Annales Academiæ Scientiarum Fennicæ
Series A. I. Mathematica
Volumen 17, 1992, 39–50

WEAK CONVERGENCE OF THE SIMULATED ANNEALING PROCESS IN GENERAL STATE SPACE

Heikki Haario and Eero Saksman

University of Helsinki, Department of Mathematics,
Hallituskatu 15, SF-00100 Helsinki, Finland

0. Introduction

Rigorous proofs of the convergence of the simulated annealing process in the *original* formulation have only been given in the case of a finite state space. In Geman and Hwang [5] the idea of a stochastic search of a minimum was cast in the form of solving a stochastic differential equation describing diffusion, together with a proof of the weak convergence of the process. Kushner [11] and others study generalizations of the process considered in [5].

The aim of this paper is to generalize the definition of the stochastic process corresponding to the original simulated annealing algorithm to the case of an arbitrary state space and study the weak convergence of the process. The results are expressed in terms of the generating distributions and the sequence of successive temperature parameters. At the same time we obtain estimates for the (Dobrushin) coefficient of ergodicity for an nth iterate of some important kinds of generating distributions. These estimates are also of independent interest.

For most of the results the proofs are only sketched here. For full proofs see [6].

1. Generalization of the discrete simulated annealing

We recall the discrete simulated annealing algorithm. The simulated annealing process is a time-inhomogeneous Markov chain on the configuration space $\mathscr{S} = \{1, 2, \dots, N\}$, generated by $N \times N$-matrices Q and q. Here Q is a given (transition) probability matrix, the *generating distribution.* The name derives from the fact that Q_{ij} gives the probability of testing the state j after the current state i. The *acceptance probability* q controls the acceptance of new states for the process. In the original simulated annealing algorithm q is specified according to the Boltzmann statistics

$$q_{ij}^{n} = \min\left\{1, e^{-(f(j)-f(i))/T_n}\right\} \tag{1.1}$$

The work of the second author was partially supported by the Academy of Finland.

doi:10.5186/aasfm.1992.1706

where T_n is the *temperature parameter* at the nth step of the process.

The nth step transition probabilities P_{ij}^n are given by

$$P_{ij}^n = \begin{cases} Q_{ij}q_{ij}^n, & \text{if } i \neq j, \\ 1 - \sum_{k \neq i} P_{ik}^n, & \text{if } i = j. \end{cases} \tag{1.2}$$

In other words, the process moves from i to j $(j \neq i)$ after testing and accepting j, or otherways stays at i.

For the purposes of generalization we rewrite (1.2) in an equivalent form:

$$P_{ij}^n = Q_{ij}q_{ij}^n + \delta_{ij} \sum_k Q_{ik}(1 - q_{ik}^n). \tag{1.3}$$

The very idea behind the simulated annealing algorithm shows clearly how to generalize (1.3) to the case of a non-discrete state space $\mathscr{S}$ and a measurable bounded f: replace Q with a transition probability function, q_{ik} with a function given by an expression similar to (1.1) and the Kronecker delta with a characteristic function so that one writes

$$P^n(x, A) = \int_A Q(x, d\omega)q^n(x, \omega)\, d\omega + \chi_A(x) \int_{\mathscr{S}} Q(x, d\omega)\bigl(1 - q^n(x, \omega)\bigr)\, d\omega \tag{1.4}$$

where $x \in \mathscr{S}$ and $A \subset \mathscr{S}$.

In the sequel we study the convergence of the Markov process associated with (1.4) allowing $\mathscr{S}$ to be completely arbitrary. More specifically, let $\mathscr{S} = (\mathscr{S}, \mathscr{A}, m)$ be a *state space*, i.e. m is a probability measure on a σ-algebra $\mathscr{A} \subset P(\mathscr{S})$ of the set $\mathscr{S}$. By $\Lambda = \Lambda(\mathscr{S})$ we denote the set of all probability measures on the space $(\mathscr{S}, \mathscr{A})$. The norm $\|\mu\|$ of a measure μ is always the total variation norm. We denote by $Op(\mathscr{S})$ the set of all transition probability functions (t.p.f's) on the space $(\mathscr{S}, \Lambda)$. The function to be minimized (the energy function) can be any bounded measurable mapping $f\colon \mathscr{S} \to \mathbf{R}$. For the sake of convenience we assume that the absolute minimum of f will always be scaled to zero, this does not affect the process.

A function $f\colon \mathscr{S} \to R$ is called an *energy function* if it is $\mathscr{A}$-measurable, positive, bounded and if, in addition, $\text{ess.inf}_{\omega \in \mathscr{S}} f = 0$. We denote by $M_f = \{\omega \in \mathscr{S} \mid f(\omega) = 0\}$ the *minimum set* of f and by Δf the maximum variation of f: $\Delta f = \sup_{\omega \in \mathscr{S}} f(\omega)$. Given the energy function f, the formula

$$q_T(x, y) = e^{\frac{1}{T}\min\{0, f(x) - f(y)\}} \tag{1.5}$$

defines the *acceptance probability function* $q_T\colon \mathscr{S} \times \mathscr{S} \to R$. Here $T > 0$ is the *temperature parameter.*

Clearly the set M_f and the function q_T are measurable, $0 < q_T \leq 1$. Note that the set M_f may well be empty.

The equality of the conditions given in the following definition is apparent.

Definition 1.1. We call a t.p.f $Q \in Op(\mathscr{S})$ *symmetric*, if the measures $Q(x, d\omega)m(dx)$ and $Q(\omega, dx)m(d\omega)$ on the product space $S \times \mathscr{S}$ coincide. In other words: for every $A_1, A_2 \in \mathscr{A}$ one has

$$\int_{A_1} m(dx)Q(x, A_2) = \int_{A_2} m(dx)Q(x, A_1).$$

Examples of symmetric t.p.f's are e.g. operators of the form $Q(x, d\omega) = g(x, \omega)\cdot$ $\cdot m(d\omega)$, where g is symmetric. A symmetric t.p.f can be seen as a generalization of a symmetric matrix in a discrete state space. A symmetric t.p.f $Q \in Op(\mathscr{S})$ is called a *generating distribution.*

Suppose that $(P_i)_{i=1}^{\infty}$ is a sequence of t.p.f's on the space $(\mathscr{S}, \mathscr{A})$ and $\mu_0 \in \Lambda(\mathscr{S})$ is a given *initial distribution.* For our purposes the sequence $(P_i)_{i=1}^{\infty}$ and the distribution μ_0 together define a discrete time *inhomogeneous Markov process* with the state space $\mathscr{S}$, since we are interested only in the *distributions* μ_i of random variables X_i. For the successive distributions one writes $\mu_i = \mu_{i-1}P_i$. We will use the notation

$$P^{(m,k)} = P_{m+1}P_{m+2}\cdots P_k$$

so that one also has $\mu_k = \mu_m P^{(m,k)}$. Often we simply identify a Markov process and the corresponding sequence $(P_i)_{i\in\mathbf{N}}$ of t.p.f's with the understanding that the process depends on the initial distribution.

We are now able to define the general simulated annealing process.

Definition 1.2. Let f be an energy function defined in state space $(\mathscr{S}, \mathscr{A}, m)$, $(Q_i)_{i\in\mathbf{N}}$ a sequence of generating distributions, $(T_i)_{i\in\mathbf{N}}$ a sequence of temperature parameters, for which

$$T_1 \geq T_2 \geq \cdots \geq T_{i-1} \geq T_i \xrightarrow[i\to\infty]{} 0.$$

A *simulated annealing process* is the nonhomogeneous Markov process $(P_i)_{i\in\mathbf{N}}$ in the state space $(\mathscr{S}, \mathscr{A}, m)$, defined by

$$P_i(x, A) = \int_A Q_i(x, d\omega)q_{T_i}(x, \omega)\, d\omega + \chi_A(x)\int_{\mathscr{S}} Q_i(x, d\omega)\big(1 - q_{T_i}(x, \omega)\big)d\omega$$

where $x \in \mathscr{S}$, $A \in \mathscr{A}$, and χ_A is the characteristic function of A.

Routine verifications show that t.p.f's P_i are well defined. The symmetry of Q implies that the Bolzman distribution is an equilibrium distribution for P_i. By considering separately the cases $f(x) \leq f(\omega)$, $f(x) \geq f(\omega)$ it is easy to see that the 'detailed balance equation' $e^{-f(x)/T}q(x, \omega) = e^{-f(\omega)/T}q(\omega, x)$ holds. Using this and the symmetry of Q it is easy to see that the operator P_i has the equilibrium distribution π_i,

$$\pi_i(d\omega) = C_i e^{-f(\omega)/T_i} m(d\omega), \tag{1.6}$$

where C_i is the normalization constant, $C_i = \big(\int_{\mathscr{S}} e^{-f(\omega)/T_i} m(d\omega)\big)^{-1} > 0$.

2. The behaviour of the sequence $(\pi_n)_{n=1}^{\infty}$

In the case of a continuous state space it is no more possible to prove convergence results along the lines of [4] and [1] since their method relies on the convergence of the sequence $\sum_{i=1}^{\infty} \|\pi_i - \pi_{i+1}\|$. For that reason this section contains a more careful study on the properties of the equilibrium distributions.

In Theorem 2.2 we estimate the distance between two equilibrium distributions which correspond to different values of temperature. The estimate plays a fundamental role in later considerations. In order to be able to state the theorem we first define a quantity which describes the behaviour of an energy function f in state space $(\mathscr{S}, \mathscr{A}, m)$ near its absolute minima.

We recall the definition of the distribution function λ_f of f: for $x \in \mathbf{R}$ one writes

$$\lambda_f(x) = m\{\omega \mid f(\omega) \leq x\}. \tag{2.1}$$

Note that λ_f is increasing, $\lambda_f(x) = 0$ for $x < 0$, $\lambda_f(x) > 0$ for $x > 0$, and $\lambda_f(x) = 1$ for $x \geq \Delta f$. Thus λ_f is also the distribution function of a probability measure $\lambda_f(dx)$ concentrated on the interval $[0, \Delta f]$.

Definition 2.1. We denote by $\mathscr{L}_f$ the Laplace transform of the measure $\lambda_f(dx)$ and call it the *steepness indicator* of the energy function f. In other words, for every $z \in \mathbf{C}$

$$\mathscr{L}_f(z) = \int_{\mathbf{R}} e^{-zx} \lambda_f(dx). \tag{2.2}$$

Theorem 2.2. *Let f be an energy function. Let π_i, $0 < k \leq i \leq n$, be given as in* (1.6) *and let $T_k \geq T_{k+1} \geq \cdots \geq T_n > 0$. Then the following estimate is valid:*

$$\sum_{i=k+1}^{n} \|\pi_i - \pi_{i-1}\| \leq 2 \log\left(\frac{\mathscr{L}_f(1/T_k)}{\mathscr{L}_f(1/T_n)}\right). \tag{2.3}$$

Proof. Because of the logarithmic form of the estimate (2.3) it is clearly enough to consider the case $k = n - 1$. First we note that it follows from the definitions that $\int_{\mathscr{S}} e^{-f(\omega)/T} m(d\omega) = \mathscr{L}_f(1/T)$. We may now write

(2.4)
$$\|\pi_n - \pi_{n-1}\| = \int_{\mathscr{S}} \left| e^{-f(x)/T_n} \left(\mathscr{L}_f(1/T_n)\right)^{-1} - e^{f(x)/T_{n-1}} \left(\mathscr{L}_f(1/T_{n-1})\right)^{-1} \right| m(dx),$$

and estimate the integrand as follows:

$$\begin{aligned}
&\left| e^{-f(x)/T_n} \left(\mathscr{L}_f(1/T_n)\right)^{-1} - e^{f(x)/T_{n-1}} \left(\mathscr{L}_f(1/T_{n-1})\right)^{-1} \right| \\
&\quad = \left| \int_{1/T_{n-1}}^{1/T_n} \frac{d}{du} \left(e^{-f(x)u} \left(\mathscr{L}_f(u)\right)^{-1} \right) du \right| \\
&\quad \leq \int_{1/T_{n-1}}^{1/T_n} \left(\mathscr{L}_f(u)\right)^{-2} e^{-f(x)u} \left(\left|f(x)\mathscr{L}_f(u)\right| + \left|\mathscr{L}_f{}'(u)\right| \right) du.
\end{aligned}$$

Substituting the estimate in (2.4), using $\int_{\mathscr{S}} e^{-f(\omega)/T} f(\omega) m(d\omega) = -\mathscr{L}_f{}'(1/T)$ and the fact that $f \geq 0$ we get (2.3).

To state conditions for weak ergodicity, we recall (Dobrushin [3]) the concept of *coefficient of ergodicity.* Let $P \in Op(\mathscr{S})$. The coefficient of ergodicity of P, denoted by $\delta(P)$, is defined as

$$\delta(P) = \sup_{\lambda, \mu \in \Lambda,\, \lambda \neq \mu} \frac{\|\mu P - \lambda P\|}{\|\lambda - \mu\|}.$$

Clearly $0 \leq \delta(P) \leq 1$. In the case $\delta(P) < 1$ the mapping P is a contraction on the space Λ in the metric defined by $\|\cdot\|$. From the definition it easily follows that (cf. [9])

$$\delta(P_1 P_2 \cdots P_n) \leq \prod_{i=1}^{n} \delta(P_i).$$

The following result (c.f. [10, Theorem 1]) follows rather easily from the definitions: The nonhomogeneous Markov process $(P_i)_{i \in N}$ is weakly ergodic if there exists an increasing sequence (n_i) of positive integers for which $\sum_{i=1}^{\infty} \left[1 - \delta(P^{(n_i, n_{i+1})})\right] = \infty$.

Simple estimates prove a useful lemma:

Lemma 2.3. *Consider a simulated annealing process as in Definition* 1.2. *Write* $P = P_k P_{k+1} \cdots P_n$ *and* $Q = Q_k Q_{k+1} \cdots Q_n$, *where* $1 \leq k < n$. *Then*

$$1 - \delta(P) \geq \exp\Big(- \sum_{i=k}^{n} \Delta f / T_i \Big) (1 - \delta(Q)).$$

In the case $m(M_f) = 0$ it is easy to prove that in general one cannot ensure that the annealing process converges in norm. For completeness, we include the following theorem, which can be proved using Theorem 2.2, Lemma 2.3 and [9, Theorem 2.1]:

Theorem 2.4. *Let* $(P_i)_{i=1}^{\infty}$ *be a simulated annealing process. Suppose that* $m(M_f) > 0$ *and the following conditions hold: there exists an increasing sequence of indices* $(n_i)_{i \in N}$ *and a number* $0 < d < 1$ *such that for every* $i \geq 1$ *the following conditions are satisfied:*

(i) $$\delta(Q^{(n_i, n_{i+1})}) = \delta(Q_{n_i+1} Q_{n_i+2} \cdots Q_{n_{i+1}}) \leq d,$$

(ii) $$\sum_{i=1}^{\infty} e^{-\sum_{j=n_i+1}^{n_{i+1}} \Delta f \frac{1}{T_j}} = \infty.$$

Then the simulated annealing process $(P_i)_1^{\infty}$ *converges in the norm and has the limit distribution* π, *with* $\pi(A) = m(A \cap M_f)/m(M_f)$.

As a corollary, we obtain a basic result, familiar for discrete simulated annealing process: Suppose that $m(M_f) > 0$. If $\delta(Q_{ik+1}Q_{ik+2}\cdots Q_{(i+1)k}) \leq \delta < 1$ for all $i \in N$, then the simulated annealing process is strongly ergodic with the limit distribution π, provided $T_i \geq k\Delta f/\log(i+2)$. We remark that the conditions for convergence in the discrete case have been carefully analyzed, cf. e.g. [2] and [7].

3. The weak convergence of the simulated annealing process

The starting point of this section is Theorem 3.1, which states that even in the case $m(M_f) = 0$ we have $\lim_{i\to\infty} \|\mu_i - \pi_i\| = 0$ under suitable conditions. The weak convergence of the process is thus reduced to the weak convergence of the equilibrium distributions.

Theorem 3.1. *Let $(P_i)_{i=1}^{\infty}$ be a simulated annealing process. Suppose that the following conditions hold: there exist sequences $(n_i)_{i\in N}$ and $(r_i)_{i\in N}$ of indices and a number $0 < d < 1$ such that $(n_i)_{i\in N}$ is increasing and $\lim_{i\to\infty} i - r_i = \infty = \lim_{i\to\infty} r_i$. Suppose also that the following three conditions are satisfied:*

(i) $$\delta(Q^{(n_i,n_{i+1})}) = \delta(Q_{n_i+1}Q_{n_i+2}\cdots Q_{n_{i+1}}) \leq d \text{ for each } i \geq 1,$$

(ii) $$\lim_{k\to\infty} \sum_{i=r_k}^{k} e^{-\sum_{j=n_i+1}^{n_i+1} \Delta f \frac{1}{T_j}} = \infty,$$

(iii) $$\lim_{k\to\infty} \frac{\mathscr{L}_f(1/T_{n_{r_k}})}{\mathscr{L}_f(1/T_{n_{k+2}})} = 1.$$

Then $\lim_{i\to\infty} \|\mu_i - \pi_i\| = 0$.

Proof. (i) and (ii) are the conditions for weak ergodicity. In contrast to the earlier cases, we no more have $\sum_{i=1}^{\infty} \|\pi_i - \pi_{i+1}\| < \infty$. However, it is still possible to verify that $\|\mu_i - \pi_i\|$ is small by writing it as a sum of appropriate terms including sums of the form $\sum_{r(k)}^{k} \|\pi_i - \pi_{i+1}\|$, which can be controlled by suitable $r(k)$ using Theorem 2.2.

Theorem 3.1 is perhaps not too transparent. To express it in terms of more concrete conditions, we may state the result e.g. in the following form: Suppose that $\delta(Q_{is+1}Q_{is+2}\cdots Q_{(i+1)s}) \leq d < 1$ for all $i \geq 1$ and some $s \geq 1$. Then $\lim_{n\to\infty} \|\mu_n - \pi_n\| = 0$, provided that there exists $\varepsilon \in]0,1[$ such that for $i \geq 1$ one of the following two conditions is satisfied

(iii′) $$T_i = \frac{(1+\varepsilon)s\Delta f}{\log(i+2)};$$

(iii$''$) $$T_i \geq \frac{(1+\varepsilon)s\Delta f}{\log(i+2)} \text{ and the mapping } i \mapsto T_i \text{ is convex.}$$

Again, we meet the logarithmic rule, familiar from the discrete case. Another consequence is the following, expressing a basic demand for a stochastic minimization algorithm:

Corollary 3.2. *Suppose that the conditions of Theorem* 3.1 *are fulfilled and* $\varepsilon > 0$. *Then*
$$\lim_{n\to\infty} \mathbf{P}\{f(X_n) \leq \varepsilon\} = 1.$$

Proof. Using the explicit formula (1.6) we easily see that $\lim_{n\to\infty} \pi_n\{\omega \mid f(\omega) \geq \varepsilon\} = 0$. This observation together with Theorem 3.1 imply the claim.

We next give examples on weak convergence of the Metropolis process. For that end we suppose for the last part of this section that the state space $\mathscr{S}$ has a *Hausdorff topology* θ and the σ-algebra $\mathscr{A}$ is generated by the open sets $U \in \theta$. We denote the weak convergence (which is defined exactly as in the case of a metric space) by $\lambda_n \xrightarrow[w]{} \lambda$. Suppose that the conditions of Theorem 3.1 are fulfilled. Suppose in addition that $\pi_n \xrightarrow[w]{} \pi$, where $\pi \in \Lambda(\mathscr{S})$. Then obviously $\mu_n \xrightarrow[w]{} \pi$.

Example 1. Suppose that the assumptions of Theorem 3.1 are satisfied for the simulated annealing process $(P_i)_{i=1}^{\infty}$. Suppose also that f achieves its absolute minimum at $x_0 \in \mathscr{S}$. Then $\mu_i \xrightarrow[w]{} \delta_{x_0}$.

The example generalizes the situation where a continuous function on a compact subset of $\mathbf{R}^n$ acquires its absolute minimum in one point only. Here we do not require any regularity of the function f, consequently we have to be more specific: we say that an energy function f *acquires its absolute minimum at* $x_0 \in \mathscr{S}$, if for every neighbourhood B of x_0 there exists $\varepsilon > 0$ such that

$$m\big(\{\omega \mid f(\omega) \leq \varepsilon\} \setminus B\big) = 0.$$

Evidently it is now enough to verify the following: for every neighbourhood B of x_0 we have $\lim_{i\to\infty} \pi_i(B) = 1$.

Example 2. Suppose that the state space $\mathscr{S}$ of a simulated annealing process $(P_i)_{i=1}^{\infty}$ is a compact subset of $\mathbf{R}^n$ equipped with inherited topology, and let m be the normalized restriction of the Lebesgue measure on $\mathscr{S}$; we assume also that $\mathscr{S}$ has interior points. The energy function f is assumed to be continuous and $M_f = \{x_1, x_2, \ldots, x_r\}$, where all the x_i are interior points of $\mathscr{S}$. Suppose also that for each $i \in \{1, \ldots, r\}$ we have

$$f(x) \underset{x\to x_i}{\sim} g_i(x - x_i),$$

where $g_i\colon \mathbf{R}^n \mapsto \mathbf{R}$ is homogeneous of degree $a_i > 0$. We may assume that $a_1 = a_2 = \cdots = a_q > a_{q+1} \geq \cdots \geq a_r$, where $1 \leq q \leq r$. Suppose that $\delta(Q_{is+1}Q_{is+2}\cdots Q_{(i+1)s}) \leq d < 1$ for all $i \geq 1$ and some $s \geq 1$. Suppose also that there exists $\varepsilon \in]0,1[$ such that for $i \geq 1$ either one of the conditions (iii′) and (iii″) is valid or that

$$\text{(iii}^*\text{)} \qquad T_i \geq \frac{(1+\varepsilon)s\Delta f}{\log(i+2)} \qquad \text{and} \quad \lim_{i\to\infty} \frac{T_i}{T_{[i-i^{1-\varepsilon/2}]}} = 1.$$

Then $\mu_i \xrightarrow[w]{} \pi$, for

$$\pi = \sum_{j=1}^{q} D_j \delta_{x_j}.$$

The coefficients D_j have a representation of the form

$$D_j = \frac{v_j}{\sum_{i=1}^{q} v_i}, \qquad j = 1, \ldots, q,$$

with

$$v_j = m(\{y \in \mathbf{R}^n \mid g_j(y) \leq 1\}), \qquad j = 1, \ldots, q.$$

Remark. We may apply the above theorem to the case where f is twice continuously differentiable in some neighbourhood of the set $M_f = \{x_1 \ldots, x_r\}$ with positive definite Hessian matrixes in M_f. Then it is easily seen that the coefficients D_j are inversely proportional to the square roots of the Hessian determinants at corresponding points x_j. In the case when M_f consists of lower dimensional manifolds the convergence of the sequence π_i may be ascertained by assuming f to be smooth enough (c.f. [8]).

The final example shows that the weak convergence of the sequence μ_i does not necessarily follow if f is only assumed to be continuous—even if cooling is logarithmic and the generating distributions Q_i behave well. Note that the example furnishes an counterexample where M_f consists only of two interior points of $\mathscr{S}$. Thus the non-existence of the weak limit is *not* due to non-tightness or irregular boundary behaviour.

Example 3. Suppose that $\mathscr{S} = [0,1]$ and m is the Lebesque measure restricted on $\mathscr{S}$. There exists a continuous energy function $f\colon \mathscr{S} \mapsto [0,\infty[$ such that $M_f = \{\frac{1}{3}, \frac{2}{3}\}$, but the sequence π_i is not weakly convergent for the choice $T_i = C/\log(i+2)$, $C > 0$ being arbitrary.

We point out that under suitable extra conditions it is possible to get rid of the asumption that the generating distributions are symmetric. The considerations become more difficult, but still one obtains similar theorems. We also believe that it should be possible to obtain more delicate conditions for convergence, comparable to those presented for the discrete case (cf. eg. [7], [2]); evidently the assumption on general state space makes the proofs much more difficult even for partial results.

4. Estimates for $\delta(Q^r)$

In the enunciation of Theorems 2.4 and 3.1 we face the following condition:

$$\delta(Q^{(n_i,n_{i+1})}) = \delta(Q_{n_i+1}Q_{n_i+2}\cdots Q_{n_{i+1}}) \leq d. \tag{4.1}$$

The question arises: what is the connection between the n_i and the properties of Q_j in (4.1). If we assume, loosely speaking, that the sequence Q_i is slowly varying, we may assume in (4.1) that $Q_{n_i+1} \approx Q_{n_i+2} \approx \cdots \approx Q_{n_{i+1}}$. Thus we restrict ourselves to the study of the condition

$$\delta(Q^r) < d < 1, \tag{4.2}$$

which we shall encounter in this section. It is to be noted that, regardless of the simulated annealing process, the estimates obtained for $\delta(Q^r)$ are of certain independent interest.

Certainly the most important type of a continuous state space is a bounded subset of $\mathbf{R}^n$; usually $\mathscr{S}$ is a parallelepiped. In this case the generating distributions Q_i, used in practice, typically have the form

$$Q_i(x, A) = \mu_i(A - x), \tag{4.3}$$

where μ_i does not depend on x. Naturally (4.3) does not necessarily make sense if x is near the boundary $\partial\mathscr{S}$, so that (4.3) must be supplemented by a condition which states that the boundary $\partial\mathscr{S}$ is reflecting. This simply means that if the chosen new point lies out of $\mathscr{S}$, it is carried back by suitable reflections with respect to the hyperplanes which determine the boundary $\partial\mathscr{S}$.

The main results of this section are Theorems 4.2 and 4.3. Theorem 4.2 estimates $\delta(Q^r)$ for general μ. Theorem 4.3 states, roughly speaking, that for (4.2) to hold r must be chosen asymptotically as

$$r \sim (\text{width of } \mu)^{-2}, \tag{4.4}$$

in the case that μ is an equidistribution or an (approximately) normal distribution of a given size. Some additional interest for (4.4) comes from the fact that in some practical implementations 'convergence' of the simulated annealing process would mean that the 'width' of the distribution μ_i in (4.3) finally decreases to zero along with T_i.

Now we turn to the details. We choose $\mathscr{S} = [0,1]^n$, where $n \in \mathbf{N}$. The measure m is the usual Lebesgue measure on $\mathscr{S}$. In order to be able to handle the reflecting boundary, we interpret $\mathscr{S}$ as a (topological) subset of $\mathscr{T}^n$, where $\mathscr{T}^n = \mathbf{R}^n/2\mathbf{Z}^n$ (this means that the points $x, y \in \mathbf{R}^n$ are indentified if $\frac{1}{2}(x-y) \in \mathbf{Z}^n$). Addition on the torus is defined in the usual way. We denote by $\mathscr{I}$ the canonical surjection $\mathscr{I}$: $R^n \to \mathscr{T}^n$ and by $\widetilde{m}$ the usual Haar measure on $\mathscr{T}^n$

multiplied by 2^n. Clearly $\tilde{m}\,|\,\mathscr{S}=m$. As usual, we denote the family of all the Borel sets of a topological space X by $\mathscr{B}(X)$.

We observe first that the reflections we need are easy to define on $\mathscr{T}^n$: in fact the mappings R_k: $\mathscr{T}^n \to \mathscr{T}^n$,

$$R_k : x=(x_1,\dots,x_n)\mapsto(x_1,\dots,x_{k-1},-x_k,x_{k+1},\dots,x_n) \quad k=1,\dots,n$$

and all their possible combinations together comprise all the reflections with respect to the boundary of $\mathscr{S}$ that are needed to bring any point $x\in\mathscr{T}^n$ into $\mathscr{S}$. We denote by R_i, $i=n+1,\dots,2^n$, the different combinations of the reflections R_i, other than $R_1,\dots,R_n$. In total the set $\{R_i\}_{i=1}^{2^n}$ consist of the mappings

$$x=(x_1,\dots,x_n)\mapsto(\pm x_1,\pm x_2,\dots,\pm x_n), \qquad x\in\mathscr{T}^n$$

where the signs are chosen in all possible ways. Write $S^o=]0,1[^n$. It is clear that we may write

$$\mathscr{T}^n=\sum_{i=1}^{2^n} R_i(S^o)+S_1, \tag{4.5}$$

where $\sum$ stands for disjoint union and S_1 is a subset of $\mathscr{T}^n$ with the property $\tilde{m}(S_1)=0$.

Definition 4.1. We denote by $\Lambda_s(\mathscr{T}^n)$ the probability measures μ on $\mathscr{T}^n$ which are absolutely continuous with respect to $\tilde{m}$ and remain *invariant* under the basic reflections:

$$\mu(R_i(A))=\mu(A), \qquad i=1,\dots,n,\ A\in\mathscr{B}(\mathscr{T}^n).$$

From the above definition it readily follows that each $\mu\in\Lambda_s$ remains invariant under all the reflections R_i, $i=1,\dots,2^n$.

Given any $\mu\in\Lambda_s(\mathscr{T}^n)$ we may define a transition probability Q_μ on $\mathscr{S}$ as follows:

$$Q_\mu(x,A)=\sum_{i=1}^{2^n}\mu(R_i(A)-x), \tag{4.6}$$

where $x\in\mathscr{S}$ and $A\in\mathscr{B}(\mathscr{S})$. The verification that Q_μ is a transition probability is straightforward using (4.5) and the fact $\mu\ll\tilde{m}$. It is not difficult to show that Q_μ is symmetric, and thus a generating distribution of the type outlined in the beginning of this section: $Q_\mu(x,\cdot)$ is the measure μ centered at x with the overlapping part reflected.

A small computation, using definition (4.6) and the properties of the reflections R_j, provides us with the result that for $\mu_1,\mu_2\in\Lambda_s(\mathscr{T}^n)$ one has $Q_{\mu_1}Q_{\mu_2}=Q_{\mu_1*\mu_2}$, $\mu_1*\mu_2$ standing for the convolution on $\mathscr{T}^n$.

Theorem 4.2 below shows that considering $\mathscr{S}$ as a subset of $\mathscr{T}^n$ leads us to useful estimates for $\delta(Q_{\mu_1}Q_{\mu_2}\cdots Q_{\mu_k})$.

Theorem 4.2. *Let us denote by a_k the kth Fourier-coefficient of the measure $\mu \in \Lambda_s(\mathcal{T}^n)$; $a_k = \int_{\mathcal{T}^n} \mu(dx) \exp(-\pi i k \cdot x)$, $k \in \mathbf{Z}^n$. Denote $(1,0,0,\ldots,0) = e_1 \in \mathbf{Z}^n$. Let $r \in \mathbf{Z}^+$. Then*

$$\sup_{k \in \mathbf{Z}^+} |a_{2ke_1}|^r \leq \delta((Q_\mu)^r) \leq \Big(\sum_{\substack{k \neq 0 \\ k \in \mathbf{Z}^n}} |a_k|^{2r} \Big)^{1/2}. \tag{4.7}$$

Proof. The upper estimate is obtained by showing that $\delta(Q_\mu) \leq \|\mu - 2^{-n}\tilde{m}\|_{\mathcal{T}^n}$ and estimating the right hand side by Hölder's inequality and Parseval's formula. The lower estimate is obtained by establishing

$$2\delta(Q_\mu) \geq \tfrac{1}{2}\|\mu_0 - m\| \geq \tfrac{1}{2}|a_{2ke_1}|.$$

Here $\mu_0 = \delta_0 Q_\mu$, δ_0 being the Dirac delta measure at 0.

Suppose next that $\mu \in \Lambda(\mathbf{R}^n)$. In a natural way μ defines a measure $\tilde{\mu}$ on $\mathcal{T}^n$ via the canonical surjection $\mathcal{S}$: $R^n \to \mathcal{T}^n$:

$$\tilde{\mu}(A) = \mu(\mathcal{S}^{-1}(A)).$$

We call $\tilde{\mu} \in \Lambda(\mathcal{T}^n)$ the retarded distribution corresponding to the distribution $\mu \in \Lambda(\mathbf{R}^n)$ (actually we need the concept of the retarded distribution only in order to define the normal distribution on $\mathcal{T}^n$).

The following theorem involves the most important types of generating distributions.

Theorem 4.3.

a) *Let $d \in]0,1[$. Let $\mu \in \Lambda(\mathbf{R}^n)$ be the normal distribution with zero mean and covariance matrix $\sigma^2 I$, $\sigma > 0$. Then there exist constants $C_1(d,n)$, $C_2(d,n) > 0$ such that*

$$\begin{aligned} &\delta((Q_{\tilde{\mu}})^r) \leq d, \qquad \textit{if} \quad r \geq C_1(d,n)\sigma^{-2} \quad \textit{and} \\ &\delta((Q_{\tilde{\mu}})^r) \geq d, \qquad \textit{if} \quad r \leq C_2(d,n)\sigma^{-2}. \end{aligned}$$

b) *Let $d \in]0,1[$. Let $\mu \in \Lambda(\mathbf{R}^n)$ be the equidistribution on the set $A_\varepsilon = \{x \mid |x_i| \leq \varepsilon,\ i = 1,\ldots,n\}$, where $\varepsilon \in]0,\frac{1}{2}[$. Then there exist constants $C_1'(d,n)$, $C_2'(d,n) > 0$ such that*

$$\begin{aligned} &\delta((Q_{\tilde{\mu}})^r) \leq d, \qquad \textit{if} \quad r \geq C_1'(d,n)\varepsilon^{-2} \quad \textit{and} \\ &\delta((Q_{\tilde{\mu}})^r) \geq d, \qquad \textit{if} \quad r \leq C_2'(d,n)\varepsilon^{-2}. \end{aligned}$$

A proof is obtained by applying the previous theorem. We skip the details.

Remark. Since ε and σ measure the width of the generating distribution, the results obtained above give rise to formula (4.4). Also, suppose that $\mathscr{S} = [0,1]^n$ and $Q_i = Q_\mu$, where μ is as in Theorem 4.3.a, otherwise we make the same assumptions as in Example 3.1. Then an appropriate condition for the temperature sequence is

$$T_k \sigma^2 \log(k+2) \geq C,$$

(together with condition (iii$''$) of Section 3) where

$$C = \frac{\tilde{\Delta} f \left[\coth^{-1}(2^{1/2n}) + 1\right]}{\pi^2}.$$

The main content of the above formulae is that they describe the relation of the width σ and an appropriate choice of the temperature sequence (T_k).

References

[1] Anily, S., and A. Federgruen: Simulated annealing methods with general acceptance probabilities. - J. Appl. Probab. 24, 1987, 657–667.

[2] Chiang Tzuu-Shuh and Chow Yunshyong: On the convergence rate of annealing processes. - SIAM J. Control Optim. 26, 1988, 1455–1470.

[3] Dobrushin, R.: Central limit theorems for non-stationary Markov chains, II. - Theory Probab. Appl. 1, 1956, 329–383 (English translation of Teor. Veroyatnost. i Primenen. 1, 1956, 365–425).

[4] Geman, S., and D. Geman: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. - IEEE Trans. Pattern Anal. Mach. Intell. 6, 1984, 721–741.

[5] Geman, S., and C. Hwang: Diffusions for global optimization. - SIAM J. Control Optim. 24, 1986, 1031–1043.

[6] Haario, H., and E. Saksman: Simulated annealing process in general state space. - Adv. Appl. Probab. 23, 1991, 866–893.

[7] Hajek, B.: Cooling schedules for optimal annealing. - Math. Oper. Res. 13, 1988, 311–329.

[8] Hwang, C.: Laplace's method revisited: weak convergence of probability measures. - Ann. Probab. 8, 1980, 1177–1182.

[9] Isaacson, D., and R. Madsen: Markov chains: Theory and applications. - John Wiley & Sons, New York, 1976.

[10] Iosifescu, M.: On two recent papers on ergodicity in nonhomogeneous Markov chains. - Ann. Math. Statist. 43, 1972, 1732–1736.

[11] Kushner, H.J.: Asymptotic global behaviour for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo. - SIAM J. Appl. Math. 47, 1987, 169–185.