Simulated Annealing with Time dependent Energy Function via Sobolev inequalities

Matthias Löwe
Universität Bielefeld
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld

Abstract

We analyze the Simulated Annealing Algorithm with an energy function U_t that depends on time. Assuming some regularity conditions on U_t (especially that U_t does not change too quickly in time), and choosing a logarithmic cooling schedule for the algorithm, we derive bounds on the Radon-Nikodym density of the distribution of the annealing algorithm at time t with respect to the invariant measure π_t at time t. Moreover we estimate the entrance time of the algorithm into typical subsets V of the state space in terms of $\pi_t(V^c)$.

Keywords: Simulated Annealing, Sobolev inequalities, Spectral gap, Markov processes

Short title: Simulated Annealing with Time dependent Energy

1 Introduction

Let X be a finite set. The well known Simulated Annealing (SA) algorithm is an inhomogeneous Markov process Y_t on X with the aim to minimize a given function $U : X \to \mathbb{R}$. The idea behind SA is to think of U as an energy function and to choose the Markov process in such a way that the corresponding time-homogeneous Markov process at time t has at its invariant measure π_t, the Gibbs distribution with energy U at temperature $1/\beta_t$. So, if β_t diverges to infinity in the limit for large times the invariant measure becomes concentrated on the set of minima of U. On the other hand one has to "freeze" the process slowly enough to guarantee convergence of its distribution at time t towards equilibrium.

Since the introduction of SA by Kirkpatrick et al. !983) it has found a wide range of applications (see Geman and Geman (1984), Vidal (1993)), and the problems mentioned above, i.e. the optimal choice of β_t and the asymptotic behavior of SA have been studied by several authors. The problem of an asymptotically optimal "cooling schedule" (i.e. the dependence of β_t from t) has been settled by Hajek (1988), who showed the optimality of a logarithmic freezing schedule $\beta_t = \frac{1}{m}\log(A + Bt)$, where A and B are positive constants and m is, roughly speaking, the largest hill to be climbed to reach the global minimum.

The rate of convergence has been investigated by Gidas (1985) and Mitra et al . (1985) on the basis of Dobrushin coefficients for Markov chains. Catoni (1990) has obtained bounds of the type $P(Y_t \geq E) \leq O(t^{-\alpha})$ by using Freidlin-Wentzell techniques (see Freidlin and Wentzell (1984)) on perturbed dynamical systems.

Holley and Stroock (1988) were able to give bounds for $P_x(Y_t = y)/\pi_t(y)$ for reversible SA processes based on Sobolev inequalities. Additionally they obtained sharp estimates for the spectral gap of the operators associated to the SA process. Based on refined spectral gap estimates Götze (1992) proved effective bounds for the waiting time of escaping from an unlikely set (i.e. a set of high energy). These techniques were generalized by Deuschel and Mazza (1994) to Markov processes on finite sets similar to SA.

In this paper we are going to analyze that variant of the SA algorithm, where U depends on t, i.e. $U = U_t$. These processes have first been studied by Frigerio and Grillo (1993). Not only they are a natural generalization of SA, but also they can be regarded as a discrete version of quantum diffusions (see Frigerio and Grillo (1993), Section 4).
Moreover SA can be thought of as the simplest form of an genetic algorithm: A population – consisting of one individual only – creates a child, and if this child is better than its ancestor in terms of the "fitness function" U it is accepted and the ancestor dies, otherwise it is accepted with a certain probability (the reader is referred to the book of Holland (1975) for an introduction and a survey on these algorithms and to the paper by the author (1996) for an application of the L^2-techniques to a wide class of genetic algorithms). In such a model the time dependence of U comes up naturally, e.g., if one thinks of the fitness of an individual being given by its environment, that usually changes in time.

A more concrete situation where our results apply comes from the theory of neural networks. The Hopfield model of neural networks (introduced by Hopfield (1982)). connects the ideas of neural computing with those of statistical mechanics. The basic idea behind this model is to choose the information to be stored – which usually is referred to as patterns – as the local minima of an appropriate Hamiltonian (this is going to be made made more formal in the next section) on the set $\{1, \ldots, N\}$ which is thought of as the set of neurons. According to the other rules of neural networks, this only is possible up to a certain accuracy depending on the number of patterns to be stored. If this number is less than $\frac{N}{6 \log N}$ the patterns are the exact local minima of the Hamiltonian and hence noised patterns can be retrieved by a gradient dynamics on the energy landscape given by the Hamiltonian. But if the number of patterns grows like αN for an $\alpha \leq 0.085$ the energy landscape becomes much more difficult and it only can be shown that the patterns are surrounded by an energy barrier of height proportional to N (see Talagrand (1995)). In these situations a stochastic retrieval dynamics like SA has been proposed (see e.g. Aarts and Kost (1989)). Now in some situations one would like the Hopfield model not only to recognize single patterns but deterministic sequences of patterns and to show each element of the sequence for a certain period of time and then jump to the next element (see e.g. van Hemmen and Kühn (1991)). This leads to a Hamiltonian (or energy function) that depends on time. The corresponding retrieval dynamics is SA with an energy function that varies in time.

Frigerio and Grillo (1993) proved that for SA with time dependent energy function the distribution of the process and its invariant measure become indistinguishable in the weak topology for large times t, if the following catalogue of conditions is fulfilled: the SA process is driven by a bistochastic matrix, $U_t(x)$ is uniformly bounded in x and t by a nonnegative number M and has a derivative with respect to t bounded by $\frac{C}{t^\alpha}$ for some $C, \alpha > 0$, the generator L_t associated to the process at time t admits a bound on its spectral gap of the form $\gamma e^{-\beta_t m}$ for some $\gamma, m > 0$, and a logarithmic cooling schedule $\beta_t = \frac{1}{h} \log(1+t)$ with $h > m$ and $m/h < \alpha$ is chosen. These conditions, that are especially satisfied when U_t admits an uniform limit U_∞ look rather natural from a point of view of Sobolev inequalities and spectral gap estimates. On one hand one has chosen a cooling schedule that also is close to optimal, when U is constant in t. On the other hand in some sense m and M are the only values which enter in these techniques for time independent energy function. So if the variation of U_t is moderate one should expect spectral gap estimates to work.

The aim of this paper is to show that under the conditions formulated by Frigerio and Grillo (1993) this is indeed the case. By Sobolev inequalities we obtain bounds for the density of the distribution of the process with respect to the invariant distribution at time t and are able to analyze the speed of convergence towards equilibrium. It turns out the conditions formulated above cannot essentially be weakened. Our main tools for this approach are those of Holley and Stroock (1988) and Götze (1992). In our results we benefit from the major advantage of the Sobolev inequality approach over the other approaches in that we get explicit bounds for the waiting time of the algorithm to enter sets of high probability (which especially means that we get explicit constants in these bounds). This brings the results of Frigerio and Grillo (1993) into an applicable form.

We organize the rest of the paper in the following way. Section 2 contains our basic definitions and a formal description of the model. In Section 3 we describe the behavior of the Radon-Nikodym derivative of the distribution of the process at time t with respect to the Gibbs-distribution at that time with the help of Sobolev inequalities and check, under which conditions these are fulfilled. Moreover we use spectral gap estimates to bound the probability of large entrance times into typical sets.

2 The Model

Let us now give a formal description of the algorithm we have mind. To this end for each $x \in X$ let $q(x, \cdot)$ be a probability distribution on X. Assume that the random walk associated with $q(x, y)$ is irreducible (i.e. for all pairs $x, y \in X$ there is a number $n := n(x, y)$ such that $q^n(x, y) > 0$ where we have put $q^n(x, y) = \sum_{z \in X} q(x, z) q^{n-1}(z, y)$). Moreover for simplicity we assume that $q(x, y)$ is time-reversible with respect to a probability measure π on X, i.e.

$$\alpha(x, y) := \pi(x) q(x, y) = \pi(y) q(y, x) = \alpha(y, x) \tag{1}$$

for all $x, y \in X$. Note that π is q-stationary and therefore irreducibility of q implies that π charges every point in X. This assumption of reversibility often makes life easier in convergence theorems for stochastic processes. On the other hand a comparison of the spectral gap estimates for time homogeneous, reversible Markov chains by Diaconis and Stroock (1991) with those for homogeneous, irreversible chains by Fill (1991) shows that there is some hope to drop the reversibility condition in future work.

Let $(U_t)_{t \in \mathbb{R}^+}$ be a family of functions $U_t : X \to \mathbb{R}$. Moreover let us assume that $t \mapsto \beta_t$ is a monotonely increasing function in t, which is strictly positive (β_t will play the role of the inverse temperature at time t). Corresponding to q, U_t, and β_t we may define a transition probability $q_t(x, y)$ by

$$q_t(x, y) = \begin{cases} \exp(-\beta_t (U_t(y) - U_t(x))^+) q(x, y) & \text{if } x \neq y \\ 1 - \sum_{z \neq y} q_t(x, z) & \text{if } x = y \end{cases} \tag{2}$$

where we have put $(x)^+ := \max(x, 0)$.
It is easy to check that q_t is reversible with respect to the Gibbs measure π_t (with reference measure π) with energy U_t at inverse temperature β_t defined by

$$\pi_t(y) := \frac{e^{-\beta_t U_t(y)}}{Z_t} \pi(y) \tag{3}$$

where $Z_t := \sum_y e^{-\beta_t U_t(y)} \pi(y)$ is the partition function. Note also that q_t again is irreducible and reversible and that π_t is its stationary distribution, or, equivalently, that the operator Q_t given by

$$[Q_t \Phi](x) := \sum_{y \in X} \Phi(y) q_t(x, y), \quad x \in X \tag{4}$$

is a self-adjoint contraction on $L^2(\pi_t)$.

If we now define the operator $L_t := Q_t - I$ on $L^2(\pi_t)$, i.e.

$$[L_t\Phi](x) := \sum_{y\in X}(\Phi(y) - \Phi(x))q_t(x,y), \quad x \in X \tag{5}$$

we are able to describe the continuous time SA algorithm. Namely let its transition probabilities $P_{s,t}(x,y)$, i.e. the probabilities to arrive in y at time t when starting in x at time s for $s \leq t$ be given by the Fokker-Planck equation

$$\frac{\partial}{\partial t}[P_{s,t}\Phi](x) = [P_{s,t}[L_t\Phi]](x) \quad \text{for } s < t \tag{6}$$

and

$$[P_{s,s}\Phi](x) := \Phi(x)$$

for $\Phi : X \to \mathbb{R}$. Here

$$[P_{s,t}\Phi](x) := \sum_{y\in X}\Phi(y)P_{s,t}(x,y), \quad x \in X.$$

Let us reserve the symbol Y_t for the process described by 6. The intuitive meaning of 6 is that starting in x at time s the process waits for exponential time with mean one and then chooses a neighbor y according to $q(x,y)$. If then $U_t(y) \leq U_t(x)$ (where t is the instantaneous time after having waited) the algorithm moves to y. Otherwise it moves to y with probability $\exp(-\beta_t(U_t(y) - U_t(x)))$ and stays in x with probability $1 - \exp(-\beta_t(U_t(y) - U_t(x)))$. Obviously this is the continuous time analogue of the ordinary discrete time SA process where we have chosen U to depend on the time. The advantage of using continuous time instead of discrete time becomes especially apparent when studying the time evolution of f_t, the Radon-Nikodym derivative of the distribution of Y_t with respect to π_t, which then can be described by means of differential equations instead of difference equations that are harder to solve.

Let us remark that together with Q_t also L_t is a self-adjoint operator on $L^2(\pi_t)$ with largest eigenvalue 0. Due to the irreducibility of q_t the eigenspace of the eigenvalue 0 is one-dimensional and the constants are the only eigenfunctions. The key observation on which most of the following techniques are based is the relation of the second smallest eigenvalue λ_t of $-L_t$ and the Dirichlet form associated with $-L_t$. This relation is given by

$$\lambda_t = \inf\{\frac{\mathcal{E}_t(\Phi,\Phi)}{Var_t(\Phi)} : \Phi : X \to \mathbb{R} \quad \gneqq \text{non constant}\}. \tag{7}$$

Here

$$\mathcal{E}_t(\Phi,\Phi) := - < \Phi, [L_t\Phi] >_t = \sum_{x,y\in X}\frac{1}{2}(\Phi(x) - \Phi(y))^2\alpha_t(x,y) \tag{8}$$

is the discrete Dirichlet form associated to L_t, $< \cdot, \cdot >_t$ denotes the scalar product in $L^2(\pi_t)$, $Var_t(\Phi)$ is the variance of Φ with respect to π_t and α_t is given by

$$\alpha_t(x, y) := \pi_t(x) q_t(x, y) = \pi_t(y) q_t(y, x) = \alpha_t(y, x). \tag{9}$$

Note that $\mathcal{E}_t$ is related to the Dirichlet form
$\mathcal{E}_0(\Phi, \Phi) := \sum_{x,y \in X} \frac{1}{2}(\Phi(x) - \Phi(y))^2 \alpha(x, y)$ via the relation

$$\mathcal{E}_t(\Phi, \Phi) = \sum_{x,y \in X} (\Phi(x) - \Phi(y))^2 \alpha(x, y) \frac{\exp(-\beta_t \max(U_t(x), U_t(y)))}{2Z_t}. \tag{10}$$

(7) is derived by the well known variational characterization of eigenvalues (see e.g. Horner and Johnson (1985)).

For the rest of the paper let us assume that the $(U_t)_t$ fulfill the following conditions:
First we may without loss of generality assume that $\min_x U_t(x) = 0$ for all $t \in \mathbb{R}^+$. Moreover let us choose such sequences of energy functions that there exist universal constants $\gamma \geq 0$ and $\alpha, C, m, M > 0$ such that

$$0 \leq U_t(x) \leq M \qquad \text{for all } x \in X, \text{ for all } t \geq 0\ , \tag{11}$$

$$|\frac{\partial}{\partial t} U_t(x)| \leq \frac{C}{(1+t)^\alpha} \qquad \text{for all } x \in X, \tag{12}$$

and

$$\lambda_t \geq \gamma e^{-\beta_t m} \qquad \text{for all } t \geq 0. \tag{13}$$

We will conclude this section with the observation that assumption (13) is automatically fulfilled if we choose m similar to Hajek (1988) as the maximal hill to be climbed at any time to reach to global minimum. More formally for fixed t and any two points $x, y \in X$ let us call a sequence $(x_i)_{1 \leq i \leq k}$ a path from x to y, if $x_1 = x$, $x_k = y$ and $q_t(x_{i-1}, x_i) > 0$ for all $i = 2, \ldots, k$. Define

$$\begin{aligned} m_t(x) := \min\{E \geq 0 : \quad & \exists x \neq y \in X \text{ with } U_t(y) \leq U_t(x) \\ & \text{and a path } (x_i)_{1 \leq i \leq k} \text{from } x \text{ to } y \text{ with} \\ & U_t(x_i) \leq U_t(x) + E \text{ for all } 1 \leq i \leq k\}. \end{aligned}$$

Finally put

$$m_t := \max_{x \in X} m_t(x). \tag{14}$$

With these definitions we are able to give the precise order (up to a factor) of the spectral gap λ_t:

Theorem 2.1 *There exist constants $\gamma, \Gamma > 0$ such that for all $t \geq 0$*

$$\gamma e^{-\beta_t m_t} \leq \lambda_t \leq \Gamma e^{-\beta_t m_t} \tag{15}$$

Proof: Note that the statement of Theorem 2.1 is the analogue to the spectral gap estimate Theorem 2.1 in Holley and Stroock (1988) for the ordinary SA. As a matter of fact, since we analyze the spectral gap of the instantaneous time operator L_t, which does not "feel" the change of U_t in t, the proof stays the same for our case. Since we are not going to make use of the upper bound in the rest of the paper, we will skip its proof here.
The proof of the lower bound follows the lines of the proof of Lemma 2.7 in Holley and Stroock (1988). By (7) it suffices to show the existence of a constant γ, such that for all t, and for all $f : X \to \mathbb{R}$

$$\frac{\mathcal{E}_t(f,f)}{Var_t(f)} = \frac{-\sum_{x\in X} f(x)[L_t f](x)\pi_t(x)}{Var_t(f)} \geq \gamma e^{-\beta_t m_t}. \tag{16}$$

This is done by bounding the variance as follows :
For any two $x, y \in X$ take a path $p_t^{x,y}$ with vertices $p_t^{x,y}(i)$ from x to y such that

$$\max_{p_t^{x,y}(i)} U_t(p_t^{x,y}(i)) = \min_{\varrho_t^{x,y}} \max_{\varrho_t^{x,y}(j)} U_t(\varrho_t^{x,y}(j))$$

where the $\varrho_t^{x,y}$ are paths form x to y with vertices $\varrho_t^{x,y}(j)$. Let $n_t(x,y)$ be the length of $p_t^{x,y}$ and put

$$N_t := \max_{x,y\in X} n_t(x,y).$$

We introduce the indicator

$$\chi_{z,w}^t(x,y) = \begin{cases} 1 & \text{if there exists } 0 \leq i \leq n_t(x,y) : \\ & p_t^{x,y}(i) = z \text{ and } p_t^{x,y}(i+1) = w \\ 0 & \text{otherwise.} \end{cases}$$

In the following we put $\frac{\chi_{z,w}^t(x,y)}{\alpha(x,y)} = 0$ if $\chi_{z,w}^t(x,y) = 0$. Then for any $f : X \to \mathbb{R}$

$$\begin{aligned} 2Var_t(f) &= \sum_{x,y\in X} (f(x) - f(y))^2 \pi_t(x)\pi_t(y) \\ &= \sum_{x,y\in X} \left(\sum_{i=1}^{n_t(x,y)} f(p_t^{x,y}(i)) - f(p_t^{x,y}(i-1)) \right)^2 \pi_t(x)\pi_t(y) \\ &\leq \sum_{x,y\in X} n_t(x,y) \sum_{i=1}^{n_t(x,y)} \left(f(p_t^{x,y}(i)) - f(p_t^{x,y}(i-1)) \right)^2 \pi_t(x)\pi_t(y) \end{aligned}$$

by Jensen's inequality. Therefore

$$2Var_t(f) \leq N_t \sum_{x,y,z,w\in X} \chi_{z,w}^t(x,y)(f(z) - f(w))^2 e^{-\beta_t \max(U_t(z),U_t(w))} \alpha(z,w)$$

$$
\begin{aligned}
&\frac{\pi_t(x)\pi_t(y)}{e^{-\beta_t \max(U_t(z),U_t(w))}\alpha(z,w)} \\
\leq\ & N_t \left[\max_{z,w\in X} \sum_{x,y\in X} \chi^t_{z,w}(x,y) \frac{\pi_t(x)\pi_t(y)Z_t}{e^{-\beta_t \max(U_t(z),U_t(w))}\alpha(z,w)}\right] \\
& \left[\sum_{z,w\in X} (f(z)-f(w))^2 \frac{e^{-\beta_t \max(U_t(z),U_t(w))}\alpha(z,w)}{Z_t}\right]. \qquad (17)
\end{aligned}
$$

Since N_t is bounded by the size of the vertex set $|X|$ (because of irreducibility) and the second term in brackets in (17) is twice the Dirichlet form $\mathcal{E}_t(f,f)$ (due to (10)) it suffices to calculate that

$$
\begin{aligned}
& \chi^t_{z,w}(x,y) \frac{\pi_t(x)\pi_t(y)Z_t}{e^{-\beta_t \max(U_t(z),U_t(w))}\alpha(z,w)} \\
=\ & \frac{\chi^t_{z,w}(x,y)}{\alpha(z,w)} \frac{\pi(x)\pi(y)}{Z_t} e^{\beta_t(\max(U_t(z),U_t(w))-U_t(x)-U_t(y))} \\
\leq\ & e^{\beta_t m_t}[\frac{\chi^t_{z,w}(x,y)}{\alpha(z,w)} \frac{\pi(x)\pi(y)}{\min_{v\in X}\pi(v)}],
\end{aligned}
$$

since by definition $m_t \geq (\max(U_t(z),U_t(w)) - U_t(x) - U_t(y))$ whenever $\chi^t_{z,w}(x,y) = 1$ for all $x,y,z,w \in X$. Since $[\frac{\chi^t_{z,w}(x,y)}{\alpha(z,w)} \frac{\pi(x)\pi(y)}{\min_{v\in X}\pi(v)}]$ is bounded independently of t, this proves the lower bound. □

Remark 2.2 *Note that under condition (2.11) especially $m := \max_t m_t$ exists and therefore we obtain*

$$\lambda_t \geq \gamma e^{-\beta_t m}$$

for all t.

3 Convergence of the Algorithm via Sobolev inequalities

In this section we are going to analyze the behavior of the density of the distribution of the process with respect to the invariant measure on the basis of a Sobolev inequality, which we first assume and then show to hold true. The results of this section are similar to those for ordinary SA proved by Holley and Stroock in Holley and Stroock (1988). Like they we start with a result on a differential inequality, on which the rest of the section is based.

Lemma 3.1 *Let a,b,c,ε be non-negative numbers witha > 0 and $\varepsilon > 0$, and let $0 < \eta \leq 1$. Let u be a differentiable function $u : \mathbb{R}^+ \to \mathbb{R}^+_{\not 0}$ fulfilling*

$$\frac{d}{dt}u(t) =: u'(t) \leq (\frac{1}{1+t})^\eta[-au^{1+\varepsilon} + bu + cu^{\frac{1}{2}}]$$

for $t > 0$ (where we adopt the convention that primes denote derivatives with respect to t). Then

$$u(t) \leq \left[\frac{4^{\varepsilon}\lambda}{a} \frac{1}{1-\exp(\frac{\varepsilon\lambda}{1-\eta}(1-(1+t)^{1-\eta}))} \right]^{\frac{1}{\varepsilon}}$$

where $\lambda := \max(b, a+c)$.

Remark 3.2 *Lemma 3.1 generalizes Lemma 2.1 in Holley and Stroock (1988) where $\eta = 1$. Our proof is similar.*

Proof: Put $v(t) := 2(u(t))^{\frac{1}{2}}$. Then

$$\begin{aligned} v'(t) &= \frac{u'(t)}{(u(t))^{\frac{1}{2}}} \\ &\leq (\frac{1}{1+t})^{\eta}[-a(\frac{v(t)}{2})^{1+2\varepsilon}) + b\frac{v(t)}{2} + c] \\ &= (\frac{1}{1+t})^{\eta}[-a(1+(\frac{v(t)}{2})^{1+2\varepsilon}) + b\frac{v(t)}{2} + a + c] \\ &\leq (\frac{1}{1+t})^{\eta}[\frac{-a}{4^{\varepsilon}}(1+\frac{v(t)}{2})^{1+2\varepsilon} + \lambda(1+\frac{v(t)}{2})]. \end{aligned}$$

Set $w(t) := (1+\frac{v(t)}{2})\exp(-\frac{\lambda}{2}\frac{(1+t)^{1-\eta}}{1-\eta})$. Then

$$\begin{aligned} w'(t) &= \exp(-\frac{\lambda}{2}\frac{(1+t)^{1-\eta}}{1-\eta})[\frac{v'(t)}{2} - \frac{\lambda}{2}\frac{1}{(1+t)^{\eta}}(1+\frac{v(t)}{2})] \\ &\leq \exp(-\frac{\lambda}{2}\frac{(1+t)^{1-\eta}}{1-\eta})[-\frac{a}{4^{\varepsilon}2}(1+\frac{v(t)}{2})^{1+2\varepsilon}](\frac{1}{1+t})^{\eta} \\ &= \frac{-a}{4^{\varepsilon}2}\exp(\varepsilon\lambda\frac{(1+t)^{1-\eta}}{1-\eta})(w(t))^{1+2\varepsilon}(\frac{1}{1+t})^{\eta}. \end{aligned}$$

Define $z(t) := (w(t))^{-2\varepsilon}$ and obtain

$$z'(t) = \frac{-2\varepsilon w'(t)}{(w(t))^{1+2\varepsilon}}$$

and therefore

$$z'(t) \geq \frac{a\varepsilon}{4^{\varepsilon}}\exp(\varepsilon\lambda\frac{(1+t)^{1-\eta}}{1-\eta})(\frac{1}{1+t})^{\eta}.$$

Hence

$$\begin{aligned} z(t) &\geq \frac{a\varepsilon}{4^{\varepsilon}}\int_0^t \exp(\varepsilon\lambda\frac{(1+s)^{1-\eta}}{1-\eta})(\frac{1}{1+s})^{\eta}ds \\ &= \frac{a\varepsilon}{4^{\varepsilon}}\int_{\frac{1}{1-\eta}}^{\frac{(1+t)^{1-\eta}}{1-\eta}} \exp(\varepsilon\lambda s)ds \\ &= \frac{a}{\lambda 4^{\varepsilon}}[\exp(\varepsilon\lambda\frac{(1+t)^{1-\eta}}{1-\eta}) - \exp(\varepsilon\lambda\frac{1}{1-\eta})]. \end{aligned}$$

Recalling the definitions of z, w and v this yields

$$\exp(\varepsilon\lambda\frac{(1+t)^{1-\eta}}{1-\eta})(u(t))^{-\varepsilon} \geq \frac{a}{\lambda 4^{\varepsilon}}[\exp(\varepsilon\lambda\frac{(1+t)^{1-\eta}}{1-\eta}) - \exp(\varepsilon\lambda\frac{1}{1-\eta})],$$

which implies

$$u(t) \leq \left[\frac{4^{\varepsilon}\lambda}{a}\frac{1}{1-\exp(\frac{\varepsilon\lambda}{1-\eta}(1-(1+t)^{1-\eta}))}\right]^{\frac{1}{\varepsilon}}.$$

□

To make use of Lemma 3.1 let us assume that the following hypothesis on the SA are fulfilled:
There is a $0 \leq \eta \leq 1$ and

(A1) there is a constant $B < \infty$ and such that

$$\beta_t' M + \beta_t R_t \leq \frac{B}{2(1+t)^{\eta}}$$

where $R_t := \sup_x U_t'(x)$ for all $t \geq 0$.

(A2) there is a constant A and a $2 < p < \infty$ such that for all $t \geq 0$

$$||f- < f >_t ||_{p,t}^2 \leq A(1+t)^{\eta}\mathcal{E}_t(f,f)$$

for all $f \in L^p(\pi_t)$, where $< \cdot >_t$ denotes expectation with respect to π_t and $|| \cdot ||_{p,t}$ denotes the norm in $L^p(\pi_t)$.

The following lemma shows that condition (A2) also implies a bound for $||f- < f >_t ||_{2,t}$.

Lemma 3.3 *If (A2) holds then for all $f \in L^2(\pi_t)$*

$$||f- < f >_t ||_{2,t}^{2+\varepsilon} \leq A(1+t)^{\eta}\mathcal{E}_t(f,f)||f- < f >_t ||_{1,t}^{2\varepsilon}$$

Proof: Use Hölder's inequality as in the proof of Lemma 1.5 Holley and Stroock (1988). □

Nex we combine Lemma 3.1 and assumption (A1) as well as the Sobolev inequality (A2) together with Lemma 3.3 to bound the density of the distribution of the process at time t with respect to the invariant measure at time t. To this end set

$$f_t(y) := \frac{P_x(Y_t = y)}{\pi_t(y)}$$

where x is any starting point of the SA process.

Lemma 3.4 *Assume that hypotheses (A1) and (A2) are fulfilled and set* $\varepsilon := (p-2)/p$ *as well as*

$$K := 4^{\varepsilon}(1 + 2AB4^{\varepsilon})\left(1 - \exp\left(\frac{-\varepsilon(4B + 2/(A4^{\varepsilon}))}{1-\eta}\right)\right)^{-1}.$$

Then

$$||f_t||_{2,t} \leq 1 + K^{\frac{1}{2\varepsilon}}$$

for all starting points $x \in X$ *after a waiting time of* $t \geq 2^{\eta-1} - 1$.

Proof: Note that

$$\frac{d}{dt}Z_t = -Z_t(\beta_t' < U_t >_t + \beta_t < U_t' >_t).$$

Therefore

$$\frac{d}{dt}\pi_t(y) = -\pi_t(y)(\beta_t'\hat{U}_t(y) + \beta_t\hat{U}_t'(y))$$

where $\hat{U}_t(y) := U_t(y) - < U_t >_t$. Using the Fokker-Planck equations (2.6) we arrive at

$$\frac{d}{dt}f_t(y) = \frac{1}{\pi_t(y)}\sum_{z\in X}(f_t(z) - f_t(y))\alpha_t(z,y) + f_t(y)(\beta_t'\hat{U}_t(y) + \beta_t\hat{U}_t'(y)).$$

Putting $u(t) := ||f_t - 1||_{2,t}^2 = ||f_t||_{2,t}^2 - 1$,this implies

$$\begin{aligned}
u'(t) &= -2\mathcal{E}_t(f_t, f_t) + \sum_{y\in X} f_t^2(y)(\beta_t'\hat{U}_t(y) + \beta_t\hat{U}_t'(y))\pi_t(y) \\
&= -2\mathcal{E}_t(f_t, f_t) + \sum_{y\in X}(f_t(y) - 1)^2(\beta_t'\hat{U}_t(y) + \beta_t\hat{U}_t'(y))\pi_t(y) \\
&+ 2\sum_{y\in X}(f_t(y) - 1)(\beta_t'\hat{U}_t(y) + \beta_t\hat{U}_t'(y))\pi_t(y) \\
&\leq -2\mathcal{E}_t(f_t, f_t) + \sum_{y\in X}(f_t(y) - 1)^2 2(\beta_t' M + \beta_t R_t)\pi_t(y) \\
&+ 2\sum_{y\in X}(f_t(y) - 1)2(\beta_t' M + \beta_t R_t)\pi_t(y) \\
&\leq (\frac{1}{1+t})^{\eta}[\frac{-2}{A4^{\varepsilon}}u^{1+\varepsilon} + 2Bu + 4Bu^{\frac{1}{2}}].
\end{aligned}$$

Thus the estimate of the lemma follows from Lemma 3.1. □

It remains to show that under conditions (2.11)-(2.13) we have modeled the SA in such a way that assumptions (A1) and (A2) are fulfilled. This is the content of the following lemma.

Lemma 3.5 *Under the assumptions (2.11)-(2.13) for any $p > 2$ there exists a cooling schedule β_t such that the associated SA algorithm satisfies (A1) and (A2).*

Proof: We follow the ideas of the proof of Theorem 2.11 in Holley and Stroock (1988). So we choose $\eta < \alpha$ and α according to assumption (2.12) and we put

$$\beta_t = \frac{\eta}{m+\varepsilon}\log(1+t),$$

where $\varepsilon = \frac{p-2}{p}M$. Now for any fixed $t \geq 0$ on one hand we have

$$\begin{aligned}\beta_t' M + \beta_t R_t &\leq \frac{\eta\, M}{(m+\varepsilon)(1+t)} + \frac{\log(1+t)^\eta}{m+\varepsilon}\frac{C}{(1+t)^\alpha} \\ &\leq const.\frac{1}{(1+t)^\eta}\end{aligned}$$

with a constant depending on η for any $0 < \eta < \alpha$. On the other hand for every $f : X \to \mathbb{R}$

$$\begin{aligned}\sum_{y\in X}(f(y)- < f >_t)^p \pi_t(y) &= \sum_{y\in X}(f(y)- < f >_t)^2 |(f(y)- < f >_t)|^{p-2}\pi_t(y) \\ &\leq \sum_{y\in X}(f(y)- < f >_t)^2 \pi_t(y)||(f- < f >_t)||_\infty^{p-2} \\ &\leq \frac{1}{\gamma}e^{m\beta_t}\mathcal{E}_t(f,f)||f- < f >_t ||_\infty^{p-2}\end{aligned}$$

where the last inequality follows from (2.16).
Since $Z_t \leq 1$ we obtain

$$\begin{aligned}||f- < f >_t ||_\infty^p e^{-\beta_t M}\min_{x\in X}\pi(x) &\leq \\ ||f- < f >_t ||_\infty^p \min_{x\in X}\pi_t(x) &\leq ||f- < f >_t ||_{p,t}^p.\end{aligned}$$

Hence

$$||f- < f >_t ||_{p,t}^p \leq \frac{1}{\gamma}e^{m\beta_t}\mathcal{E}_t(f,f)e^{\beta_t \frac{M(p-2)}{p}}\frac{||f- < f >_t ||_{p,t}^{p-2}}{(\min_{x\in X}\pi(x))^{\frac{p-2}{p}}}.$$

Thus for the finite constant $A = \frac{1}{\gamma(\min_x \pi(x))^{p-2/p}}$

$$||f- < f >_t ||_{p,t}^2 \leq Ae^{\beta_t(m+\frac{M(p-2)}{p})}\mathcal{E}_t(f,f).$$

So the choice of β_t and ε implies the statement of the lemma. □

Summarizing we have proved the following theorem

Theorem 3.6 *Let Y_t be the SA process with time dependent energy function U_t. Under conditions (2.11)-(2.13) the cooling schedule $\beta_t = \frac{\eta}{m+\varepsilon}\log(1+t)$ with $0 < \eta < \alpha$ implies the following bound on the Radon-Nikodym density $f_t(y) := \frac{P_x(Y_t=y)}{\pi_t(y)}$:*

$$||f_t(y)||_{2,t} \leq 1 + K^{\frac{1}{2\varepsilon}}$$

for $t \geq e^{1-\eta} - 1$, and any starting point $x \in X$. Here the constants ε and K are chosen according to lemmas above.

Remarks 3.7 *1. Following the same ideas as in Holley and Stroock (1988), Section 2, it is possible to derive uniform bounds for f_t as well. We omit a detailed discussion here.*

2. *The conditions of Theorem 3.6 are not only sufficient to bound f_t in $L^2(\pi_t)$ (and therefore to ensure convergence of the algorithm), but also are nearly necessary conditions to make SA sensible. Especially if m_t and therefore any bound on U_t diverges to infinity with $t \to \infty$ even the cooling schedule $\beta_t = \frac{\alpha}{m}\log(1+t)$, which we expect to be optimal, is in fact a "heating" schedule, i.e. $\beta_t \to 0$ and therefore the equilibrium measures at time t converge to the uniform distribution on X. Obviously, in this situation SA does not make sense any longer. Also note that M being finite does not imply that $M_t' = 0$ (where $M_t = \sup_x U_t(x)$), but only that, roughly speaking, $M_t' \leq t^{-\delta}$ for some $\delta > 1$, i.e M_t changes noticeably slower than U_t does.*

3. *Note that if in condition (2.12) $\alpha > 1$ U_t admits a limit (for $t \to \infty$) U_∞. As one would expect, SA with time dependent energy function U_t then asymptotically behaves like SA with energy function U_∞.*

Let us finally show, how to use Theorem 3.6 to derive a bound on the entrance time of the algorithm into typical sets. Note that the following theorem especially becomes interesting, when the minima of U_t become concentrated on a subset $V \subset X$.

Theorem 3.8 *Let $V \subset X$ and denote by τ_{V^c} the first entrance time of the algorithm into V^c when starting in $x \in V$. Then under the conditions and with the notations of Theorem 3.6*

$$P(\tau_{V^c} > t) \leq (1 + K^{\frac{1}{2\varepsilon}})\pi_t^{\frac{1}{2}}(V) + \pi_t(V)$$

for $t \geq 2^{\eta-1} - 1$.

Proof: We have

$$\begin{aligned} P(\tau_{V^c} > t) &= P(Y_s \in V \quad \forall s \leq t) \\ &\leq \inf_{s \leq t} P(Y_s \in V), \end{aligned}$$

because of $\{Y_r \in V \quad \forall r \leq t\} \subseteq \{Y_s \in V\}$ if $s \leq t$. Moreover

$$\begin{aligned} (P(Y_t \in V) - \pi_t(V))^2 &= \left(\sum_{y \in X} (f_t - 1) 1_V(y) \pi_t(y) \right)^2 \\ &\leq ||f_t - 1||_{2,t}^2 \pi_t(V). \end{aligned}$$

Hence the assertion follows from Theorem 3.6. $\square$

Surprisingly the SA with time dependent energy function (and $\alpha < 1$) seems to resist calculations as in Götze (1992) for the ordinary SA, where more applicable constants than our $(1 + K^{\frac{1}{2\varepsilon}})$ were obtained.

Remark 3.9 *Throughout the paper we have used a bound for the spectral gap of the form $e^{-\beta_t m}$. As Theorem 2.1 shows these bounds are optimal up to a constant. On the other hand, since usually m, or m_t resp. are as hard to determine as the argument where U_t attains its minimum, these bounds as well as the associated cooling schedules are not very applicable. For practicle use Poincaré-type or Cheeger-type bounds as have been obtained by Diaconis and Stroock (1991) and Götze (1992) seem to be preferable.*

References

E. Aarts, J. Korst, Simulated Annealing and Boltzmann machines, Wiley and Sons, New York (1989).

O. Catoni, Large deviations for annealing algorithms, Ph.D. Thesis, Université Paris-Sud, Orsey (1990).

J.D. Deuschel, C. Mazza, L^2-convergence of time non-homogeneous Markov processes I: Spectral Estimates, Ann. Appl. Prob. 4, (1994), 1012-1056.

P. Diaconis, D. Stroock, Geometric bounds for eigenvalues of Markov chains, Ann. Appl. Prob. 1, (1991), 36-61.

J.A. Fill, Eigenvalue bounds on convergence to stationarity for non-reversible Markov chains, with an application to exclusion processes, Ann. Appl. Prob. 1, (1991), 62-87.

M.I. Freidlin, A.D. Wentzell, Random Pertubations of Dynamical Systems, Springer, Grundlehren d. math. Wissenschaften 260 (1984).

A. Frigerio, G. Grillo, Simulated annealing with time-dependent energy function, Math. Zeit. 213, (1993), 97-116.

S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, IEEE Trans. on Pattern Anal. and Machine Intelligence, PAMI-6,(1984), 721-741.

B. Gidas, Nonstationary chains and the convergence of the annealing algorithm, J. Stat. Phys. 39, (1985), 73-131.

F. Götze, Rate of convergence for simulated annealing processes, Preprint 92-008, SFB 343, Bielefeld (1992).

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

L. van Hemmen, R. Kühn, Collective phenomena in neural networks; in Models of neural networks, E. Domany, L. v. Hemmen, R. Schulte (eds.), Springer, Berlin (1991), pp 1-105.

J. Holland, Adaption in natural and artificial systems, Ann Arbor, University of Michigan Press (1975).

J.J. Hopfield; Neural networks and physical systems with emergent collective computational abilities; Proc. nat. Acad. Sci. USA 79, (1982), 2554-2558.

R. Holley, D. Stroock, Simulated annealing via Sobolev inequalities , Comm. Math. Phys. 115, (1988), 553-569.

R. Horn, C. Johnson, Matrix analysis, Camb. Univ. Press (1985).

S. Kirkpatrick, C. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220, (1983), 621-680.

M. Löwe, On the convergence of genetic algorithms, to appear in Expo. Math. (1996).

D. Mitra, F. Romeo, A. Sangiovanni-Vincentelli, Convergence and finite time behavior of simulated annealing, Proc. 24th Conf. Dec. a. Control, (1985), 761-767.

M. Talagrand; Résultats rigoureux pour le modèle de Hopfield, C.R. Acad. Sci. 321, I,(1995), 309-312.

R.V.V. Vidal (Ed.), Applied Simulated Annealing, Lecture Notes in Economics and Math. Syst. 369, Springer, Berlin (1993).