-\frac{1}{T_{kd+d-1}}\right),$$

and

$$\alpha(i,j) = V_d(i,j) \qquad \forall\, i,j \in \Sigma . \tag{2.69}$$

Then $\{\tilde{\xi}_k\}$ is a Markov chain with 1-step transition matrices $\{\tilde{P}^{(k,k+1)}\} = \{[\tilde{p}_{ij}^{(k,k+1)}]\}$ which satisfy

$$\tilde{p}_{ij}^{(k,k+1)} \ge A\, \theta_k^{\alpha_{ij}} \qquad \forall\, i,j \in \Sigma .$$

Let

$$a = \max_{i \in \Sigma \backslash S} \min_{j \in S} \alpha_{ij} .$$

By (2.68) and (2.69) $a = V^*$. Hence since $\{T_k\}$ is nonincreasing the divergence of the series in (2.67) implies

$$\sum_{k=1}^{\infty} \theta_k^a = \infty .$$

Hence we may apply Theorem 2.4 to $\{\tilde{\xi}_k\}$ with $J = S$ to get $P\{\tilde{\xi}_k \in S \text{ i.o.}\} = 1$ and so $P\{\xi_k \in S \text{ i.o.}\} = 1$. □

Remark Clearly $V^* > \Delta^*$, the optimal constant (see (2.10), (2.66)). Hence (assuming strong irreducibility and weak reversibility) Theorem 2.2 is a much stronger result. However, the importance of Theorem 2.9 is that it can be extended to a general state version of the annealing algorithm under essentially the condition that the state space be a compact metric space and the energy function be continuous. This will be done in Chapter 3.

The next theorem gives conditions under which the annealing chain $\{\xi_k\}$ visits S with probability strictly less than one. Let

$$V_1 = \max_{K \supset S} \min_{i \in \Sigma \backslash K} \min_{j \in S} V(i,j) . \qquad (2.70)$$

Note that (P2) and (2.47) imply $V_1 > 0$.

Theorem 2.10 Assume (P2). Let $T_k \to 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V_1}{T_k}\right) < \infty .$$

Then there exists an $i \in \Sigma$ such that

$$P_i \bigcup_{k=1}^{\infty} \{\xi_k \in S\} < 1 .$$

Proof From Theorem 2.7 there exists a positive number B such that

$$p_{ij}^{(k,k+1)} \leq B \exp\left[-\frac{V(i,j)}{T_k}\right] \qquad \forall\ i,j \in \Sigma .$$

Theorem 2.5 may be applied to $\{\xi_k\}$ in an obvious manner. □

Finally, we give a theorem which gives conditions such that the annealing chain $\{\xi_k\}$ converges to S with probability one, provided it visits S infinitely often with probability one. Let

$$V_2 = \min_{j \in S} \min_{i \in \Sigma \backslash S} V(j,i) . \qquad (2.71)$$

Theorem 2.11 Let $T_k \to 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V_2}{T_k}\right) < \infty .$$

If $P\{\xi_k \in S \text{ i.o.}\} = 1$ then $P\{\xi_k \in S \text{ a.a.}\} = 1$.

Proof From Theorem 2.7 there exits a positive number B such that

$$p_{ij}^{(k,k+1)} \leq B \exp\left[-\frac{V(i,j)}{T_k}\right] \qquad \forall\ i,j \in \Sigma .$$

Theorem 2.6 may be applied to $\{\xi_k\}$ in an obvious manner. □

Remark Theorem 2.2 or 2.9 may be combined with Theorem 2.11 to give conditions under which the annealing chain $\{\xi_k\}$ converges to S with probability one. Note, however, that is is not always possible to do this since it is not in general true that $V_2 > V^*$ or even $V_2 > \Delta^*$ (see (2.10), (2.66), (2.71)).

2.4 Annealing Algorithm with Noisy Energy Measurements

In this Section we consider a modification of the annealing algorithm so as to allow for noisy measurements of the energy differences which are used in selecting successive states. This is important when the energy differences cannot be computed exactly or when it is simply too costly to do so. Using the notation introduced in 2.1 we construct the modified annealing chain as follows. At time k, given the current state is i we select a candidate state j with probability q_{ij}. We assume that the energy difference $U(j) - U(i)$ is measured with (additive) noise, which is independent of states and candidate states at times less than or equal to k, and noise at times less than k. The exponent of the energy difference plus noise is then used to determine whether a transition is made from i to j. More precisely, let $\{w_k\}$ be a sequence of $\mathbb{R}$-valued independent random variables. Construct a Σ-valued discrete-time process $\{\hat{\xi}_k\}$ with $\hat{\xi}_{k+1}$ conditionally independent of $\hat{\xi}_1,\ldots,\hat{\xi}_{k-1}$ and $w_1,\ldots,w_{k-1}$ given $\hat{\xi}_k$ and w_k, and

$$P\{\hat{\xi}_{k+1} = j \,|\, \hat{\xi}_k = i,\ w_k = w\}$$

$$= \begin{cases} q_{ij} \exp\left[-\dfrac{U(j) - U(i) + w}{T_k}\right] & \text{if } U(j) - U(i) + w > 0,\ j \neq i , \\ q_{ij} & \text{if } U(j) - U(i) + w \leq 0,\ j \neq i, \end{cases}$$

for all $i,j \in \Sigma$. It is easy to see that $\{\hat{\xi}_k\}$ is a Markov chain. Let $\{\hat{P}^{(k,k+1)}\} = \{[\hat{p}_{ij}^{(k,k+1)}]\}$ be the 1-step transition matrices for $\{\hat{\xi}_k\}$. Then since w_k is independent of $\hat{\xi}_k$ we have

$$\begin{aligned} \hat{p}_{ij}^{(k,k+1)} &= E\{P\{\hat{\xi}_{k+1} = j \,|\, \hat{\xi}_k,\ w_k\} \,|\, \hat{\xi}_k = i\} \\ &= E\{P\{\hat{\xi}_{k+1} = j \,|\, \hat{\xi}_k = i,\ w_k\}\} \\ &= \int_{\{w > U(i) - U(j)\}} q_{ij} \exp\left[-\frac{U(j) - U(i) + w}{T_k}\right] dF_k(w) \\ &\quad + q_{ij}\, P\{w_k \leq U(i) - U(j)\} \qquad \forall\ j \neq i , \end{aligned} \qquad (2.72)$$

where $F_k(\cdot)$ is the distribution function for w_k. We shall call $\{\hat{\xi}_k\}$ the *annealing chain modified for noisy energy measurements.* In the sequel we shall only consider the case where w_k is Gaussian with mean 0 and variance $\sigma_k^2 > 0$. Hence (2.72) can be written as

$$\hat{p}_{ij}^{(k,k+1)} = \int_{U(i)-U(j)}^{\infty} q_{ij} \exp\left[-\frac{U(j)-U(i)+w}{T_k}\right] dN(0,\sigma_k^2)\,(w)$$

$$+\, q_{ij}\, N(0,\sigma_k^2)\,(-\infty, U(i)-U(j)] \qquad \forall\, j \neq i\,. \tag{2.73}$$

The following theorem shows that if the noise variance goes to zero fast enough then the 1-step transition probabilities for the annealing chain modified for (Gaussian additive) noisy measurements are asymptotically equivalent to the 1-step transition probabilities for the unmodified annealing chain.

Theorem 2.12 If

$$\sigma_k^2 = o(T_k^4) \qquad \text{as } k \to \infty$$

then

$$\hat{p}_{ij}^{(k,k+1)} \sim \begin{cases} q_{ij} \exp\left[-\dfrac{U(j)-U(i)}{T_k}\right] & \text{if } U(j) > U(i) \\ q_{ij} & \text{if } U(j) \leq U(i)\,,\ j \neq i\,, \end{cases} \tag{2.74}$$

as $k \to \infty$, for all $i,j \in \Sigma$.

Corollary 2.3 If

$$\sigma_k^2 = o(T_k^4) \qquad \text{as } k \to \infty$$

then Theorems 2.1, 2.2, 2.7-2.11 hold with $\{\xi_k\}$ by $\{\hat{\xi}_k\}$.

Remarks

(1) The Corollary is more or less obvious, since the convergence in (2.74) is uniform for $i,j \in \Sigma$ (since Σ is finite); we leave the details to the reader.

(2) We have reason to believe that $\sigma_k^2 = o(T_k^4)$ is quite conservative and that $\sigma_k^2 = o(T_k^2)$ may suffice.

CHAPTER III
DIFFUSION TYPE ALGORITHMS

3.1 Introduction to the Langevin Algorithm

In Chapter 2 we discussed the annealing algorithm proposed by Kirkpatrick et. al. [19] and Cerny [3] for combinatorial optimization. In Chapter 3 we extended the annealing for optimization on general spaces. Motivated by image processing problems with continuous variables, Geman and independently Grenander [13] have recently proposed using diffusions for optimization on multidimensional Euclidean space. In this Section we describe this method. Like the annealing algorithm, this approach to global optimization has generated alot of interest and there already exists a significant literature on the subject.

Let $U(\cdot)$ be a nonnegative continuously differentiable function on $\mathbb{R}^r$. The goal is to find a point in $\mathbb{R}^r$ which minimizes or nearly minimizes $U(\cdot)$. Let $T(\cdot)$ be a positive Borel function on $[0,\infty)$. As with the annealing algorithm we shall refer to $U(\cdot)$ as the *energy function* and $T(\cdot)$ as the *temperature schedule*. Let $w(\cdot)$ be a standard r-dimensional Wiener process and let $x(\cdot)$ be a solution of the stochastic differential equation

$$dx(t) = -\nabla U(x(t))dt + \sqrt{2T(t)}\, dw(t)\,, \qquad t \geq 0\,, \tag{3.1}$$

for some initial condition $x(0) = x_0$ (by a solution we mean that $x(\cdot)$ is a separable process with continuous sample paths with probability one, $x(\cdot)$ is nonanticipative with respect to $w(\cdot)$, and $x(\cdot)$ satisfies the Ito integral equation corresponding to (3.1)). For a fixed temperature $T(t) = T > 0$, (3.1) is the Langevin equation, proposed by Langevin in 1908 to describe the motion of a particle in a viscous fluid. Geman and Grenander suggested that (3.1) could be used to minimize $U(\cdot)$ by letting $T(t) \to 0$. Following Gidas' [11] notation, we shall call the algorithm which simulates the sample paths of $x(\cdot)$ with $T(t) \to 0$ the *Langevin algorithm*.

The motivation behind the Langevin algorithm is similar to that of the annealing algorithm. Let $x^T(\cdot)$ be the solution of (3.1) with $T(t) = T$, a positive constant, and let $P^T(\cdot,\cdot,\cdot)$ be its (stationary) transition function, i.e.,

- for every $t \geq 0$ and $A \in \mathbb{B}^r$ $P^T(t,\cdot,A)$ is a Borel function on $\mathbb{R}^r$
- for every $t \geq 0$ and $x \in \mathbb{R}^r$ $P^T(t,x,\cdot)$ is a probability measure on $(\mathbb{R}^r,\mathbb{B}^r)$
- $P^T(t,x,A) = \int P^T(s,x,dy)\, P^T(t-s,y,A)$ for all $0 \leq s < t$, $x \in \mathbb{R}^r$, and $A \in \mathbb{B}^r$
- $P\{x^T(t) \in A | x^T(s)\} = P^T(x^T(s),A)$ w.p.1 for all $0 \leq s < t$ and $A \in \mathbb{B}^r$

Under certain conditions (c.f. [31]), $P^T(\cdot,\cdot,\cdot)$ has an invariant Gibbs measure $\Pi^T(\cdot)$, i.e.,

$$\Pi^T(A) = \int \Pi^T(dx)\, P^T(t,x,A) \qquad \forall\, t \geq 0\,, \quad \forall\, A \in \mathbb{B}^r\,,$$

where

$$\Pi^T(A) = \frac{\int_A \exp(-\,U(x)/T)\, dx}{\int \exp(-\,U(y)/T)\, dy} \qquad \forall\, A \in \mathbb{B}^r\,,$$

and furthermore

$$P\{x^T(t) \in \cdot\} \rightarrow \Pi^T(\cdot) \quad \text{weakly} \quad \text{as} \quad t \rightarrow \infty\,. \tag{3.2}$$

Now for suitable $U(\cdot)$

$$\Pi^T(\cdot) \rightarrow \Pi^*(\cdot) \quad \text{weakly} \quad \text{as} \quad T \rightarrow 0 \tag{3.3}$$

where $\Pi^*(\cdot)$ is a probability measure on $(\mathbb{R}^r,\mathbb{B}^r)$ with support in the set S of global minima of $U(\cdot)$; see [17] for conditions under which (3.3) holds and a characterization of $\Pi^*(\cdot)$ in terms of the Hessian of $U(\cdot)$. In view of (3.2) and (3.3) the idea behind the Langevin algorithm is that by choosing $T = T(t) \rightarrow 0$ slowly enough hopefully

$$P\{x(t) \in \cdot\} \approx \Pi^{T(t)}(\cdot) \qquad (t \text{ large})$$

and then perhaps

$$P\{x(t) \in \cdot\} \rightarrow \Pi^*(\cdot) \quad \text{weakly} \quad \text{as} \quad t \rightarrow \infty \tag{3.4}$$

and consequently $x(t)$ converges to S in probability.

The Langevin and the annealing algorithms both have a stochastic descent behavior whereby "downhill" moves are modified probabilistically by "uphill" moves with fewer and fewer uphill moves as time tends to infinity and temperature tends to zero. However, the simulations of these Monte Carlo algorithms are quite different. To simulate sample paths of $x(\cdot)$ we might discretize (in time) the Langevin algorithm as

$$x_{k+1}^{\epsilon} = x_k^{\epsilon} - \nabla U(x_k^{\epsilon})\epsilon + \sqrt{2T(k\epsilon)\epsilon}\; w_k\,, \tag{3.5}$$

where $\{w_k\}$ is a sequence of standard $\mathbb{R}^r$-valued Gaussian random variables and ϵ is a (positive) discretization interval, and simulate sample paths of $\{x_k^{\epsilon}\}$ by generating pseudorandom Gaussian variates. $\nabla U(\cdot)$ may be computed from an analytical formula or approximated in a standard fashion. Compare this simulation with that of the annealing algorithm (see Chapter 2).

Geman reports some encouraging numerical results have been obtained by Aluffi-Pentini et. al. [32] with a modified Langevin algorithm which uses an interactive temperature schedule. Tests have been run on $U(\cdot)$ defined on $\mathbb{R}^r$ with $r = 1,\ldots,14$. Gidas also reports a numerical experiment with a single $U(\cdot)$ defined on $\mathbb{R}$ with 400 local minima. He suggests that a combination of the Langevin algorithm with the popular multistart technique (c.f. [29]) might improve the performance obtained by using either approach alone. We remark here that comparing different global optimization algorithms is in general a very difficult problem. Rubenstein [29] discusses some analytical methods for comparing different algorithms. Dixon and Szego [5] have attempted to define a standard set of test functions which might be used to empirically compare different algorithms. It is not clear that either of these methods are suitable for evaluating the performance of the Langevin algorithm. These tools it seems were designed to compare algorithms which in some way take advantage of the structure of smooth functions on low dimensional spaces. We regard the Langevin algorithm as a "universal" algorithm which may be used on functions defined on high dimensional space whose structure is essentially unknown or cannot be simply characterized. It seems that the best test for the Langevin algorithm is the particular problem one wishes to solve.

We shall now outline those convergence results for the Langevin algorithm which are known to us. We refer the reader to the specific paper for full details.

Geman and Hwang [9] were the first to obtain a convergence result for the Langevin algorithm. They consider a version of the Langevin algorithm restricted to a compact subset of $\mathbb{R}^r$ (using reflection barriers). They show that for a temperature schedule of the form

$$T(t) = \frac{c}{\log t} \qquad (t \text{ large})$$

that if c is no smaller than the difference between the maximum and minimum values of $U(\cdot)$ then (3.4) is obtained.

Gidas [11] has obtained necessary and sufficient conditions for the convergence of the Langevin algorithm in all of $\mathbb{R}^r$, using partial differential equation methods. He shows that there exists a constant Δ^* such that for temperature schedules $T(t)\downarrow 0$, (3.4) holds iff

$$\int_0^\infty \exp\left(-\frac{\Delta^*}{T(t)}\right) dt = \infty$$

Furthermore, the constant Δ^* is the natural continuous analog of Hajek's constant (see (2.10)). Chiang et. al. [4] have also obtained sufficient conditions for the convergence of the Langevin algorithm in all of $\mathbb{R}^r$ using large deviations theory.

Kushner [21] has obtained a detailed picture of the asymptotic behavior of a class of diffusions related to the Langevin algorithm and certain discrete-time approximations as well. Kushner considers (in discrete-time) an algorithm of the form

$$X_{k+1} = X_k + a_k b(X_k,\xi_k) + \sqrt{2}\, a_k \sigma(X_k) w_k \tag{3.6}$$

where $\{\xi_k\}$ is a sequence of bounded random variables and

$$a_k = \frac{c}{\log k} \qquad (\text{k large}) .$$

In the special case where $\bar{b}(\cdot) = E\{b(\cdot,\xi_k)\} = -\nabla U(\cdot)$ and $\sigma(\cdot) = I$, (3.6) is a stochastic approximation version of the Langevin algorithm with noisy measurements of $\nabla U(\cdot)$. We shall refer to the Monte Carlo algorithm which simulates the sample paths of $\{X_k\}$ as *Kushner's algorithm.*

We remark that the conditions under which the above results are obtained typically include

(i) $U(\cdot)$ has continuous second-order partial derivatives

(ii) The local minima of $U(\cdot)$ consist of a finite number of compact sets; for Gidas' result it is actually required that the local minima be isolated and nondegenerate.

In this Chapter we shall examine certain issues concerning the Langevin and annealing algorithms which seem important to us and apparently have not been considered elsewhere. We proceed as follows. We have seen that the motivation behind the annealing and Langevin algorithms is quite similar. The first question we would like to answer is:

- what more can be said about the relationship between the annealing and Langevin algorithms?

In 3.2 we shall show that an annealing chain driven by white Gaussian noise converges in a certain sense to a Langevin diffusion. Now it seems clear that the annealing algorithm and the Langevin algorithm each have certain advantages. The Langevin algorithm, for example, looks like (for large time and small temperature) a gradient descent algorithm, and gradient descent algorithms and their higher order generalizations such as Newton's algorithm, which are "local" algorithms in the sense that they use only the value of the objective function and a finite number of derivatives at the current iterate to obtain the next iterate, are efficient at finding local minima. The annealing algorithm, on the other hand, is not strictly "local" in that it uses the value of the objective function in some set containing the current iterate to obtain the next iterate. In this sense, the annealing algorithm might be called "semilocal" or even "global" depending on how much of the objective function is used. Following the usual thinking behind both the annealing and Langevin algorithms, the idea is to make large fluctuations initially and small descent-like moves eventually. In view of these considerations, the second question we would like to answer is:

- is there a natural hybrid algorithm whose initial behavior resembles the annealing algorithm an whose large time behavior is similar to the Langevin algorithm?

3.2 Convergence of the Annealing Chain to a Langevin Diffusion

In this Section we shall examine the relationship between the annealing and Langevin algorithms. We shall show using a result of Kushner's [22] on the weak convergence of interpolated Markov chains to diffusions that a parameterized family of annealing chains driven by white Gaussian noise interpolated into piecewise constant processes converge weakly to a time-scaled solution of the Langevin equation. The weak convergence here is in the sense that the probability measures induced by the interpolated chains on the path space of functions without discontinuities of the second kind converge weakly to the probability measure induced by the limit diffusion. This technique is the same one used to justify the popular diffusion approximation method, whereby a complicated possibly non-Markovian process is approximated by a simpler diffusion process (c.f. [23]).

Let $D^r[0,\bar{T}]$ denote the space of $\mathbb{R}^r$-valued càdlàg functions on $[0,\bar{T}]$ with $0 < \bar{T} < \infty$, i.e., functions which are right-continuous on $[0,\bar{T}]$, have left-hand limits on $(0,\bar{T}]$, and are left continuous at $\bar{T}$. The following elementary results on weak convergence of probability measures may be found in [2]. There is a metric $d_T(\cdot,\cdot)$ on $D^r[0,\bar{T}]$ with respect to which $D^r[0,\bar{T}]$ is a complete separable metric space, and if $f(\cdot)\in D^r[0,T]$ and $\{f_n(\cdot)\}$ is a sequence in $D^r[0,\bar{T}]$ then the convergence of $f_n(\cdot)$ to $f(\cdot)$ in $D^r[0,\bar{T}]$ implies convergence at all points of continuity of $f(\cdot)$ (convergence of $f_n(\cdot)$ to $f(\cdot)$ in $D^r[0,\bar{T}]$ is roughly equivalent to uniform convergence outside of any neighborhood of the discontinuity points of $f(\cdot)$). Let $\xi(\cdot)$, $\{\xi_\epsilon(\cdot) : \epsilon > 0\}$ be processes with sample paths in $D^r[0,\bar{T}]$, or equivalently, random variables which take values in $D^r[0,\bar{T}]$, and let $\mu(\cdot)$, $\{\mu_\epsilon(\cdot) : \epsilon > 0\}$ be the probability measures they induce on the Borel subsets of $D^r[0,\bar{T}]$. We shall say that $\xi_\epsilon(\cdot)$ converges weakly to $\xi(\cdot)$ in $D^r[0,\bar{T}]$ and write $\xi_\epsilon(\cdot)\to\xi(\cdot)$ weakly (in $D^r[0,\bar{T}]$) if $\mu_\epsilon(\cdot)$ converges weakly to $\mu(\cdot)$ as $\epsilon\to 0$, i.e., if

$$\lim_{\epsilon\to 0} \int f(x)\, d\mu_\epsilon(x) = \int f(x)\, d\mu(x)$$

for all bounded continuous $f(\cdot)$ on $D^r[0,\bar{T}]$. Let $D^r[0,\infty)$ denote the set of $\mathbb{R}^r$-valued functions on $[0,\infty)$ which are right-continuous on $[0,\infty)$ and have left-hand limits on $(0,\infty)$. Let

$$d(f,g) = \sum_{n=1}^{\infty} \frac{1}{2^n}\, d_n(f,g) \qquad \forall\, f,g\in D^r[0,\infty)\ .$$

$d(\cdot,\cdot)$ is a metric on $D^r[0,\infty)$ with respect to which $D^r[0,\infty)$ is a complete separable metric space, and we can define the weak convergence of processes with sample paths in $D^r[0,\infty)$ analogously to $D^r[0,\bar{T}]$ with $\bar{T}$ finite.

Suppose $\xi_\epsilon(\cdot)\to\xi(\cdot)$ weakly (in $D^r[0,\bar{T}]$) as $\epsilon\to 0$ with $0 < \bar{T} \le \infty$. Then it can be shown that the set of points $t\in[0,\bar{T}]$ such that $\mu(\{\xi(t_-) \ne \xi(t)\}) > 0$ is at most countable. Let

$$C = \{t\in[0,\bar{T}] : \mu(\{\xi(t_-) \ne \xi(t)\}) = 0\}\ .$$

Then it can also be shown that for any points $t_1,\dots,t_k\in C$ the multivariate distributions of $\{\xi_\epsilon(t_1),\dots,\xi_\epsilon(t_k)\}$ converge to the multivariate distributions of $\{\xi(t_1),\dots,\xi(t_k)\}$ as $\epsilon\to 0$. But the weak convergence of $\xi_\epsilon(\cdot)$ to $\xi(\cdot)$ says much more than this: if $f(\cdot)$ is a continuous functional on $D^r[0,\bar{T}]$ (or just μ-a.s. continuous) then $f(\xi_\epsilon(\cdot))\to f(\xi(\cdot))$ weakly as $\epsilon\to 0$.

Let $C^r[0,\bar{T}]$ denote the space of $\mathbb{R}^r$-valued continuous functions on $[0,\bar{T}]$ with $0 \le \bar{T} \le \infty$. If we equip $C^r[0,\bar{T}]$ with the uniform topology for $\bar{T} < \infty$ and with the topology of uniform convergence on compacts for $\bar{T} = \infty$, then $C^r[0,\bar{T}]$ is a complete separable metric space and we can define weak convergence of processes with sample paths in $C^r[0,\bar{T}]$. Our reason for using $D^r[0,\bar{T}]$ is simply that we shall make use of Kushner's result on the weak convergence of Markov chains interpolated into $D^r[0,\bar{T}]$. Kushner's stated reason for working with $D^r[0,\bar{T}]$ as opposed to $C^r[0,\bar{T}]$ is that it is easier to verify tightness (relative compactness) for a sequence of probability measures on the Borel subsets of $D^r[0,\bar{T}]$. If the limit process is a jump diffusion then of course it would be necessary to work with $D^r[0,\bar{T}]$, but this is not an issue here since our limit processes are assumed to be ordinary (continuous sample paths with probability one) diffusions.

We now set up the notation necessary to state Kushner's Theorem on the weak convergence of interpolated Markov chains. It will be notationally convenient in the sequel to assume that all processes are defined on a common probability space (Ω, F, P) and we shall do so without further comment. Let $0 < \bar{T} < \infty$. Let $F(\cdot,\cdot)$ and $F_\epsilon(\cdot,\cdot)$, $\epsilon > 0$, be $\mathbb{R}^r$-valued Borel functions on $\mathbb{R}^r\times[0,\bar{T}]$, and let $G(\cdot,\cdot)$ and $G_\epsilon(\cdot,\cdot)$, $\epsilon > 0$, be r×r matrix-valued Borel functions on $\mathbb{R}^r\times[0,\bar{T}]$. For each $\epsilon > 0$ let $\{\xi^\epsilon_k\}$ be a Markov chain with state-space $\mathbb{R}^r$ such that

$$E\{\xi^\epsilon_{k+1} - \xi^\epsilon_k | \xi^\epsilon_k\} = F_\epsilon(\xi^\epsilon_k,k\epsilon)\epsilon\ ,$$

$$E\{(\xi^\epsilon_{k+1} - \xi^\epsilon_k)\otimes(\xi^\epsilon_{k+1} - \xi^\epsilon_k) | \xi^\epsilon_k\} = G_\epsilon(\xi^\epsilon_k,k\epsilon)\, G'_\epsilon(\xi^\epsilon_k,k\epsilon)\epsilon\ ,$$

with probability one. Interpolate $\{\xi^\epsilon_k\}$ into a process $\xi_\epsilon(\cdot)$ with sample paths in $D^r[0,\bar{T}]$ by

$$\xi_\epsilon(t) = \xi^\epsilon_k \qquad \forall\ (k-1)\epsilon \le t < k\epsilon\ , \qquad \forall\ k = 1,\dots,\left\lceil \frac{\bar{T}}{\epsilon} \right\rceil .$$

Here is Kushner's Theorem in slightly modified form.

Theorem 3.1 (Kushner [22]). Assume

(K1) $F(\cdot,\cdot)$, $G(\cdot,\cdot)$ are bounded and continuous

(K2) $F_\epsilon(\cdot,\cdot)$, $G_\epsilon(\cdot,\cdot)$ are uniformly bounded for small $\epsilon > 0$

$$\text{(K3)}\quad E\left\{\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor}\left[|F_\epsilon(\xi_k^\epsilon,k\epsilon)-F(\xi_k^\epsilon,k\epsilon|^2+|G_\epsilon(\xi_k^\epsilon,k\epsilon)-G(\xi_k^\epsilon,k\epsilon)|^2\right]\epsilon\right\}\longrightarrow 0$$

as $\epsilon \longrightarrow 0$

$$\text{(K4)}\quad E\left\{\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor}\left[|\xi_{k+1}^\epsilon-\xi_k^\epsilon-F_\epsilon(\xi_k^\epsilon,k\epsilon)\epsilon|^{2+\alpha}\right\}\longrightarrow 0\right.$$

as $\epsilon \longrightarrow 0$ for some $\alpha > 0$.

Let $v(\cdot)$ be a standard r-dimensional Wiener process and assume that

$$d\xi(t) = F(\xi(t),t)dt + G(\xi(t),t)dv(t), \qquad 0 \le t \le \bar{T},$$

has a unique solution $\xi(\cdot)$ (in the sense of multivariate distributions) with initial condition $\xi(0) = \xi_0$. Assume that

$$\xi_1^\epsilon \longrightarrow \xi_0 \quad \text{weakly as } \epsilon \longrightarrow 0 .$$

Then

$$\xi_\epsilon(\cdot) \longrightarrow \xi(\cdot) \quad \text{weakly (in } D^r[0,\bar{T}]) \text{ as } \epsilon \longrightarrow 0 .$$

Consider now the following family of Markov chains. Let $U(\cdot)$ and $T(\cdot)$ be defined as in .1. For each $\epsilon > 0$ let $\{z_k^\epsilon\}$ be a Markov chain with state space $\mathbb{R}^r$ and 1-step transition functions $\{P_k^\epsilon(\cdot,\cdot)\}$ given by†

$$P_k^\epsilon(x,A) = \int_A s_k^\epsilon(x,y)\, dN(x,\epsilon I)(y) + \gamma_k^\epsilon(x)\, \delta(x,A) \tag{3.7}$$

for all $x \in \mathbb{R}^r$ and $A \in \mathbb{B}^r$, where

$$s_k^\epsilon(x,y) = \begin{cases} \exp\left[-\dfrac{U(y)-U(x)}{T(k\epsilon)}\right] & \text{if } U(y) > U(x) \\ 1 & \text{if } U(y) \le U(x) , \end{cases} \tag{3.8}$$

$$\gamma_k^\epsilon(x) = 1 - \int s_k^\epsilon(x,y)\, dN(x,\epsilon I)\,(y) , \tag{3.9}$$

and $\delta(x,\cdot)$ is the unit measure concentrated at x, for all $x,y \in \mathbb{R}^r$. It is easy to see that $\{z_k^\epsilon\}$ is infact an annealing chain with state space the measure space (Σ,B,ϕ) where $\Sigma = \mathbb{R}^r$, $B = \mathbb{B}^r$, $\phi(\cdot)$ is Lebesque measure, and

$$Q(x,A) = \int_A q(x,y)\, \phi(dy) = N(x,\epsilon I)\,(A) \qquad \forall\, A \in \mathbb{B}^r$$

(hence the annealing chain is "driven" by white Gaussian noise). It will be convenient to introduce the following notation. For each $\epsilon > 0$ let

$$s(x,y,t) = \begin{cases} \exp\left[-\dfrac{U(y)-U(x)}{T(t)}\right] & \text{if } U(y) > U(x) \\ 1 & \text{if } U(y) \le U(x) , \end{cases}$$

$$\gamma_\epsilon(x,t) = 1 - \int s(x,y,t)\, dN(x,\epsilon I)\,(y) ,$$

for all $x,y \in \mathbb{R}^r$ and $t \ge 0$, and let

$$P_\epsilon(x,A,t) = \int_A s(x,y,t)\, dN(x,\epsilon I)\,(y) + \gamma_\epsilon(x,t)\, \delta(x,A)$$

for all $x \in \mathbb{R}^r$, $A \in \mathbb{B}^r$, and $t \ge 0$. Then

$$P_\epsilon(x,A,k\epsilon) = P_k^\epsilon(x,A) \qquad \forall\, x \in \mathbb{R}^r ,\ \forall\, A \in \mathbb{B}^r .$$

For each $\epsilon > 0$, $x \in \mathbb{R}^r$ and $t \ge 0$ let

$$b_\epsilon(x,t) = \frac{1}{\epsilon} \int (y-x)\, P_\epsilon(x,dy,t) ,$$

$$a_\epsilon(x,t) = \frac{1}{\epsilon} \int (y-x) \otimes (y-x)\, P_\epsilon(x,dy,t) ,$$

and $\sigma_\epsilon(x,t)$ be a positive square root of $a_\epsilon(x,t)$ i.e.

$$\sigma_\epsilon(x,t)\sigma_\epsilon'(x,t) = a_\epsilon(x,t) .$$

Since $P_\epsilon(\cdot,\cdot,k\epsilon) = P_k^\epsilon(\cdot,\cdot)$ is a (regular wide-sense) conditional distribution for

z^ϵ_{k+1} given z^ϵ_k,

$$E\{z^\epsilon_{k+1} - z^\epsilon_k | z^\epsilon_k\} = b_\epsilon(z^\epsilon_k,k\epsilon)\epsilon$$
$$E\{(z^\epsilon_{k+1} - z^\epsilon_k) \otimes (z^\epsilon_{k+1} - z^\epsilon_k) | z^\epsilon_k\} = \sigma_\epsilon(z^\epsilon_k,k\epsilon)\sigma'_\epsilon(z^\epsilon_k,k\epsilon)\epsilon$$

with probability one. Interpolate $\{z^\epsilon_k\}$ into $z_\epsilon(\cdot)$ with sample paths in $D^r[0,\overline{T}]$ by

$$z_\epsilon(t) = z^\epsilon_k \qquad \forall\,(k-1)\epsilon \le t < k\epsilon\,, \qquad \forall\, k = 1,\ldots,\left\lceil \frac{\overline{T}}{\epsilon} \right\rceil .$$

Here is our convergence theorem.

Theorem 3.2 Assume

(A1) $U(\cdot)$ is continuously differentiable, $\nabla U(\cdot)$ is bounded and Lipshitz

(A2) $T(\cdot)$ is continuous

Let $w(\cdot)$ be a standard r-dimensional Wiener process, and let $z(\cdot)$ be a solution of†

$$dz(t) = -\frac{\nabla U(z(t))}{2T(t)}\,dt + dw(t)\,, \qquad 0 \le t \le \overline{T}\,, \tag{3.10}$$

with initial condition $z(0) = z_0$. Assume that

$$z^\epsilon_1 \longrightarrow z_0 \quad \text{weakly} \quad \text{as } \epsilon \longrightarrow 0\,.$$

Then

$$z_\epsilon(\cdot) \longrightarrow z(\cdot) \quad \text{weakly} \quad (\text{in } D^r[0,\overline{T}]) \text{ as } \epsilon \longrightarrow 0\,.$$

REFERENCES

[1] Ash, R.: *Real Analysis and Probability,* Academic Press (1972).

[2] Billingsley, P.: *Convergence of Probability Measures,* Wiley (1968).

[3] Cerny, V.: "A Thermodynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm," Preprint, Inst. of Phys. and Biophys., Comenius Univ., Bratislava (1982).

[4] Chiang, T.S., C.R. Hwang and S.J. Shew: "Diffusion for Global Optimization in $\mathbb{R}^n$," Preprint, Institute of Mathematics, Academia Sinica, Tapei, Taiwan (1985).

[5] Dixon, L.C.W. and G.P. Szegö: *Towards Global Optimization,* North Holland (1978).

[6] Doob, J.: *Stochastic Processes,* Wiley (1953).

[7] Feller, W.: *An Introduction to Probability Theory and Its Applications,* (Vol. 1), Wiley (1957).

[8] Geman, S. and D. Geman: "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Trans. PAMI 6 (1984), 721-741.

[9] Geman, S. and C.R. Hwang: "Diffusions for Global Optimization," SIAM J. Cntrl. Opt. 24 (1986) 1031-1043.

[10] Gidas, B.: "Nonstationary Markov Chains and Convergence of the Annealing Algorithm," J. Stat. Phys. 39 (1985) 73-131.

[11] Gidas, B.: "Global Optimization via the Langevin Equation," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[12] Golden B. and C. Skiscim: "Using Simulated Annealing to Solve Routing and Location Problems," Naval Res. Log. Quarterly 33 (1986) 261-279.

[13] Grenander, U.: *Tutorial in Pattern Theory,* Brown University (1983).

[14] Hajek, B.: "Cooling Schedules for Optimal Annealing," Preprint, Dept. of Elec. Eng. and Coord. Science Lab., U. Illinois at Champaign-Urbana (1985).

[15] Hajek, B.: "Tutorial Survey of Theory and Applications of Simulated Annealing," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[16] Hammersley, J. and D. Handscomb: *Monte Carlo Methods,* Chapman and Hall (1964).

[17] Hwang, C.R.: "Laplace's Method Revisited: Weak Convergence of Probability Measures," Ann. Prob. 8 (1980) 1177-1182.

[18] Johnson, D.S., C.R. Aragon, L.A. McGeoch, and C. Schevon: "Optimization by Simulated Annealing: An Experimental Evaluation," Preprint (1985).

[19] Kirkpatrick, S., C.D. Gelatt, and M. Vecchi: "Optimization by Simulated Annealing," Science 220 (1983) 621-680.

[20] Knopp, K.: *Theory and Application of Infinite Series,* Hafner (1971).

[21] Kushner, H.: "Asymptotic Behavior for Stochastic Approximations and Diffusion with Slowly Decreasing Noise Effects: Global Minimization via Monte Carlo," Preprint, Div. Applied Math., Brown University (1985).

[22] Kushner, H.: "On the Weak Convergence of Interpolated Markov Chains to a Diffusion," Ann. Prob. 2 (1974) 40-50.

[23] Kushner, H.: *Approximation and Weak Convergence Methods for Random Processes,* MIT Press (1984).

[24] Kushner, H. and D. Clark: *Stochastic Approximation for Constrained and Unconstrained Systems,* Springer Verlag (1978).

[25] Metropolis, M., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller: "Equations of State Calculations by Fast Computing Machines," J. Chem. Phys. 21 (1953) 1087-1091.

[26] Mitra, D., F. Romeo, and A. Sangiovanni-Vancentelli: "Convergence and Finite-time Behavior of Simulated Annealing," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[27] Orey, S.: *Limit Theorem for Markov Chain Transition Probabilities,"* Van Nostrand (1971).

[28] Royden, H.: *Real Analysis,* Macmillan (1964).

[29] Rubenstein, R.: *Simulation and the Monte-Carlo Method,* Wiley (1981).

[30] Tsitsiklis, J.: "Markov Chains with Rare Transitions and Simulated Annealing," Preprint, Lab. for Info. and Decision Systems (1985).

[31] Varadhan, S.R.S.: *Lectures on Diffusion Problems and Partial Differential Equations,* Tata Institute, Bombay (1980).

[32] Aluffi-Pentini, F., V. Parisi, and F. Zerilli: "Global Optimization and Stochastic Differential Equations," J. Opt. Th. Applic. (1985) (to appear).