J. Appl. Prob. **33**, 1127–1140 (1996)
Printed in Israel

CONVERGENCE PROPERTIES OF SIMULATED ANNEALING FOR CONTINUOUS GLOBAL OPTIMIZATION

M. LOCATELLI,* *Universitá degli studi di Milano*

Abstract

In this paper conditions for the convergence of a class of simulated annealing algorithms for continuous global optimization are given. The previous literature about the subject gives results for the convergence of algorithms in which the next candidate point is generated according to a probability distribution whose support is the whole feasible set. A class of possible cooling schedules has been introduced in order to remove this restriction.

SIMULATED ANNEALING; CONTINUOUS GLOBAL OPTIMIZATION; COOLING SCHEDULE

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60C05
SECONDARY 60J24

1. Introduction

This paper deals with some aspects of simulated annealing (SA) applied to continuous global optimization. In Section 2 we will consider in more detail the problem of continuous global optimization and the simulated annealing algorithm applied to it. Sections 3, 4 and the appendix will be dedicated to new results. In this section we consider briefly the origins of and the problems connected with simulated annealing. The name comes from a physical process called annealing, the process for growing crystals, which can be simulated by the Metropolis Monte Carlo method (see [13]). It was first applied to combinatorial global optimization independently by Kirkpatrick [10] and Černy [4]. The main idea (the details will be considered in the next section) is to generate a candidate point according to a certain probability distribution; accept it if it represents an improving move; accept or reject it if it is a hill-climbing move according to a certain criterion which depends on a parameter called the temperature, in analogy with the physical case. The choice of the criterion is one of the problems to face (see [15] for a criterion which is optimal in a defined sense under some conditions). The theory of homogeneous and inhomogeneous Markov chains (see for example [9], [5]) is extensively used for studies of convergence of the algorithm (for results about convergence see for example [1], [6], [7]). Homogeneous theory is useful for the study of an ideal algorithm which can be run for an infinite time before changing the temperature, so that it is possible to reach the stationary distribution of the chain for that temperature. In this case it is typically

Received 17 November 1994; revision received 16 June 1995.
* Postal address: Dipartimento di Scienze dell'Informazione, Via Comelico, 39/41–20135 Milano, Italy.

enough to let the temperature tend to zero, without restrictions on the rate of decrease, to get convergence. The situation is different for real algorithms. In these we start with an initial temperature, for which it is generally simple to obtain a good approximation of the stationary distribution (for its choice see, for example, [8] or [11]). Then, when the temperature is T, we choose a temperature T' close to T so that if we have a good approximation of the stationary distribution for T we can, in a few steps, obtain a good approximation for T'. Notable problems are the rate of change of the temperature and the number of steps before changing it (see [15], [12]). Another interesting problem, which is the one addressed in this paper, is where to generate the next candidate point. It is also interesting to try to understand on which functions simulated annealing performs best (see [16]). Even if born for discrete time–discrete state space problems, simulated annealing has been successively extended to discrete time–continuous state space problems, the ones we will consider, and to continuous time–continuous state space problems, with the use of diffusion processes (see for example [18]), which are of some use only if analog machines are available.

2. The main concepts of simulated annealing

This paper addresses the issue of global optimization. First, we state the problem. We have a function f which is called the objective function and a set X which is called the feasible set. We talk about continuous optimization, i.e. f is supposed to be continuous, while at the moment we only assume $X \subseteq \mathbb{R}^d$ and compact. The problem is to find the global minimum of the objective function over the feasible set, i.e. a point x^* such that

$$f(x^*) \leqq f(x), \qquad \forall x \in X.$$

In real applications we are satisfied when we get 'close' to the global minimum. 'Close' can be interpreted in different ways. Typically we look for closeness of the function values: we want to find a point $\bar{x}$ whose function value is close to the optimal value, i.e. $f(\bar{x}) - f(x^*) \leqq \varepsilon$, where ε is a small positive value; or we look for closeness of the argument values, i.e. $\|\bar{x} - x^*\| \leqq \varepsilon$. We will face the problem using simulated annealing (for more information about global optimization techniques see [17] or [19]). SA does not guarantee deterministic convergence but convergence in probability, as we shall see. Now we introduce briefly the main ideas of SA; we do this by describing a typical step of the algorithm (or rather of the meta-algorithm, since there are many things to be instantiated). Let x_k denote the current point at iteration k.

1. Generate a new point y_{k+1} according to a probability distribution $D(x_k, \cdot)$, which only depends on x_k.
2. Generate a uniform random number p over [0, 1] and then set

$$x_{k+1} = \begin{cases} y_{k+1} & \text{if } p \leqq A(x_k, y_{k+1}, c_k), \\ x_k & \text{otherwise,} \end{cases}$$

where A is a function with values in [0, 1].

3. Set $c_{k+1} = U(x_0, x_1, \cdots, x_k)$.
4. Check if the stopping criterion is true; if not, go back to 1.

The things to be instantiated to get a real algorithm are:

- the Markov kernel D, through which a candidate point is generated;
- the acceptance probability A, which gives the probability that the candidate point is accepted as the next iterate;
- the cooling schedule U, through which the c_k, positive quantities called temperatures and used in A at every step, are generated;
- the stopping criterion.

In what follows the function A will always be the Metropolis function:

$$A(x_k, y_{k+1}, c_k) = \min\left\{1, \exp\left(\frac{f(x_k) - f(y_{k+1})}{c_k}\right)\right\},$$

which always accepts the descent steps.

It is possible to show that, under some conditions, the sequence $\{f(x_k)\}$ converges in probability to f^*. Among these conditions we remember:

- $\exists d > 0 : \forall B \subseteq X$ measurable, $\forall x \in X$, $D(x, B) \geqq d\mu(B)$,
- $c_k \xrightarrow{P} 0$,

where μ denotes the Lebesgue measure and $\xrightarrow{P}$ convergence in probability. For a deeper discussion of these conditions see [2]. It is interesting to note that the first condition implies that we sample over all the set X at every step. We now cite Hajek's result for the discrete case, when X is a finite set of states and a neighbourhood structure is defined on it. Hajek's theorem gives a necessary and sufficient condition for the algorithm to converge in probability to the optimum, when the sequence of temperatures is deterministic, i.e. when the temperature at step k is already known at the beginning. We have that $\{f(x_k)\}$ converges in probability to f^* if and only if c_k is chosen so that

$$\sum_{k=1}^{\infty} \exp\left(-\frac{d^*}{c_k}\right) = \infty,$$

for a given value d^*. If c_k has the form $c_k = c/\log(k+1)$ then we must have $c \geqq d^*$. Therefore the convergence is possible if and only if the temperatures are decreased slowly enough (for more information see [7]). The interesting fact is that if we are at point x_k, the next candidate point y_{k+1} is generated in the neighbourhood of x_k and not in all the feasible region. This prompts us to answer the question if, in the continuous case, it is possible to generate points around x_k at step k, without imposing to give positive probability of generation to any set of positive Lebesgue measure. We then ask if we can move from a situation in which any step can be considered a global one, towards the situation of the discrete case where every step is a local one. That is the main issue of the paper and will be treated in the two following sections, the first one dediciated to the simpler case in which the optimum value is *a priori* known, and the second dedicated to the more general case in which the optimum value is not known.

3. The case of known optimum value

We begin with the case in which the optimum value is already known because it contains the idea that makes the more general case work. The assumptions are the following.

1. The candidate point y_{k+1} is generated according to a distribution $D(x_k, \circ)$ whose support is a sphere of radius R with center in x_k, giving positive probability to any set whose intersection with the sphere has positive Lebesgue measure (from globality towards locality of the steps); in a formal way this condition can be stated as:

$$\exists d, D > 0 : \forall B \subseteq S(x, R) \cap X, \forall x \in X, d\mu(B) \leqq D(x, B) \leqq D\mu(B).$$

A consequence of this assumption is that:

$$\mu(A) = 0 \Rightarrow \forall x \in X, \quad D(x, A) = 0.$$

2. The feasible region X is connected.
3. The interior of the feasible region, denoted by X°, is non-empty and connected.
4. The function has a finite number of global minima.
5. Let O be the set of global minima; for any $x_i^* \in O$ we consider the set

$$M_i = \{x : \|x - x_i^*\| < \rho\} \cap X, \tag{1}$$

where ρ is chosen so that $M_i \cap M_j = \emptyset$, $\forall i \neq j$ and $\rho \leqq \frac{1}{4}R$; the sets

$$B_\varepsilon \cap M_i = \{x \in X : f(x) \leqq f^* + \varepsilon\},$$

must have, for any positive ε, positive Lebesgue measure.

6. Denoting with $X' = X - X^\circ$ the border of the feasible region, we must have that $\mu(X') = 0$, i.e. the border has null Lebesgue measure.
7. The starting point of the algorithm must be with probability 1 in X°.

We notice that, under Assumption 6, Assumption 7 is satisfied, for instance, if we sample x_0 from a uniform distribution over X; Assumption 7, together with Assumptions 1 and 6, implies that with probability 1 the sequence of points generated by the algorithm never goes out of X°.

Notice that $X - \cup M_i$ is still compact. Then we can find the minimum of f in this set, which we indicate with $\bar{f} > f^*$. It is then obvious that all the points whose function value differs from f^* less than $\bar{f} - f^* = \gamma > 0$, must be in the M_i's. We choose $\bar{\varepsilon} > 0$ so that:

$$2\bar{\varepsilon} \leqq \gamma. \tag{2}$$

With the assumptions above it is possible to prove the following theorem.

Theorem 1. *Using the following temperatures*

$$c_k = \begin{cases} f(x_k) - f^* & \text{if } f(x_k) - f^* > \bar{\varepsilon}, \\ 0 & \text{otherwise}, \end{cases}$$

we have

$$\lim_{k\to\infty} \boldsymbol{P}\{x_k \in B_\varepsilon\} = 1, \qquad \forall \varepsilon > 0.$$

The proof of this theorem is given in the appendix. We notice that in the cooling schedule we could also multiply $[f(x_k)-f^*]$ by a positive weight M (possibly depending on $[f(x_k)-f^*]$) which controls the probability of accepting ascent steps when we are not in $B_{\bar{\varepsilon}}$. We do not consider this possiblity in what follows, but it can be interesting in the implementation of an algorithm. In [3] a similar cooling schedule is presented and numerically investigated, but no proof of convergence is given.

4. The general case

In this section we will not assume knowledge of the optimal value. We keep the assumptions of the previous section. What we cannot keep from the previous section is the rule to update the temperature, because before we needed to know the value of the optimum. On the other hand, we can think of using the best value so-far observed, which is indicated by f_k^*, in place of f^*, so that the new updating rule is:

$$c_k = \begin{cases} f(x_k) - f_k^* & \text{if } f(x_k) - f_k^* > \bar{\varepsilon}, \\ t_k & \text{otherwise.} \end{cases}$$

We introduced the deterministic non-increasing sequence of emperatures $\{t_k\}$, in place of the 0 of the previous section, because this would act like a trap (not accepting non-improving points) in a place which, very likely, is the wrong one. Of course t_k must decrease to 0. The point is if there exists a sequence $\{t_k\}$ through which we can ensure convergence of $f(x_k)$ to f^* in probability and, if yes, which should be the rate of decrease (remember that for the discrete case we had the inverse of the logarithm). The quantities N and ΔF will have the same meaning as in the appendix. Let us start by computing the expected time of the first visit in B_ε.

Lemma 1. *The expected time until the first visit in B_ε is finite if*

$$t_k \geqq (1+\mu)\frac{N\Delta F}{\log k}, \qquad \mu > 0.$$

Proof. The expected time until the first visit in B_ε is equal to

$$\sum_{k=1}^{\infty} k\boldsymbol{P}\{x_1, \cdots, x_k \notin B_\varepsilon \mid x_0 \notin B_\varepsilon\} \times \boldsymbol{P}\{x_{k+1} \in B_\varepsilon \mid x_0, x_1, \cdots, x_k \notin B_\varepsilon\}.$$

We want to know when this is a finite quantity, and in order to do that we must see when we can find a finite upper bound for it. We simply limit from above the second probability in the argument of the series with one. The first probability in the argument is the probability of never visiting B_ε in the first k steps. It can also be written in the following way:

$$\boldsymbol{P}\{x_1,\cdots, x_N \notin B_\varepsilon \mid x_0 \notin B_\varepsilon\} \times \cdots \times \boldsymbol{P}\{x_k,\cdots, x_{\lfloor k/N\rfloor N} \notin B_\varepsilon \mid x_{\lfloor k/N\rfloor N-1},\cdots, x_0 \notin B_\varepsilon\}.$$

Let us consider the probability of a visit in B_ε in the first N steps; we have that $\exists \delta$, $\gamma(\varepsilon) > 0$ and an integer N such that

$$\boldsymbol{P}\{\text{at least one visit in } B_\varepsilon \text{ in the first } N \text{ steps}\} \geqq \boldsymbol{P}\{x_N \in B_\varepsilon\} \geqq \delta^N \gamma(\varepsilon) \exp\left(-\frac{N\Delta F}{t_N}\right),$$

where the last lower bound is obtained in the same way we obtained (8) in the appendix. So we have

$$\boldsymbol{P}\{x_1,\cdots, x_N \notin B_\varepsilon \mid x_0 \notin B_\varepsilon\} \leqq 1 - \delta^N \gamma(\varepsilon) \exp\left(-\frac{N\Delta F}{t_N}\right),$$

and in a similar way we can show that

$$\boldsymbol{P}\{x_{iN+1},\cdots, x_{(i+1)N} \notin B_\varepsilon \mid x_0,\cdots, x_{iN} \notin B_\varepsilon\} \leqq 1 - \delta^N \gamma(\varepsilon) \exp\left(-\frac{N\Delta F}{t_{(i+1)N}}\right).$$

In this way we can bound from above the expected time before the first visit to B_ε by

$$\sum_{k=1}^{\infty} k \prod_{i=1}^{\lfloor k/N\rfloor} \left[1 - \delta^N \gamma(\varepsilon) \exp\left(-\frac{N\Delta F}{t_{(i+1)N}}\right)\right].$$

Now we use the fact that Π can be written as $\exp(\Sigma \log)$ and the fact that $\log(1-2x) < -x$ if $0 < x < 1/2$ to bound from above the previous sum by:

$$\sum_{k=1}^{\infty} k \exp\left[-\sum_{i=1}^{\lfloor k/N\rfloor} \frac{\delta^N \gamma(\varepsilon)}{2} \exp\left(-\frac{N\Delta F}{t_{(i+1)N}}\right)\right].$$

We choose $t_k = (1+\mu)(N\Delta F/\log k)$, $\mu > 0$. The sum becomes:

$$\sum_{k=1}^{\infty} k \exp\left[-\frac{\delta^N \gamma(\varepsilon)}{2} \sum_{i=1}^{\lfloor k/N\rfloor} \left[\frac{1}{(i+1)N}\right]^{1/(1+\mu)}\right],$$

which can be bounded from above by

$$\sum_{k=1}^{\infty} k \exp\left[-\frac{\delta^N \gamma(\varepsilon)}{2} \lfloor k/N\rfloor \left[\frac{1}{k+N}\right]^{1/(1+\mu)}\right], \tag{3}$$

which is a convergent series.

Before proceeding we consider the constant $N\Delta F$ by which the inverse of the logarithm is multiplied in the result of the previous lemma and in following lemmas, and whose origin is explained in the appendix: its value can possibly be decreased and we propose to explore in a forthcoming paper the possibility of finding a lower constant which is related to the form of the problem and by which we can give not only a sufficient but also a necessary condition for convergence, as done in [7] for the discrete case. We notice

that the constant given here is analogous to the constant which can be found in [1] for the discrete case. The following lemma gives us a limitation from above of the expected time before a new visit to B_ε after it has already been visited at least once.

Lemma 2. *After we have visited B_ε once, with $\varepsilon \leqq \bar{\varepsilon}$, the expected time before a new visit is limited above by a constant E independent of the number of steps.*

Proof. Let us assume we are outside B_ε and we have already visited it once. There are two possibilities:

(i) we are in $B_{\varepsilon+\bar{\varepsilon}} \backslash B_\varepsilon$: in this case the value of the temperature can be t_k, but because of the choice (2) for the value of $\bar{\varepsilon}$ and of Assumptions 1 and 5, we can get to B_ε in one step with positive probability, independent of k;

(ii) we are outside $B_{\varepsilon+\bar{\varepsilon}}$: in this case the temperature is always greater than ε and we can follow the proof of Lemma 6 to show that there is a positive probability, independent of k, of getting in N steps to B_ε.

From the fact that there is always a positive probability, independent of k, of getting to B_ε in N steps, it follows that the expected number of steps outside B_ε is limited above by a constant E.

The next thing we want to show is that the expected time inside B_ε is finite but increases to infinity with the number of steps. In this way we have that we expect a finite time before getting to B_ε for the first time and after that we expect to stay alternatively inside and outside B_ε, but while we do not expect to stay outside more than a fixed time, we expect to stay inside for more and more time, so that as time goes by we spend a lower and lower percentage of time outside.

Lemma 3. *The expected time spent inside B_ε increases to infinity as the number of steps increases to infinity.*

Proof. Inside B_ε, $\varepsilon \leqq \bar{\varepsilon}$, the temperature is always given by the t_k. Now assume that we get to B_ε at time K. We can limit from below the expected time inside B_ε with the expected time inside B_ε for the case in which the temperature is kept fixed to t_K (this is because the t_k are non-increasing and it becomes more and more difficult to come out of B_ε). We set

$$\boldsymbol{P}(t_K) = \boldsymbol{P}\{x_{k+1} \in B_\varepsilon^{\mathrm{C}} \mid x_k \in B_\varepsilon;\, t_K\},$$

where A^{C} denotes the complement of A. When t_K is fixed, the number of steps inside B_ε follows a geometric distribution with parameter $\boldsymbol{P}(t_K)$. Its expected value is thus $1/\boldsymbol{P}\{t_K\}$ (see [14]). We have that $\boldsymbol{P}(t_K) \to 0$ as $K \to \infty$. Indeed, we can write $\boldsymbol{P}(t_K)$ as

$$\boldsymbol{P}\{x_{k+1} \in B_{\varepsilon+t_K^{1/2}} \backslash B_\varepsilon \mid x_k \in B_\varepsilon\} + \boldsymbol{P}\{x_{k+1} \in B^{\mathrm{C}}_{\varepsilon+t_K^{1/2}} \mid x_k \in B_\varepsilon\}.$$

The measure of $\bar{B}_K = B_{\varepsilon+t_K^{1/2}} \backslash B_\varepsilon$ tends to 0 as K increases. Indeed $\bigcap_{K=1}^{\infty} \bar{B}_K = \emptyset$. Then $\mu(\bar{B}_K) \to 0$ and in view of Assumption 1 we must have that the first probability tends to 0. The second probability tends to 0 as well since the acceptance probability outside $B_{\varepsilon+t_K^{1/2}}$ is not greater than $\exp(-1/t_K^{1/2}) \to 0$. Therefore we have that the limitation from below of the expected time inside B_ε diverges as $K \to \infty$, which proves the result.

Now we set:

$$\boldsymbol{E}_K[B_\varepsilon] = \text{expected time spent in } B_\varepsilon \text{ after step } K.$$

$$\boldsymbol{E}_K[B_\varepsilon^{\mathrm{C}}] = \text{expected time spent outside } B_\varepsilon \text{ after step } K.$$

$$A_K = \text{the event that } B_\varepsilon \text{ has been already visited in the first } K \text{ steps.}$$

We want to find the fraction of time spent in B_ε after instant K.

Lemma 4. *The quantity*

$$\boldsymbol{E}_K[B_\varepsilon^{\mathrm{C}} \mid A_K^{\mathrm{C}}] \times \boldsymbol{P}[A_K^{\mathrm{C}}],$$

can be bounded above by a constant E'.

Proof. We show that this quantity goes to 0 as K goes to infinity, when t_k decreases to 0 slowly enough. The fact is not so surprising, because analyzing the two factors above, it is possible to notice that the slower we decrease t_k the smaller both become. We prove this in what follows. Remembering the way we found the expected time before the first visit in B_ε when $t_k = (1+\mu)(N\Delta F/\log k)$, we have that, being at time K outside B_ε, the expected time before the first visit to it is bounded above by

$$\sum_{k=1}^{\infty} k \prod_{i=1}^{\lfloor k/N \rfloor} \left[1 - \delta \exp\left(-\frac{N\Delta F}{t_{K+iN}}\right)\right]$$

$$\leqq \sum_{k=1}^{\infty} k \exp\left(-\frac{\delta}{2} \sum_{i=1}^{\lfloor k/N \rfloor} \left[\frac{1}{K+iN}\right]^{1/(1+\mu)}\right)$$

$$\leqq \sum_{k=1}^{\infty} k \exp\left(-\frac{\delta}{2N} \frac{k}{[K+k]^{1/(1+\mu)}}\right)$$

$$= \sum_{k=1}^{K} k \exp\left(-\frac{\delta}{2N} \frac{k}{[K+k]^{1/(1+\mu)}}\right) + \sum_{k=K}^{\infty} k \exp\left(-\frac{\delta}{2N} \frac{k}{[K+k]^{1/(1+\mu)}}\right)$$

$$\leqq \sum_{k=1}^{K} k \exp\left(-\frac{\delta}{2N} \frac{k}{[2K]^{1/(1+\mu)}}\right) + \sum_{k=K}^{\infty} k \exp\left(-\frac{\delta}{2N} \frac{k}{[2k]^{1/(1+\mu)}}\right).$$

The first sum can be bounded from above by $\Sigma_{k=1}^{K} k = \frac{1}{2}K(K+1)$, which is probably not the best bound, but it is enough for our interests. About the second sum we only need to observe that it is the sum of the tail of a finite series, then it tends to 0 as K increases to infinity. Then, the expected time till the first visit to B_ε, being outside at time K, is at most $O(K^2)$. Now we compute the probability of no visit in B_ε until time K. Remembering that the probability of no visit in K steps is lower than

$$\prod_{i=1}^{\lfloor k/N \rfloor} \left[1 - \delta^N \exp\left(-\frac{N\Delta F}{t_{iN}} \right) \right],$$

and in a way which is the same as in the proof of Lemma 1, we can see that this quantity is of the order of $\exp(-K^{\mu/(1+\mu)})$. We can finally say that the product of the expected time until the first visit to B_ε, being outside of it at time K, and the probability of no visit in B_ε until time K, is of order not greater than $K^2 \exp(-K^{\mu/(1+\mu)})$ and then it tends to 0 as K tends to infinity. We can then conclude that this product can be limited by a constant, as we wished to prove.

We can finally prove the following theorem.

Theorem 2. *The fraction of time spent inside B_ε after time K tends to* 1 *as* $K \to \infty$.

Proof. First we see $\boldsymbol{E}_K[B_\varepsilon^C]$ as

$$\boldsymbol{E}_K[B_\varepsilon^C \mid A_K] \times \boldsymbol{P}[A_K] + \boldsymbol{E}_K[B_\varepsilon^C \mid A_K^C] \times \boldsymbol{P}[A_K^C].$$

As already seen in Lemma 2, the first addend can be bounded above by a constant E. We have also shown in Lemma 4 that the same is true for the second addend, limited by a constant E', so that all the sum can be bounded above by a constant E''. Remembering, as seen in Lemma 3, that $\boldsymbol{E}_K[B_\varepsilon]$ increases as K increases (and its limit for $K \to \infty$ is infinite), we have that the fraction of time spent inside B_ε after time K can be limited from below by

$$\frac{\boldsymbol{E}_K[B_\varepsilon]}{\boldsymbol{E}_K[B_\varepsilon] + E''},$$

which goes to 1 as K increases to infinity.

It is also possible to work with probabilities and prove the following theorem.

Theorem 3. *We have that*

$$\forall \varepsilon > 0 \qquad \lim_{n \to \infty} P[x_n \in B_\varepsilon] = 1,$$

if

$$t_r \geqq (1+\mu) \frac{N\Delta F}{\log r}, \qquad \mu > 0.$$

Proof. First of all we notice that, for any $n \geqq K$,

$$\boldsymbol{P}[x_n \in B_\varepsilon] = \boldsymbol{P}[x_n \in B_\varepsilon, \textit{ and } B_\varepsilon \text{ visited in the first } K \text{ steps}]$$
$$+ \boldsymbol{P}[x_n \in B_\varepsilon, \textit{ and } B_\varepsilon \text{ not visited in the first } K \text{ steps}].$$

We want to bound from below this sum and show that this limitation goes to 1 as K increases. We can simply limit from below the second term with zero and consider only the first term, written as follows:

$$\boldsymbol{P}[x_n \in B_\varepsilon \mid B_\varepsilon \text{ visited in the first } K \text{ steps}] \tag{4}$$

$$\times \boldsymbol{P}[B_\varepsilon \text{ visited in the first } K \text{ steps}]. \tag{5}$$

In the proof of Lemma 4 we have seen that when $t_k = (1+\mu)(N\Delta F/\log k)$, the probability of never visiting B_ε is $O(\exp(-K^{\mu/(1+\mu)}))$. Then the probability of no visit in B_ε goes to 0 as K increases and then the probability in (5) goes to 1. The same is true for any $t_k \geqq (1+\mu)(N\Delta F/\log k)$. Now we need to show that, for any K, the probability in (4) tends to 1 as $n \to \infty$. We use the following notation.

N_n = the event of not coming out of B_ε in steps from n till $n+N$, where N is the same as in Lemma 6.

M_n = the event of getting to B_ε at step $n+N$.

P_K = the probability conditioned on the event that B_ε has been already visited in the first K steps.

We have that:

$$\begin{aligned}\boldsymbol{P}_K(x_{n+N} \in B_\varepsilon) &\geqq \boldsymbol{P}_K(x_n \in B_\varepsilon \text{ and } N_n) + \boldsymbol{P}_K(x_n \notin B_\varepsilon \text{ and } M_n) \\ &= \boldsymbol{P}_K(N_n \mid x_n \in B_\varepsilon)\boldsymbol{P}_K(x_n \in B_\varepsilon) + \boldsymbol{P}_K(M_n \mid x_n \notin B_\varepsilon)(1 - \boldsymbol{P}_K(x_n \in B_\varepsilon)).\end{aligned}$$

We define

$$a_i = \boldsymbol{P}_K(x_i \in B_\varepsilon),$$

$$\delta_n = \boldsymbol{P}_K(N_n \mid x_n \in B_\varepsilon),$$

and we observe that δ_n does not decrease as n increases, since the temperatures decrease, and tends to 1 as can be proven in a way absolutely analogous to the proof that $\boldsymbol{P}(t_K) \to 0$ in Lemma 3. Moreover we have seen in Lemma 2 that, independently of the position of x_n outside B_ε, there exists a probability $\eta > 0$ which limits from below $\boldsymbol{P}_K(M_n \mid x_n \notin B_\varepsilon)$. Then we have

$$a_{n+N} \geqq a_n \delta_n + \eta(1 - a_n). \tag{6}$$

If we subtract a_n from both terms of this inequality, we have $a_{n+N} - a_n \geqq a_n(\delta_n - 1) + \eta(1 - a_n)$. The term on the right is not negative if

$$a_n \leqq 1 - \frac{1-\delta_n}{\eta}, \tag{7}$$

and if it is negative it cannot be smaller than $(\delta_n - 1)$. Let us assume that (7) is not true only for a finite number of steps. Then we have that the sequence $\{a_n\}$ is definitely not decreasing and, since it is bounded by 1, it must have a limit. If we take the limit on both sides of (6) and we denote with L the limit $\lim_{n\to\infty} a_n$, we have $L \geqq L + \eta(1-L)$, and then $\eta(1-L) \leqq 0$, which is possible only if $L = 1$. If (7) is not true for an infinite number of times, we have that $\{a_n\}$ can also decrease sometimes but only if we are above the threshold in (7) and, as soon as we get below it, we start increasing again. Since the

maximum decrease is $(\delta_n - 1)$, if we are above the threshold and at the following step a_n decreases, we must have that a_n cannot get below

$$1 - \frac{1-\delta_n}{\eta} - (1-\delta_n),$$

and it can never get below this limit, which tends to 1 as n increases. In this way, we have shown that $a_n \to 1$. We then have:

$$\boldsymbol{P}(x_n \in B_\varepsilon) \geqq a_n \boldsymbol{P}(B_\varepsilon \text{ visited in the first } K \text{ steps}).$$

Taking the limit of both sides as $n \to \infty$ we have that

$$\lim_{n \to \infty} \boldsymbol{P}(x_n \in B_\varepsilon) \geqq \boldsymbol{P}(B_\varepsilon \text{ visited in the first } K \text{ steps}),$$

for any K. Since the term on the right tends to 1 as K increases, we must have that the limit on the left is equal to 1, as we wanted to prove.

5. Conclusion

This paper has dealt with simulated annealing applied to continuous global optimization. In Section 1 we briefly analyzed the origins of and the problems connected with simulated annealing, giving some references for a deeper discussion about them. In Section 2 we defined the continuous global optimization problem and gave details of the simulated annealing algorithm as applied to this problem. We pointed out that existing literature gives convergence results only for the case in which at any step the next candidate point is generated sampling from a distribution whose support is the whole feasible set. Inspired by the results of Section 3 for the case in which the optimum value is *a priori* known, in Section 4 we showed that, by choosing appropriately the cooling schedule and under some conditions on the function and on the feasible set, it is possible to obtain convergence (in probability) to the optimum, even when the next candidate point is sampled from a distribution whose support is not the whole feasible set, but, for example, only a sphere around the present point. This provides better opportunities of exploiting local information.

Acknowledgments

I would like to thank Professor Fabio Schoen and an anonymous referee for the help they gave me in improving the quality of the paper.

Appendix. Proof of Theorem 1

First, we need a lemma.

Lemma 5. *If X is compact and Assumptions* 1–7 *of Section* 3 *are true, given a generic $r > 0$ it is possible to find a finite number $M(r) > 0$ and a finite set of points $x_1, \cdots, x_{M(r)} \in X^\circ$ with the following characteristics*:

1. *for any* $x \in X^\circ$ *there exists a point* x_k *of this set such that* $d(x, x_k) < r$, *where* d *gives the distance between its arguments*;

2. *the set of points contains at least one point inside* B_ε;

3. *if we construct a graph whose nodes are the points of this set and the arc* (i, j), $i \neq j$ *exists if and only if* $d(x_i, x_j) \leqq 2r$, *this graph is connected.*

Proof. Because of the compactness of X it is possible to find a $\frac{1}{2}r$-net, i.e. a finite set of points $y_1, \cdots, y_{M(r)-1}$, such that for any $x \in X^\circ$ we have $d(x, y_i) < \frac{1}{2}r$ for at least one point y_i. The y_k do not necessarily belong to X°. So we construct a new set of $M(r)-1$ points in this way: $x_k = y_k$ if $y_k \in X^\circ$, otherwise we choose as x_k one of the points belonging to X° whose distance from y_k is lower than $\frac{1}{2}r$ (one must exist otherwise y_k can be removed). We have that for any point $x \in X^\circ$ such that $d(x, y_k) < \frac{1}{2}r$

$$d(x, x_k) \leqq d(x, y_k) + d(x_k, y_k) < \tfrac{1}{2}r + \tfrac{1}{2}r = r,$$

where we used the triangular inequality. In this way the set of the x_k satisfies the first characteristic. The second characteristic is easily satisfied. Assumption 5 guarantees the existence of a point in B_ε belonging to X°; if we add this point to the set of the x_k, the first characteristic is still satisfied. About the third characteristic, we start considering the point that we will denote by x_1. Let us assume by contradiction that no other point of the set is connected in the graph with this one. But that means that all around the sphere $S(x_1, r)$ there must be a crown where no point of X° can fall; indeed, if one point of X° were there, it would be at a distance from any x_k greater than r, which is not possible. But this can mean two things: either there are no more points x_k and in this case we are done, or the region with x_1 is disconnected from the rest of X°, which contradicts Assumption 3 of Section 3. So if there are other points than x_1, at least one must be connected through an arc to x_1. We denote this point by x_2 and we repeat the same reasoning to show that, if there are other points, at least one must be connected with x_1 or x_2. Going on until we have exhausted the points x_k, we obtain a connected graph.

We then need the following lemma.

Lemma 6. *For any* $\varepsilon > 0$ *there exist* $\delta, \gamma(\varepsilon) > 0$ *and an integer* N *such that the probability of getting to* B_ε *in a finite number* N *of steps, i.e.* $\boldsymbol{P}\{x_N \in B_\varepsilon\}$, *is at least*

$$\delta^N \gamma(\varepsilon) \exp(-N\Delta F/\varepsilon),$$

where

$$\Delta F = \max_{x \in X \setminus B_\varepsilon} \max_{y \in S(x,R) \cap X} [f(y) - f(x)].$$

We remark that ΔF *is finite because of the compactness of* X *and the continuity of* f.

Proof. Let us take both $r > 0$ in the previous lemma and a $\rho > 0$ small enough so that if we consider the $M(r)$ points x_i with the characteristics expressed in the previous lemma we have that

(i) $S(x_k, \rho) \subset X^\circ$ which is possible because the x_k's belong to X°;

(ii) any sphere $S(y, R)$, $y \in S(x_k, \rho)$, contains the spheres $S(x_j, \rho)$ for any j such that (k, j) is an arc of the graph built in the previous lemma.

Now we start from a generic point in X° From it we can get to $S(x_k, \rho)$ with probability at least $\delta > 0$, where x_k is one of the points of the set of the previous lemma. The value of δ is, in view of Assumption 1, equal to $d\mu(S(x_k, \rho)) > 0$ and is actually independent of x_k. Since the acceptance probability can be limited below by $\exp(-\Delta F/\varepsilon)$, we have that the probability of getting to $S(x_k, \rho)$ is at least $\delta \exp(-\Delta F/\varepsilon)$. Since the graph is connected we can find a path from x_k to the point $\bar{y}$ in the graph belonging to $B_{\bar{\varepsilon}}$. We denote by $l(x_k, \bar{y})$ the length of the minimum path on the graph between x_k and $\bar{y}$, and we set $N = \max_{k=1,\cdots,M(r)} l(x_k, \bar{y})$. From the sphere $S(x_k, \rho)$ we can go to the sphere of radius ρ and center in the next point of the minimum path to $\bar{y}$ with probability at least $\delta \exp(-\Delta F/\bar{\varepsilon})$, where we already considered the acceptance probability. Continuing like this, we finally get in the neighbourhood of $B_{\bar{\varepsilon}}$ in no more than N steps and with probability at least

$$\delta^N \exp\left(-\frac{N\Delta F}{\bar{\varepsilon}}\right). \tag{8}$$

Once we are in the neighbourhood of $B_{\bar{\varepsilon}}$ we can get with a positive probability $\gamma(\varepsilon)$ to B_ε in one step in view of the choice (2) for $\bar{\varepsilon}$ and of Assumptions 1 and 5. The computation of N is apparently complicated. In the case of X a convex set, we have $N \approx \operatorname{diam}(X)/R$.

Now we are ready to prove the theorem. Indeed we have that the probability of not falling in B_ε in N steps is lower than

$$1 - \delta^N \gamma(\varepsilon) \exp\left(\frac{N\Delta F}{\bar{\varepsilon}}\right),$$

and in nN steps is lower than:

$$\left[1 - \delta^N \gamma(\varepsilon) \exp\left(\frac{N\Delta F}{\bar{\varepsilon}}\right)\right]^n,$$

which goes to zero as n tends to infinity. So the probability of never visiting B_ε is zero and the probability of visiting it once is one. Since when we are in $B_{\bar{\varepsilon}}$ we cannot exit it, because of the zero temperature inside $B_{\bar{\varepsilon}}$ which prevents us from accepting non-improving points, we also have that x_k converges to B_ε with probability one for any choice of $\varepsilon > 0$.

References

[1] ANILY, S. AND FEDERGRUEN, A. (1985) Probabilistic analysis of simulated annealing methods. *Technical Report.* Graduate School of Business, Columbia University, New York.

[2] BELISLE, C. J. P. (1992) Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *J. Appl. Prob.* **29**, 885–892.

[3] BOHACHEVSKY, I. O., JOHNSON, M. E. AND STEIN, M. L. (1986) Generalized simulated annealing for function optimization. *Technometrics* **28**, 209–217.

[4] ČERNY, V. (1985) Thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.* **45**, 41–51.

[5] DOOB, J. L. (1953) *Stochastic Processes.* Wiley, New York.

[6] GELFAND, S. AND MITTER, S. (1985) Analysis of simulated annealing for optimization. *Proc. 24th Conf. on Decision and Control.* pp. 779–786.

[7] HAJEK, B. (1988) Cooling schedules for optimal annealing. *Math. Operat. Res.* **13**, 311–329.

[8] HUANG, M. D., ROMEO, F. AND SANGIOVANNI-VINCENTELLI, A. (1986) An efficient general cooling schedule for simulated annealing. *Proc. ICCAD.* pp. 381–284.

[9] KEMENY, J. G. AND SNELL, J. L. (1976) *Finite Markov Chains*. Springer, New York.

[10] KIRKPATRICK, S., GELATT, C. D. AND VECCHI, M. P. (1983) Optimization by simulated annealing. *Science* **220**, 671–680.

[11] LAM, J. AND DELOSME, J.-M. (1986) Logic minimization using simulated annealing. *Proc. ICCAD.* pp. 348–352.

[12] LAM, J. AND DELOSME, J.-M. (1987) An adaptive annealing schedule. *Technical Report 8608.* University of Yale, New Haven.

[13] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N. AND TELLER, A. H. (1953) Equation of state calculations by fast computer machines. *J. Chem. Pys.* **21**, 1087.

[14] MOOD, A. M., GRAYBILL, F. A. AND BOES, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, New York.

[15] ROMEO, F. (1989) Simulated annealing: theory and applications to layout problems. *PhD thesis.* University of California, Berkeley, CA.

[16] SORKIN, G. B. (1991) Efficient simulated annealing on fractal energy landscapes. *Algorithmica* **6**, 367–418.

[17] TÖRN, A. AND ZILINSKAS, A. (1987) *Global Optimization.* Springer, Berlin.

[18] WONG, E. (1991) Stochastic neural networks. *Algorithmica.* **6**, 466–478.

[19] ZHIGLJAVSKY, A. A. (1991) *Theory of Global Random Search.* Kluwer, Dordrecht.