Simulated Annealing for Noisy Cost Functions

Walter J. Gutjahr and Georg Ch. Pflug
Department of Statistics, OR and Computer Science
University of Vienna

Abstract: We generalize a classical convergence result for the Simulated Annealing algorithm to a stochastic optimization context, i.e., to the case where cost function observations are disturbed by random noise. It is shown that for a certain class of noise distributions, the convergence assertion remains valid, provided that the standard deviation of the noise is reduced in the successive steps of cost function evaluation (e.g., by repeated observation) with a speed $O(k^{-\gamma})$, where γ is an arbitrary constant larger than one.

Key words: Simulated Annealing, stochastic optimization, noisy cost functions.

1 Introduction

The Simulated Annealing algorithm, introduced about ten years ago by Kirkpatrick, Gelatt and Vecchi [8] into the area of combinatorial optimization, has developed into a well–known and thoroughly studied optimization technique with a large (and still rapidly growing) number of applications. (Cf. Laarhoven and Aarts [9], Aarts and Korst [1]; see also the bibliography in [7].)

The standard type of applications concerns optimization problems of the form

$$\underset{i \in S}{\text{Minimize}} \quad f(i),$$

where S is a finite set of feasible solutions which usually exposes a specific combinatorial structure. The algorithm works as follows (see [1], p. 16):

```
procedure SimAnn
begin
    initialize (i_start, c'_0, L_0);
    m := 0;
    i := i_start;
    repeat
        for l := 1 to L_m do
        begin
            generate (j from S_i);
            if f(j) ≤ f(i) then i := j
            else if exp((f(i) − f(j)) / c'_m) > random(0,1) then i := j
        end
    m := m + 1;
    CalculateLength(L_m);
    CalculateControl(c'_m);
    until stopcriterion
end.
```

Therein,

- i_{start}, i and j are feasible solutions from S;
- $c'_0, c'_1, \ldots$ is a (usually decreasing) sequence of values for the *control parameter*, often also called *temperature*;
- The sets S_i form the pre–defined *neighborhood structure*: to each feasible solution $i \in S$, a set $S_i \subseteq S$ of "neighbor solutions" is assigned.
- random(α, β) is a procedure selecting a uniformly distributed (pseudo–)random number from the interval $[\alpha, \beta]$;
- CalculateLength and CalculateControl are procedures updating the values L_m and c'_m; they define the so–called *cooling schedule*.

Most works on Simulated Annealing assume implicitly that to each feasible solution i, the corresponding cost value $f(i)$ can be computed *exactly*. In this article, we investigate the case where $f(i)$ can only be observed *with a random error*. It is assumed that at each time k when we want to to determine $f(i)$, a value $\tilde{f}_k(i)$ can be observed, which is obtained from $f(i)$ by the superposition of *random noise*:

$$\tilde{f}_k(i) = f(i) + \epsilon_{ik}, \tag{1}$$

where the values ϵ_{ik} are independent random variables with mean zero. This assumption is typical for *stochastic optimization problems*. For example, suppose that $f(i)$ denotes the expected costs occuring in a situation that depends on parameter i and on a random influence ω; moreover, suppose that to given i, an unbiased estimate $\tilde{f}_k(i)$ of the expected costs can be determined by means of a simulation experiment. Then we are within the problem context decribed above.

The aim of our article is to indicate conditions under which *convergence results* for the Simulated Annealing algorithm (see, e.g., Gelfand and Mitter [5], Hajek [6]) generalize to our stochastic optimization context. It has been argued that convergence results, stating that for certain cooling schedules the current solutions i converge in distribution to global optimizers, are not sufficient for a justification of Simulated Annealing (cf. [2]). Nevertheless, we think that it makes sense to gather exact information under which conditions such a convergence can be expected, and when it cannot be hoped for. It is easy to see that convergence to global optimizers is not possible, if the random noise ϵ_{ik} is, say, normally distributed with constant variance: under these circumstances, the noise will keep the probability that the current solution is suboptimal above a fixed level larger than zero. We will show that convergence *can* be obtained for suitable cooling schedules, if $Var(\epsilon_{ik})$ is decreased fast enough in the successive steps of evaluation. Notice that a reduction of the variance of the random noise ϵ_{ik} can always be achieved by *repeated, independent* observations at i, taking the average value of the observations as the estimate $\tilde{f}_k(i)$.

The problem treated here has already been formulated by Roenko [10]. His approach, however, makes it necessary to store all feasible solutions i encountered during the execution of the algorithm and to compare them with each newly generated solution j. This seems to be unrealistic for practical applications. In our approach, only information on the current solution i and a neighbor solution j is required.

2 Convergence in the Undisturbed Case

We start with a short review of a basic convergence result in the undisturbed case (cf. Aarts and Korst [1], ch. 3). First, we observe that from the viewpoint of probability theory, the algorithm SimAnn simulates an *inhomogenous Markov chain*, consisting of a sequence of *homogenous Markov chains* $\mathcal{M}_m$ $(m = 0, 1, \ldots)$ where $\mathcal{M}_m$ contains L_m state transitions. In the sequel, we assume that the Markov chain lengths L_m are kept constant, i.e., $L_m = L$. Furthermore, we assume that L is chosen as the minimum number of transitions to neighbors, required to reach an optimal solution i_{opt} from an arbitrary solution $j \in S$.

As before, c'_m denotes the value of the control parameter during the execution of the Markov chain $\mathcal{M}_m$. The sequence c'_m is supposed to satisfy the conditions $c'_{m+1} \leq c'_m$ $(m = 0, 1, \ldots)$ and $\lim_{m\to\infty} c'_m = 0$. While the index m refers to Markov chain $\mathcal{M}_m$, the index k will be used to refer to the single state transitions in the algorithm. The transition matrix $P(k) = (P_{ij}(k))_{i,j\in S}$ for the kth step is then given by

$$P_{ij}(k) = \begin{cases} |S_i|^{-1}\, I_{S_i}(j) \cdot \exp\left(-\frac{(f(j)-f(i))^+}{c_k}\right), & i \neq j, \\ 1 - \sum_{l\in S, l\neq i} P_{il}(c_k), & i = j, \end{cases} \tag{2}$$

where $|\cdot|$ denotes the cardinality of a set, I is the indicator function, and $c_k = c'_m$ for $mL < k \le (m+1)L$.

By the *underlying graph*, we understand the following graph: its nodes are the states $i \in S$ of the Markov chain, and its edges are defined by the neighborhood relation: (i, j) is edge if and only if $j \in S_i$. We always assume that the neighborhood relation is *symmetric*, i.e., $j \in S_i$ exactly if $i \in S_j$. Then, the underlying graph can be conceived as an undirected graph.

Without loss of generality, we may assume $S = \{1, \ldots, n\}$. Let the probability vector $q(k) = (q_1(k), \ldots, q_n(k))$ denote the distribution of the current solution $i = i(k)$ after the kth step of the algorithm $(k = 1, 2, \ldots)$. Obviously,

$$q(k) = q(0)P(1) \cdots P(k),$$

where $q(0) = (q_1(0), \ldots, q_n(0))$ is the distribution of the initial solution i_{start}. Aarts and Korst prove the following convergence theorem ([1], p. 46 – 50):

Theorem 2.1. Suppose that for an application of SimAnn to a combinatorial optimization problem, the following conditions hold:

(i) The underlying graph is connected.

(ii) The sequence c'_m satisfies

$$c'_m \ge \frac{(L+1)\Delta}{\log(m+2)} \qquad (m = 0, 1, \ldots),$$

where

$$\Delta := \max_{i \in S, j \in S_i} (f(j) - f(i))$$

is a Lipschitz constant.

Then, for an arbitrary initial distribution $q(0)$, the distribution of the current solution converges to the uniform distribution on the set S_{opt} of global optimizers:

$$\lim_{k\to\infty} q(k) = \lim_{k\to\infty} q(0)P(1) \cdots P(k) = q^*, \tag{3}$$

where $q^* = (q_1^*, \ldots, q_n^*)$ is given by

$$q_i^* = \frac{1}{|S_{opt}|} I_{S_{opt}}(i) \qquad (i \in S).$$

A recapitulation of the proof of this theorem in [1], ch. 3.4, immediately shows that the assertion also holds if the first k_0 transitions with transition matrices $P(1), \ldots, P(k_0)$ are skipped (k_0 fixed); the convergence result only depends on the "tail" of the infinite sequence of transitions. We have therefore even

$$\lim_{k\to\infty, k\ge k_0} qP(k_0+1) \cdots P(k) = q^* \tag{4}$$

for arbitrary fixed k_0 and arbitrary initial distribution q.

3 Convergence in the Disturbed Case

Now we assume that the evaluation step in SimAnn uses, instead of $\exp((f(i)-f(j))/c_m)$, the acceptance probabilities $\exp((\tilde{f}_k(i)-\tilde{f}_k(j))/c_m)$, where the disturbed values $\tilde{f}_k(\cdot)$ are given by (1). Moreover, let us assume first that the noise variables ϵ_{ik} are independent and $\mathcal{N}(0,\sigma_k^2)$–distributed. (This assumption will be relaxed in Section 4.) In order to prove our result, we need three lemmata:

Lemma 3.1. Let the conditions of Theorem 2.1 be satisfied, and let $\tilde{P}(k)$ be a sequence of transition matrices such that

$$\sum_{k=1}^{\infty} ||P(k)-\tilde{P}(k)|| < \infty \tag{5}$$

with

$$||A|| := \max_i \sum_j |a_{ij}|$$

for $A = (a_{ij})$. Then also

$$\lim_{k\to\infty} q(0)\tilde{P}(1)\cdots\tilde{P}(k) = q^*$$

holds for an arbitrary initial distribution $q(0)$.

Proof. It is easy to verify that $||\cdot||$ is a norm on the space of real $[n\times n]$–matrices. Moreover,

$$||AP|| \le ||A|| \qquad \text{and} \qquad ||PA|| \le ||A|| \tag{6}$$

holds for each *stochastic* $[n\times n]$–matrix P:

$$||AP|| = \max_i \sum_j \left|\sum_l a_{il}p_{lj}\right| \le \max_i \sum_l |a_{il}| \sum_j p_{lj} = ||A||,$$

and

$$||PA|| = \max_i \sum_j \left|\sum_l p_{il}a_{lj}\right| \le \max_i \sum_l p_{il} \sum_j |a_{lj}| \le \max_i \sum_l p_{il}\cdot ||A|| = ||A||.$$

We show by induction w.r.t. k that for $r \le k$,

$$||\, P(r)\cdots P(k) - \tilde{P}(r)\cdots\tilde{P}(k)\,|| \le \sum_{l=r}^{k} ||P(l)-\tilde{P}(l)||. \tag{7}$$

The case $k = r$ is clear. Let us verify that if (7) holds for some $k \ge r$, then it also holds for $k+1$:

$$||\, P(r)\cdots P(k+1) - \tilde{P}(r)\cdots\tilde{P}(k+1)\,||$$

$$\le ||\, P(r)\cdots P(k)\,[P(k+1)-\tilde{P}(k+1)]\,|| + ||\,[P(r)\cdots P(k) - \tilde{P}(r)\cdots\tilde{P}(k)]\,\tilde{P}(k+1)\,||$$

$$\leq ||P(k+1) - \tilde{P}(k+1)|| + \sum_{l=r}^{k} ||P(l) - \tilde{P}(l)|| = \sum_{l=r}^{k+1} ||P(l) - \tilde{P}(l)||,$$

where we have used (6) and the induction assumption. This proves (7) for every k.

Now, set

$$A_{r,k} := P(r)\cdots P(k) - \tilde{P}(r)\cdots\tilde{P}(k) \quad (k \geq r).$$

Let $\epsilon > 0$ arbitrary. Equation (7) yields

$$||A_{r,k}|| \leq \sum_{l=r}^{k} ||P(l) - \tilde{P}(l)||.$$

Hence it follows from condition (5) that there is a number r such that

$$||A_{r,k}|| < \epsilon \qquad \text{for all} \quad k \geq r.$$

Since

$$||xA||_1 \leq \sum_i x_i \sum_j |a_{ij}| \leq \sum_i x_i ||A|| = ||A||$$

for each probability vector $x \in R^n$ and each $[n \times n]$–matrix A, we obtain with

$$\tilde{q}(k) := q(0)\tilde{P}(1)\cdots\tilde{P}(k)$$

that

$$\begin{aligned} &|| \, q(0)\tilde{P}(1)\cdots\tilde{P}(k) - q^* \, ||_1 \\ \leq\, & || \, \tilde{q}(r-1) \, [\tilde{P}(r)\cdots\tilde{P}(k) - P(r)\cdots P(k)] \, ||_1 + || \, \tilde{q}(r-1)P(r)\cdots P(k) - q^* \, ||_1 \\ \leq\, & ||A_{r,k}|| + || \, \tilde{q}(r-1)P(r)\cdots P(k) - q^* \, ||_1 < \epsilon + || \, \tilde{q}(r-1)P(r)\cdots P(k) - q^* \, ||_1 \end{aligned}$$

for $k \geq r$. Because of (4),

$$|| \, \tilde{q}(r-1)P(r)\cdots P(k) - q^* \, ||_1 \to 0 \qquad (k \to \infty).$$

In total, this yields

$$|| \, q(0)\tilde{P}(1)\cdots\tilde{P}(k) - q^* \, ||_1 \to 0 \qquad (k \to \infty).$$

□

In the following lemma, we consider the kth state transition of the algorithm. Since the index k is kept fixed, it is be omitted in the notation for the presence.

Lemma 3.2. Let the temperature have the value c, and let P_{ij} resp. $\tilde{P}_{ij}$ denote the transition probability from state i to state j in the undisturbed Markov chain, resp. in the Markov chain disturbed according to (1), where the random noise is $\mathcal{N}(0,\sigma^2)$–distributed. Then for $i \in S$ and $j \in S_i$,

$$\tilde{P}_{ij} - P_{ij} = \frac{1}{n}\left\{\Phi\left(-\frac{\mu}{s}\right) + \exp\left(\frac{s^2}{2c^2} - \frac{\mu}{c}\right)\,\left[1 - \Phi\left(\frac{s}{c} - \frac{\mu}{s}\right)\right] - \exp\left(-\frac{\mu^+}{c}\right)\right\}, \qquad (8)$$

where

$$\mu = f(j) - f(i),$$
$$s = \sqrt{2}\sigma,$$

and Φ is the distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

Proof. By (2), $P_{ij} = (1/n)\exp(-\mu^+/c)$ for $j \in S_i$. Let us compute $\tilde{P}_{ij}$. The disturbed function values in i and j are given by

$$\tilde{f}(i) = f(i) + \epsilon_i, \qquad \tilde{f}(j) = f(j) + \epsilon_j,$$

with independent noise variables ϵ_i, ϵ_j. Hence the acceptance probability for given ϵ_i and ϵ_j is

$$\tilde{P}_{ij}^{(\epsilon_i,\epsilon_j)} = \frac{1}{n}\exp\left(-\frac{(\mu+\epsilon_j-\epsilon_i)^+}{c}\right) = \frac{1}{n}\exp\left(-\frac{(\mu+sz)^+}{c}\right),$$

where $z = (\epsilon_j - \epsilon_i)/s \sim \mathcal{N}(0,1)$. We have to take average of $\tilde{P}_{ij}^{(\epsilon_i,\epsilon_j)}$ over all (ϵ_i, ϵ_j), i.e., over all values of the variable z distributed according to $\mathcal{N}(0,1)$. Let $\varphi := \Phi'$ be the density of the $\mathcal{N}(0,1)$–distribution. Then one obtains by a short calculation:

$$n\tilde{P}_{ij} = \int_{-\infty}^{\infty} \exp\left(-\frac{(\mu+sz)^+}{c}\right)\varphi(z)\,dz$$

$$= \int_{-\infty}^{-\mu/s}\varphi(z)\,dz + \int_{-\mu/s}^{\infty}\exp\left(-\frac{\mu+sz}{c}\right)\varphi(z)\,dz$$
$$= \Phi\left(-\frac{\mu}{s}\right) + \exp\left(\frac{s^2}{2c^2} - \frac{\mu}{c}\right)\left[1 - \Phi\left(\frac{s}{c} - \frac{\mu}{s}\right)\right],$$

which yields the assertion. □

Now, we consider the index k as variable. For each k, we have a specific temperature c_k and noise ϵ_{ik} with a specific standard deviation σ_k.

Lemma 3.3. If, in the inhomogenous Markov chain, $s_k = \sqrt{2}\sigma_k$ is of order $O(k^{-\gamma})$ with $\gamma > 1$, and c_k is of order $\Omega((\log k)^{-1})$, then

$$\sum_{k=1}^{\infty} |P_{ij}(k) - \tilde{P}_{ij}(k)| < \infty$$

for all $i, j \in S$.

Proof. For $j \notin S_i$, the assertion is trivial. Assume $j \in S_i$. We consider the r.h.s. of (8) for different transitions k.

Case (i): $\mu > 0$.
The r.h.s. of (8) for a given k can be decomposed in the form $(1/n)A_k + (1/n)B_k - (1/n)C_k$, where

$$A_k = \Phi(-\frac{\mu}{s_k}),$$

$$B_k = \exp\left(\frac{s_k^2}{2c_k^2} - \frac{\mu}{c_k}\right) - \exp\left(-\frac{\mu^+}{c_k}\right),$$

$$C_k = \exp\left(\frac{s_k^2}{2c_k^2} - \frac{\mu}{c_k}\right) \cdot \Phi\left(\frac{s_k}{c_k} - \frac{\mu}{s_k}\right).$$

We show that $\sum_k |A_k| < \infty$, $\sum_k |B_k| < \infty$ and $\sum_k |C_k| < \infty$. First, observe that

$$0 \le \Phi\left(-\frac{\mu}{s_k}\right) \le \Phi(-\text{ const } \cdot k) \to 0 \quad (k \to \infty),$$

Since

$$\Phi(x) \le \frac{1}{|x|\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{9}$$

for $x < \infty$, the convergence is of exponential speed. This proves $\sum_k |A_k| < \infty$. Secondly, we have

$$0 \le B_k = \exp\left(-\frac{\mu}{c_k}\right)\left(\exp\left(\frac{s_k^2}{2c_k^2}\right) - 1\right) \le \exp\left(\frac{s_k^2}{2c_k^2}\right) - 1.$$

For $0 \le x \le 1$, the estimation $e^x - 1 \le (e-1)x$ holds, and therefore

$$\exp\left(\frac{s_k^2}{2c_k^2}\right) - 1 \le (e-1) \cdot \frac{s_k^2}{2c_k^2} \le \text{ const } \cdot \left(\frac{\log k}{k}\right)^2 \le \text{ const } \cdot k^{-3/2}.$$

Hence $\sum_k |B_k| < \infty$. Finally, $C_k \ge 0$, and for sufficiently large k, we obtain

$$\exp\left(\frac{s_k^2}{2c_k^2} - \frac{\mu}{c_k}\right) \le \exp\left(1 - \frac{\mu}{c_k}\right) \le e$$

and

$$\Phi\left(\frac{s_k}{c_k} - \frac{\mu}{s_k}\right) \le \Phi\left(\frac{s_k}{c_k} - \text{ const } \cdot k\right) \le \Phi(1 - \text{ const } \cdot k) \to 0 \quad (k \to \infty)$$

with exponential speed. Hence also $\sum_k |C_k| < \infty$. This proves the assertion for case (i).

Case (ii): $\mu < 0$.

In this case, we decompose the r.h.s. of (8) for a given k in the form $(1/n)D_k + (1/n)E_k$, where

$$D_k = \Phi\left(-\frac{\mu}{s_k}\right) - \exp\left(-\frac{\mu^+}{c_k}\right),$$

$$E_k = \exp\left(\frac{s_k^2}{2c_k^2} - \frac{\mu}{c_k}\right) \cdot \left[1 - \Phi\left(\frac{s_k}{c_k} - \frac{\mu}{s_k}\right)\right].$$

We find (using $\mu < 0$):

$$|D_k| = \left|\Phi\left(-\frac{\mu}{s_k}\right) - 1\right| = \Phi\left(\frac{\mu}{s_k}\right) \le \Phi(-\,\text{const}\,\cdot k) \to 0 \quad (k \to \infty)$$

with exponential speed, so $\sum_k |D_k| < \infty$ holds. Now, for sufficently large k,

$$0 \le E_k = \exp\left(\frac{s_k^2}{2c_k^2}\right) \cdot \exp\left(-\frac{\mu}{c_k}\right)\,\Phi\left(\frac{\mu}{s_k} - \frac{s_k}{c_k}\right) \le \text{const}\,\cdot \exp\left(-\frac{\mu}{c_k}\right)\,\Phi\left(\frac{\mu}{s_k}\right).$$

By (9), we have $\Phi(-x) \le \varphi(x)/x$ $(x > 0)$. Therefore, with constants K and C and for sufficiently large k,

$$\exp\left(-\frac{\mu}{c_k}\right)\,\Phi\left(\frac{\mu}{s_k}\right) \le \frac{s_k}{(-\mu)\sqrt{2\pi}}\,\exp\left(\frac{(-\mu)}{c_k} - \frac{\mu^2}{2s_k^2}\right)$$

$$\le \text{const}\,\cdot \exp\left(\frac{(-\mu)\log k}{K} - \frac{\mu^2 k^2}{2C^2}\right) \le \text{const}\,\cdot \exp\left(-\frac{\mu^2}{4C^2}k^2\right) \to 0 \quad (k \to \infty)$$

with exponential speed. This proves $\sum_k |E_k| < \infty$.

Case (iii): $\mu = 0$.
In this case, the r.h.s. of (8) for a given k yields

$$\frac{1}{n}\left\{\exp\left(\frac{s_k^2}{2c_k^2}\right)\left(1 - \Phi\left(\frac{s_k}{c_k}\right)\right) - \frac{1}{2}\right\} = \frac{1}{n}\left\{\exp\left(\frac{x^2}{2}\right)\Phi(-x) - \frac{1}{2}\right\} =: G(x)$$

with $x := s_k/c_k \to 0$ as $k \to \infty$. Since

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}x + O(x^2) \quad (x \to 0),$$

expansion of $G(x)$ at $x = 0$ leads to

$$G(x) = \frac{1}{n}\left(-\frac{x}{\sqrt{2\pi}} + O(x^2)\right) \quad (x \to 0),$$

and therefore $|G(x)| \le |x|$ for sufficiently small x. Now

$$x = s_k/c_k = O(k^{-\gamma} \cdot \log k) \quad (k \to \infty)$$

with $\gamma > 1$, so $\sum_k s_k/c_k$ converges, and hence also $\sum_k |G(s_k/c_k)|$, which completes the proof. □

Now we are in the position to state our main result:

Theorem 3.1. If the standard deviations σ_k of the independent, normally distributed noise variables ϵ_{ik} decrease with an order $O(k^{-\gamma})$, where $\gamma > 1$, then the assertion of Theorem 2.1 holds also for disturbed function observations $\tilde{f}_k(i) = f(i) + \epsilon_{ik}$.

Proof. The result follows immediately from Lemma 3.1 and Lemma 3.3, since

$$\sum_{k=1}^{N} ||P(k) - \tilde{P}(k)|| \leq \sum_{i,j \in S} \sum_{k=1}^{N} |P_{ij}(k) - \tilde{P}_{ij}(k)|.$$

□

Remark 1. If neighbor solutions i, j always have different objective function values, i.e., if $f(i) \neq f(j)$ for all $i \in S$, $j \in S_i$, then the condition $\sigma_k = O(k^{-\gamma})$ $(\gamma > 1)$ in Theorem 3.1 can be replaced by the weaker condition $\sigma_k = O(k^{-1})$. The reader can verify this by a recapitulation of the proof, observing that case (iii) in the proof of Lemma 3.3 is now impossible.

Remark 2. It is easy to see that, just as in the undisturbed case, the assertion of Theorem 3.1 still holds if a fixed number k_0 of initial transitions is skipped (cf. (4)).

4 Generalization to Other Noise Distributions

The question arises whether the assertion of Theorem 3.1 depends on the special assumption of *normally* distributed noise. In this section, we will show that this is not the case. For that purpose, let us start with a definition.

Following Birnbaum [3], we say that a distribution μ_1 is *more peaked around zero* than a distribution μ_2, if

$$\mu_1(\,]-t,t[\,) \geq \mu_2(\,]-t,t[\,) \quad \text{for all}\, t > 0.$$

This holds if and only if there are random variables $X_1 \sim \mu_1$ and $X_2 \sim \mu_2$ such that $|X_1| \leq |X_2|$. It is easy to show that if μ_1 is more peaked around zero than μ_2, and both μ_1 and μ_2 are symmetric around zero, then

$$\int_0^R \psi(z)\, d\mu_1(z) \leq \int_0^R \psi(z)\, d\mu_2(z)$$

for each $R > 0$ and for each function ψ that is nondecreasing in the interval $[0, R]$.

The following theorem allows a generalization of Theorem 3.1 from normal noise distributions to rectangular distributions, triangular distributions, Maxwell distributions etc.:

Theorem 4.1. Let $\bar{\epsilon}_{ik}$ be independent noise variables with $\bar{\epsilon}_{jk} - \bar{\epsilon}_{ik}$ distributed according to μ_k $(i, j \in S, i \neq j)$, and assume that for the distributions μ_k $(k \geq 1)$ the following conditions hold:

(i) μ_k is symmetric around zero,

(ii) μ_k is more peaked around zero than $\mathcal{N}(0, s_k^2)$, where $s_k = O(k^{-\gamma})$ with a constant $\gamma > 1$.

Then also in an environment with noise $\bar{\epsilon}_{ik}$, the assertion of Theorem 2.1 holds.

Proof. Let $\bar{P}_{ij}(k)$ denote the transition probabilities at step k under noise $\bar{\epsilon}_{ik}$ and $\bar{\epsilon}_{jk}$. It suffices to show that

$$\sum_{k=1}^{\infty} |P_{ij}(k) - \bar{P}_{ij}(k)| < \infty \tag{10}$$

for all $i, j \in S$ (cf. Lemma 3.3). Let k be fixed. Analogously as in the proof of Lemma 3.2, we set

$$\bar{P}_{ij}^{(\bar{\epsilon}_i,\bar{\epsilon}_j)} = \frac{1}{n} \exp\left(-\frac{(\mu + \bar{\epsilon_j} - \bar{\epsilon_i})^+}{c}\right) = \frac{1}{n} \exp\left(-\frac{(\mu + s\bar{z})^+}{c}\right),$$

where $s = s_k$ and $\bar{z} = (\bar{\epsilon_j} - \bar{\epsilon_i})/s_k$. Let $\mu_k^{(0)}$ denote the distribution of $\bar{z}$. Obviously, $\mu_k^{(0)}$ is the linear compression of the distribution μ_k by the factor s_k. Then, because of condition (ii), $\mu_k^{(0)}$ is more peaked around zero than $\mathcal{N}(0,1)$. The transition probability $\bar{P}_{ij}(k)$ is obtained from the probabilities $\bar{P}_{ij}^{(\bar{\epsilon}_i,\bar{\epsilon}_j)}$ by taking average over $(\bar{\epsilon}_i, \bar{\epsilon_j})$, resp. over $\bar{z} \sim \mu_k^{(0)}$. Using the symmetry of $\mu_k^{(0)}$, one finds after short calculation:

$$n\bar{P}_{ij} - nP_{ij} = \int_0^{\infty} \psi_k(\bar{z})\, d\mu_k^{(0)}(\bar{z})$$

with

$$\psi_k(\bar{z}) = \exp\left(-\frac{(\mu + s\bar{z})^+}{c}\right) + \exp\left(-\frac{(\mu - s\bar{z})^+}{c}\right) - 2\exp\left(-\frac{\mu^+}{c}\right).$$

Let us distinguish the cases $\mu > 0$ and $\mu \leq 0$. In the first case, $\psi_k(\bar{z})$ is nonnegative and nondecreasing in the interval $[0,\, \mu/s]$. For an arbitrary sequence R_k of nonnegative numbers, the estimation

$$\left|\int_0^{\infty} \psi_k(\bar{z})\, d\mu_k^{(0)}\right| \leq \left|\int_0^{R_k} \psi_k(\bar{z})\, d\mu_k^{(0)}\right| + 4\,\Phi(-R_k)$$

follows from the fact that $\mu_k^{(0)}$ is more peaked around zero than $\mathcal{N}(0,1)$. We choose $R_k := \mu/s_k$. Then $\sum_k \Phi(-R_k) < \infty$ (cf. the proof of Lemma 3.3, case (i)). Furthermore, one obtains

$$0 \leq \left|\int_0^{R_k} \psi_k(\bar{z})\, d\mu_k^{(0)}\right| = \int_0^{R_k} \psi_k(\bar{z})\, d\mu_k^{(0)} \leq \int_0^{R_k} \psi_k(\bar{z})\, \varphi(\bar{z}) d\bar{z},$$

again because $\mu_k^{(0)}$ is more peaked around zero than $\mathcal{N}(0,1)$, and by the monotonicity of $\psi_k(\bar{z})$ in $[0, R_k]$. Now,

$$\int_0^{R_k} \psi_k(\bar{z})\, \varphi(\bar{z}) d\bar{z} \leq \left|\int_0^{\infty} \psi_k(\bar{z})\, \varphi(\bar{z}) d\bar{z}\right| + 4\,\Phi(-R_k) = |\tilde{P}_{ij}(k) - P_{ij}(k)| + 4\,\Phi(-R_k),$$

where $\tilde{P}_{ij}(k)$ denotes the acceptance probability for normally distributed noise with standard deviation s_k. Because of Lemma 3.3, $\sum_k |\tilde{P}_{ij}(k) - P_{ij}(k)|$ converges.

If, on the other hand, $\mu \leq 0$, then one obtains $\psi_k(\bar{z}) \leq 0$ and $-\psi_k(\bar{z})$ nondecreasing on $[0, \infty[$, and hence

$$0 \leq nP_{ij} - n\bar{P}_{ij} = -\int_0^\infty \psi_k(\bar{z}) d\mu_k^{(0)}(\bar{z}) \leq -\int_0^\infty \psi_k(\bar{z})\varphi(\bar{z})d\bar{z} = nP_{ij}(k) - n\tilde{P}_{ij}(k),$$

which leads again to a convergent series. In total, this yields (10). □

5 Conclusion

Our results show that if Simulated Annealing is applied in the context of stochastic optimization, it is not efficient to spend at each evaluation step the same effort for the estimation of the (disturbed) cost function: In the first transition steps of the algorithm, we can do with relatively vague estimations; the more the temperature is decreased, the more accurate cost function evaluations are required. For the cooling schedule of Theorem 2.1, which guarantees convergence of the current solutions to global optimizers, we have found that each reduction of the standard error of the order $O(k^{-\gamma})$ ($\gamma > 1$) is sufficient for maintaining the desired convergence property.

Let us mention that a gradual increment of the precision in the cost function estimations, as prescribed by the Theorems 3.1 and 4.1, is also a common feature of other stochastic optimization techniques, such as the well–known Kiefer–Wolfowitz procedure.

Our results give first hints for Simulated Annealing in a noisy environment, but a lot of work remains to be done. The next step of research should perhaps include a generalization of *finite–time behavior* results (see, e.g., Catoni [4]) to the noisy context. Also, experimental results on the application of Simulated Annealing to different "hard" stochastic optimization problems would be of great value.

References

[1] Aarts, E., Korst, J., *Simulated Annealing and the Boltzmann Machine*, Wiley (1990).

[2] Bertsimas, D., Tsitsiklis, J., Simulated Annealing, *Statistical Science* 8, pp. 10 - 15 (1993).

[3] Birnbaum, Z. W., On random variables with comparable peakedness, *Ann. Math. Statist.* 19, pp. 76 - 81 (1948).

[4] Catoni, O., Rough large deviation estimates for Simulated Annealing: application to exponential schedules, *Annals of Probability* 20, pp. 1109 - 1146 (1992).

[5] Gelfand, S. B., Mitter, S. K., Analysis of Simulated Annealing for optimization, *Proc. 24th IEEE Conf. on Decision and Control*, Ft. Lauderdale, pp. 779 - 786 (1985).

[6] Hajek, B., Cooling schedules for optimal annealing, *Math. of Operations Research* 13, pp. 311 - 329 (1988).

[7] Horst, H., Pardalos, P.M. (Editors), *Handbook of Global Optimization*, Kluwer Academic Publishers (1995).

[8] Kirkpatrick, S., Gelatt, Jr., Vecchi, M. P., Optimization by Simulated Annealing, *Science* 220, pp. 671 - 680 (1983).

[9] Laarhoven, P. J. M. van, Aarts, E. H. L., *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht (1987).

[10] Roenko, N., Simulated Annealing under uncertainty, Technical Report, Inst. f. Operations Research, Univ. Zrich (1990).