Piecewise Constant Triangular Cooling Schedules for Generalized Simulated Annealing Algorithms

Cécile COT
Olivier CATONI

LMENS - 96 - 19
Piecewise Constant Triangular Cooling Schedules for Generalized Simulated Annealing Algorithms

Cécile COT†*
Olivier CATONI†

LMENS - 96 - 19
June 1996

†Laboratoire de Mathématiques de l’Ecole Normale Supérieure
45 rue d’Ulm 75230 PARIS Cedex 05
Tel : (33)(1) 44 32 00 00
Adresse électronique : catoni@DMI.ENS.FR

*Laboratoire de Modélisation Stochastique et Statistique
Université de Paris-Sud, Mathématiques, Bât 425, 91405 Orsay
Adresse électronique : cot@stats.MATPS.FR
Piecewise Constant Triangular Cooling Schedules for Generalized Simulated Annealing Algorithms.

Cécile Cot‡, Olivier Catoni†

June 1996

Abstract

We investigate how to tune a generalized simulated annealing algorithm with piecewise constant cooling schedule to get an optimal convergence exponent. The optimal convergence exponent of generalized simulated annealing algorithms has been computed in [6] and [25]. It is reached only with triangular sequences of temperatures, meaning that different finite sequences are used depending on the time resource available for computations (expressed by an overall number of iterations). We show first that it is possible to get close to the optimal convergence exponent uniformly over suitably bounded families of energy landscapes using a fixed number of temperature steps. Then we show that letting the number of steps increase with the time resource, we can build a cooling schedule which is universally robust with respect to the convergence exponent: a fixed triangular sequence of temperatures gives an optimal convergence exponent for any energy landscape. Piecewise constant temperature sequences are often used in practice: in favourable cases, the use of the same temperature during a large number of iterations allows to tabulate the exponential penalties appearing in the transition matrix, thus sparing a significant amount of computer time. The proofs we give rely on Freidlin and Wentzell’s closed formulas for the exit time and point from subdomains of time homogeneous Markov chains.

key words: Large deviations, Metropolis dynamics, cycle decomposition, simulated annealing, difficulty, critical depth.

‡DMI, Laboratoire de Mathématiques de l’Ecole Normale Supérieure, UA 762 du CNRS, 45 rue d’Ulm, 75230 Paris Cedex 05-France and Laboratoire de Modélisation Stochastique et Statistique, Université Paris Sud, Mathématiques, Batiment 425, 91405 Orsay Cedex, France. E-mail address: cot@dni.ens.fr and cot@stats.matups.fr

†DIAM, Laboratoire de Mathématiques de l’Ecole Normale Supérieure, UA 762 du CNRS, 45 rue d’Ulm, 75230 Paris Cedex 05-France. E-mail adress: catoni@dni.ens.fr
1 Introduction.

This paper deals with global optimization algorithms which simulate Markov processes with a finite but large state space $E$, namely the generalized Metropolis and simulated annealing algorithms. These algorithms are used to search for global minima of an energy function $U$ defined on $E$. Their laws of evolution depend on a global parameter $T$, called the temperature. The marginal distributions of the simulated Markov chains concentrate around the global minima of the energy function as temperature is lowered and the number of iterations is accordingly increased.

In the case of the Metropolis dynamic, the temperature is constant during the relaxation. We establish optimal convergence properties towards states of minimal energy when $T$ is low enough.

Simulated annealing is a method to speed up the Metropolis algorithm. The temperature is gradually lowered as the relaxation proceeds. The performance of simulated annealing depends on the way the temperature is cooled down while the algorithm is running. The good ways to cool down the temperature depend on some critical quantities related to the energy landscape $(E, U)$ and the connectivity of the used dynamic.

Among the pioneering works on simulated annealing is the paper by D. and S. Geman [16] which introduced this algorithm as a tool for statistical image processing and proved convergence results for it. They chose an annealing schedule of the type $T(k) = C/\log k$ and proved convergence for $C$ large enough. This cooling schedule is far from optimal, even for an optimal choice of the constant $C$. The optimal value for $C$ can be found in various papers using different methods of proof (see Hajek [18], Holley & Stroock [19], Miclo [22], Chiang & Chow [12], Hwang & Sheu [20], Tsitsiklis [28], Catoni [3], [4]).

Rates of convergence for simulated annealing algorithms can be established in different ways. Two main approaches dominate: the semi-group approach initiated by Holley and Stroock [19] and the large deviation approach initiated by Freidlin and Wentzell in their book about random perturbations of dynamical systems [15].

The optimal exponent of the convergence rate of decreasing cooling schedule is established in Catoni [6] for reversible simulated annealing processes and in Trouvé [25] in the general non-reversible case. It is shown in Catoni [5] that the cooling schedules optimizing the marginal distribution after $N$ iterations are triangular sequences of temperatures $(T_{n,N})_{n=1,\ldots,N}$, where all the temperatures are chosen as a function of the horizon $N$. Exponential cooling schedules of the form $(T_{n,N}^N) = A(\rho(N))^n$ are proved to be almost optimal for a proper choice of the function $\rho(N)$ in Catoni [6]. We propose here two other types of cooling schedules. They are both piecewise constant triangular sequences of the exponential form
\[ T_n^N = T_{\text{max}}(N) \left( \frac{T_{\text{min}}(N)}{T_{\text{max}}(N)} \right)^{\frac{k-1}{r}} \] for \((k - 1) \frac{N}{r} < n \leq k \frac{N}{r}, k = 1, \ldots, r\), where \(r\) is the number of temperature steps and \(N\) is the total number of iterations to be performed.

In the first cooling schedule the number of steps \(r(N) = r\) is independent of \(N\). We show that in this situation, for a proper choice of \(T_{\text{max}}(N)\) and \(T_{\text{min}}(N)\), we miss the optimal convergence exponent by a factor which tends to one when \(r\) tends to infinity uniformly on suitably bounded subsets of energy landscapes.

In the second cooling schedule, we let \(r(N)\) be a function of \(N\). We remark first that choosing \(T_{\text{max}}(N)\) and \(T_{\text{min}}(N)\) as previously, we can, for a suitable choice of \(r(N)\), get the optimal convergence exponent uniformly on bounded subsets of energy landscapes. Then we see that taking \(r(N)\) a little larger than necessary, we can let \(T_{\text{max}}(N) = T_{\text{max}}\) be a constant independent of \(N\) and get, for a proper, but fixed choice of \(T_{\text{min}}(N)\), the optimal convergence exponent for any energy landscape. This is how we build a second cooling schedule which can be said to be “universally robust” with respect to the convergence exponent.

From the technical point of view, we first prove results about the rate of convergence of the generalized Metropolis chain in the spirit of Catoni [8]. This proof rests on Freidlin and Wentzell’s closed formulas giving the mean value of the exit time from subdomains of the energy landscape and the invariant distribution. Then we apply these results to each step of a piecewise constant cooling schedule. The purpose of each step is to reach with a probability close to one a state below a given energy level, which has to be linked in a proper way with the temperature. The best strategy would ideally consist in tuning the temperature as a function of the maximum energy barrier which has to be jumped over to get down from one target energy level to the other. This strategy however requires a precise knowledge of the energy landscape and would lead to non robust choices of the temperature sequence. In order to get a robust cooling schedule, needed for practical applications, we do as if the energy landscape were of homogeneous difficulty. The difficulty of the energy landscape is by definition the inverse of its optimal convergence exponent. It is the maximum ratio between the depth of the cycles (see Freidlin and Wentzell [15]) of the state space and their ground state energy above the minimum, the cycles containing the global ground states being excluded. To choose a robust cooling schedule, we do as if all the cycles had the same ratio. This leads naturally to consider a geometrical sequence of target energy levels, corresponding to a geometrical sequence of temperatures. The nice thing in this construction is that the sequence of target energy levels is not a parameter of the cooling schedule, it is only a tool in the proof and, as it turns out, we can show that the same sequence of temperatures gives an optimal convergence exponent for different energy landscapes by tuning the target energy sequence differently, as a function of the real (changing and unknown) difficulty. We show in this way that the chain reaches lower and lower energy levels with an
almost optimally close to one probability, until it reaches a global ground state.

In all this discussion, convergence rates are measured by the rate of decrease of
the probability to be above a given energy level after a given number of iterations.
Although it is not the only possible notion of convergence (the minimum energy
level reached along the whole trajectory could also be considered), it has proved
to be a meaningful one to compare algorithms in full generality, that is without
making special assumptions about the state space.

2 Formalizing the problem.

2.1 Generalized Metropolis chain.

Let $E$ be a finite space. We consider a family of time homogeneous Markov chains
on $E$, $\mathcal{F} = (E^N, (X_n)_{n\in\mathbb{N}}, \mathcal{B}, P_\beta)_{\beta\in\mathbb{R}_+}$ indexed by a positive parameter $\beta$, called
the inverse temperature.

More precisely, we consider the coordinate process $(X_n)_{n\in\mathbb{N}}$, on $E^\mathbb{N}$ defined
by $X_n(x) = x_n$, for any $x \in E^\mathbb{N}$. We define $\mathcal{B} = \sigma(X_1^-(\mathcal{P}(E)), n \in \mathbb{N})$, the $\sigma$ field generated by $(X_n)$. We consider a family of probability distributions $(P_\beta)$ on $(E^N, \mathcal{B})$ indexed by $\beta$. Under each distribution $P_\beta$ the coordinate process is a Markov chain. The Markov chain $((X_n), P_\beta)$ is entirely characterized once we give its initial law $P_\beta \circ X_0^{-1} \in \mathcal{M}_1(E)$ and its transitions kernel

$$p_\beta(x, y) = P_\beta(X_n = y \mid X_{n-1} = x).$$

We suppose that the transitions of $P_\beta$ are "rare", meaning thereby that they obey
some Large Deviation Principle with speed $\beta$, and rate function $V : E \times E \to \mathbb{R}_+ \cup \{+\infty\}$, namely that the following hypothesis LDP($a$) holds for some positive constant $a$:

$$\forall (x, y) \in E \times E, \forall \beta \in \mathbb{R}_+, \quad a e^{-\beta V(x, y)} \leq p_\beta(x, y) \leq a^{-1} e^{-\beta V(x, y)}. \quad (1)$$

The rate function $V$ is often called the communication cost function. We assume
also that $V$ is irreducible, which means that

$$\forall x, y \in E \times E, \quad \exists i_0, \ldots, i_r, \quad i_0 = x, \quad i_r = y, \quad V(i_0, i_1) + \ldots + V(i_{r-1}, i_r) < \infty.$$

Under these assumptions, $((X_n), P_\beta)$ is called a generalized Metropolis chain.

2.2 Generalized simulated annealing chain.

Let $(\beta_n) = (\beta_n)_{n\in\mathbb{N}}$ be an arbitrary non decreasing sequence of inverse
temperatures. We consider the associated non homogeneous Markov chain $\mathcal{R} =
(E^N, (X_n)_{n\in\mathbb{N}}, \mathcal{B}, P_{(\beta_n)})$ on $E$ with transitions defined by

$$P_{(\beta_n)}(X_n = y \mid X_{n-1} = x) = p_{\beta_n}(x, y),$$

4
where \((p_\beta)_{\beta \in \mathbb{R}_+}\) is a family of transition matrices satisfying the same hypothesis as in the previous paragraph.

We call \(\mathcal{R}\) a generalized simulated annealing chain.

### 2.3 Minimizing an "energy" function.

As we need to minimize an "energy" function \(U : E \to \mathbb{R}\), it is necessary to add some constraints, linking the energy function \(U\) and the communication cost function \(V\). The proper link between \(U\) and \(V\) is ensured by the weak reversibility condition of Hajek and Trouvé: We define, for any \(x, y \in E\),

\[
\Gamma_{x,y} = \{(\gamma_i)_{i=0}^r : \gamma_0 = x, \gamma_r = y\},
\]

the set of all paths joining \(x\) to \(y\),

\[
H_U(x, y) = \min_{\gamma \in \Gamma_{x,y}} \max_i U(\gamma_i) + V(\gamma_i, \gamma_{i+1})
\]

and we assume that

\[
H_U(x, y) = H_U(y, x), \quad x, y \in E.
\]

Under this assumption, the unique invariant measure \(\mu_\beta\) of \(p_\beta\) satisfies for some positive constant \(b\) (independent of \(\beta\)), for all \(x\) in \(E\),

\[
b e^{-\beta(U(x) - U_{\min})} \leq \mu_\beta(x) \leq b^{-1} e^{-\beta(U(x) - U_{\min})}.
\]

It is proved in Trouvé [25] that the weak reversibility condition is necessary and sufficient, namely that (2) is equivalent to (3). Moreover, for any irreducible rate function \(V\), there is an energy function \(U\), uniquely defined up to an additive constant, satisfying (3) and therefore also (2).

In order to simplify notations, we will assume without loss of generality that the minimum of the energy function is zero.

**remark:** A sufficient condition to have (2) is that \(\forall x, y \in E, \ V(x, y) < \infty \implies V(y, x) < \infty\), and that \(V(x, y) = (U(y) - U(x))^+\) whenever \(V(x, y) < \infty\). This is the case of "classical" Metropolis and simulated annealing algorithms, the first one to have been studied.

After giving some more precisions on the stationary measure, we will find the time needed for the generalized Metropolis chain to reach the basin of attraction of the global minima of \(U\). For this purpose, we will first prove a general lemma about the entrance time in arbitrary subdomains of \(E\). From this we will get an upper bound for the marginal distributions of the generalized Metropolis chain. It will allow us to give its optimal rate of convergence.
3 An upper bound for the marginal distributions of generalized Metropolis chains.

Lemma 3.1 (Behaviour of the stationary measures): For any $\beta$, the generalized Metropolis Markov chain $((X_n), P_\beta)$ is irreducible and its invariant measure $\mu_\beta$ is unique. The family of invariant measures satisfies:

there exists a positive constant $b$ depending only on $a$ and $|E|$ such that for any $x$ in $E$, for any $\beta \in \mathbb{R}_+$,

$$b e^{-\beta U(x)} \leq \mu_\beta(x) \leq b^{-1} e^{-\beta U(x)}.$$  \hspace{1cm} (4)

Thus the stationary measure $\mu_\beta$ concentrates on states of minimal energy $U$ when $\beta$ is big enough.

Proof: Using (1) and the assumption that $V$ is irreducible, we deduce that the generalized Metropolis chain $((X_n), P_\beta)$ is irreducible for any $\beta$. Thus, its invariant measure $\mu_\beta$ is unique. Let us then consider the expression of the invariant measure given by Freidlin and Wentzell ([15] page 177)

$$\mu_\beta(x) = \frac{\sum_{g \in G(x)} p_\beta(g)}{\sum_{y \in E, g \in G(y)} p_\beta(g)},$$

where $p_\beta(g) = \prod_{(u,v) \in g} p_\beta(u,v)$, and where for any arbitrary non-empty subset $W$ of $E$, $G(W)$ denotes the set of all oriented graphs over $E$ that satisfy the following conditions:

• there is no arrow starting from a point of $W$,
• each point of $W^c$ is the initial point of exactly one arrow,
• there is no cycle in the graph or equivalently, for each point $x$ in $W^c$, there exists a path in $W^c$ leading from $x$ to $W$.

We then use hypothesis (1) on the transitions of the generalized Metropolis chain to find that there exists a positive constant $b$ depending only on $a$ and $|E|$, such that for any $x$ in $E$, for any $\beta$,

$$b \exp \left( -\beta \left( \min_{g \in G(x)} V(g) - \min_{y \in E, g \in G(y)} \min_{v \in \{y\}} V(g) \right) \right) \leq \mu_\beta(x) \leq b^{-1} \exp \left( -\beta \left( \min_{g \in G(x)} V(g) - \min_{y \in E, g \in G(y)} \min_{v \in \{y\}} V(g) \right) \right),$$

with $V(g) = \sum_{(u,v) \in g} V(u,v)$.  

(6)
Finally the weak reversibility condition of Hajek-Trouvé guarantees that for all $x$ in $E$,

$$U(x) - U_{\text{min}} = U(x) = \min_{g \in G(\{x\})} V(g) - \min_{y \in E} \min_{g \in G(\{y\})} V(g).$$

\[\square\]

**Lemma 3.2** (estimation of the law of the entrance time in an arbitrary subdomain of $E$, independently of the initial point):

There exists a constant $c > 0$ such that for any subdomain $A$, $A \neq E$, $A \neq \emptyset$, any $n \in \mathbb{N}$, any inverse temperature $\beta > 0$:

$$\max_{x \in E \setminus A} P_\beta(\tau(A) > n \mid X_0 = x) \leq \exp\left(-c n e^{-\beta H(E \setminus A)}\right),$$

where $\tau(A) = \inf\{n \in \mathbb{N} \mid X_n \in A\}$ denotes the entrance time in $A$ (and $\lfloor r \rfloor = \max\{n \in \mathbb{Z} \mid n \leq r\}$ denotes the integer part of the real number $r$).

**Proof:** Let $A \subset E$, $A \neq E$, $A \neq \emptyset$, Let $n \in \mathbb{N}$, and $\beta > 0$. For any integer $k$,

$$\max_{x \in E \setminus A} P_\beta(\tau(A) > n \mid X_0 = x) \leq \left(\max_{y \in E \setminus A} P_\beta(\tau(A) > k \mid X_0 = y)\right)^{\lfloor \frac{n}{k} \rfloor} \leq \left(\frac{1}{k} \max_{y \in E \setminus A} E_\beta(\tau(A) \mid X_0 = y)\right)^{\lfloor \frac{n}{k} \rfloor}.$$

The upper bound mainly depends on the expectation of the entrance time in $A$, from which Wentzell and Freidlin give the following expression (in a lemma on irreducible Markov chains, page 182 of [15]):

$$E_\beta(\tau(A) \mid X_0 = y) = \frac{\sum_{z \in E \setminus A} \sum_{g \in G_{y,z}(A \cup \{z\})} p_\beta(g)}{\sum_{g \in G(A)} p_\beta(g)}.$$

where $G_{y,z}(W)$ is the set of all graphs in $G(W)$ such that $y$ leads to $z \in W$, when $y \in E \setminus W$, and where for convenience $G_{z,z}(W) = G(W)$.

Let us recall the definition of the depth of the subset $E \setminus A$ (see Trouvé [25], [27], Catoni [6], [8], or Catoni and Cerf [11] for further explanations):

$$H(E \setminus A) = \max_{x \in E \setminus A} \min_{y \in A} H_U(x, y) - U(x) = \max_{y \in E \setminus A} \lim_{\beta \to \infty} \frac{1}{\beta} \log E_\beta(\tau(A) \mid X_0 = y) = \min_{g \in G(A)} V(g) - \min_{g \in G(A \cup \{z\})} V(g).$$

7
Using the hypothesis about the upper bound and lower bound of the transition kernel \((1)\), we see that for some constant \(C\) depending only on \(a\) and \(E\):

\[
\max_{y \in E \setminus A} E_\beta(\tau(A) \mid X_0 = y) \leq C \exp(\beta H(E \setminus A)).
\]

From this we get an upper bound for the law of entrance in the subset \(A\), independent of the initial point:

\[
\max_{x \in E \setminus A} P_\beta(\tau(A) > n \mid X_0 = x) \leq \exp \left( \left\lceil \frac{n}{k} \right\rceil \log \left( \frac{C e^{\beta H(E \setminus A)}}{k} \right) \right).
\]

Choosing the least integer \(k\) such that \(k \geq C e^{\beta H(E \setminus A)} + 1\), we get

\[
\max_{x \in E \setminus A} P_\beta(\tau(A) > n \mid X_0 = x) \leq \exp \left( - \left\lfloor \frac{n}{d e} e^{-\beta H(E \setminus A)} \right\rfloor \right)
\]

where the constant \(c\) only depends on the cardinality of the state space and on the constant \(a\). □

The next lemma will use the first critical depth of the energy landscape, which is defined to be \(H_1(V) = H(E \setminus \arg \min U)\).

**Lemma 3.3 (Upper bound for the marginal distributions of generalized Metropolis chains):**

There exists a constant \(d > 0\), depending only on \(a\) and \(|E|\) such that, for any inverse temperature \(\beta > 0\), the generalized Metropolis chain \(P_\beta\) satisfies the following condition: for all \(n \in \mathbb{N}\), and all \(y \in E\),

\[
\max_{x \in E} P_\beta(X_n = y \mid X_0 = x) \leq \exp \left( - \left\lfloor \frac{n}{d e} e^{-\beta H_1(E,V)} \right\rfloor \right) + de^{-\beta U(y)}.
\]

**Corollary 3.4 (on the probability of failure of the minimization):**

There exists a constant \(d' > 0\), depending only on \(a\) and \(|E|\), such that for any energy level \(\eta > 0\) and any inverse temperature \(\beta > 0\), the generalized Metropolis chain \((X_n)_{n \in \mathbb{N}}\) satisfies, for all \(n \in \mathbb{N}\):

\[
\max_{x \in E} P_\beta(U(X_n) \geq \eta \mid X_0 = x) \leq \exp \left( - \left\lfloor \frac{n}{d' e} e^{-\beta H_1(E,V)} \right\rfloor \right) + d' e^{-\beta \eta}.
\]

**Remark:** We can deduce the corollary from the lemma ever since we note that:

\[
\max_{x \in E} P_\beta(U(X_n) \geq \eta \mid X_0 = x) \leq \sum_{y \in E \setminus \{y \mid U(y) \geq \eta\}} \max_{x \in E} P_\beta(X_n = y \mid X_0 = x)
\]

\[
\leq 1 \wedge |E| \exp \left( - \left\lfloor \frac{n}{d' e} e^{-\beta H_1(E,V)} \right\rfloor \right) + d|E|e^{-\beta \eta}.
\]
Proof of lemma: Let $B = \text{arg min} U$ be the “bottom” of $E$, let $n \in \mathbb{N}, \beta > 0$ and $y \in E \setminus B$. We have

$$
\max_{x \in E} P_\beta(X_n = y \mid X_0 = x) \\
\leq \max_{x \in E \setminus B} P_\beta(\tau(B) > n \mid X_0 = x) \\
+ \sum_{z \in B} \max_{x \in E} P_\beta(X_n = y, X_{\tau(B)} = z, \tau(B) \leq n \mid X_0 = x).
$$

Using lemma 3.2 to estimate from above the law of the entrance time in the bottom of $E$, we find:

$$
\max_{x \in E} P_\beta(X_n = y \mid X_0 = x) \leq \exp\left(- \left|ncc^{-\beta H_1(E,V)}\right|\right) \\
+ \max_{z \in B, k \in \mathbb{N}} P_\beta(X_k = y \mid X_0 = z).
$$

We can transform the second member of the upper bound, using the following remark:

Let $f_n(t) = \frac{P_\beta(X_n = t \mid X_0 = z)}{\mu_\beta(t)}$, $t \in E$. Each $f_{n-1}$ is a convex combination of $f_n(t), t \in E$.

Indeed, we have $\sum_{t \in E} f_n(t) p_\beta(t, u) \frac{\mu_\beta(t)}{\mu_\beta(u)} = f_{n+1}(u)$ with $\sum_{t \in E} p_\beta(t, u) \frac{\mu_\beta(t)}{\mu_\beta(u)} = 1$.

Thus we find

$$
\max_{t \in E} f_n(t) \leq \max_{t \in E} f_0(t) = \frac{1}{\mu_\beta(z)},
$$

and therefore for any $y$ and $z \in E$,

$$
P_\beta(X_n = y \mid X_0 = z) \leq \frac{\mu_\beta(y)}{\mu_\beta(z)}.
$$

After using the lower bound and upper bound of the invariant measure (3) stated in lemma 3.1, we finally find the expected upper bound of the probability of failure:

$$
\max_{x \in E} P_\beta(X_n = y \mid X_0 = x) \leq \exp\left(- \left|ncc^{-\beta H_1(E,V)}\right|\right) \\
+ b^{-2} \max_{z \in B} \exp(-\beta(U(y) - U(z))) \\
\leq \exp\left(- \left|\frac{n}{d}c^{-\beta H_1(E,V)}\right|\right) + de^{-\beta U(y)},
$$

where the constant $d$ only depends on the cardinality of the state space and on the constant $a$. 
4 A lower bound for the marginal distributions of generalized Metropolis chains.

In this paragraph, we will use the decomposition of the state space into cycles introduced by Freidlin and Wentzell. Let us recall that $\Pi \subset E$ is a cycle if it is a component of one of the equivalence relations

$$\mathcal{R}_\lambda = \{(x, y) \in E^2 : H_U(x, y) \leq \lambda\} \cup \{(x, x) : x \in E\}, \quad \lambda \in \mathbb{R}.$$ 

For more details on the cycle decomposition, see Catoni [8], Trouvé [25], [26], [27] or Catoni and Cerf [11].

**Lemma 4.1** There exists a constant $C > 0$ depending only on $a$ and $|E|$, such that for any cycle $\Pi$ of $E$, there exists an inverse temperature $\beta_0$ such that for any $\beta \geq \beta_0$, any $n$, and any $x \in \Pi$,

$$P_\beta(\tau(\Pi) > n \mid X_0 = x) \geq \frac{1}{C} \exp(-C n e^{-\beta H(\Pi)}).$$

**Corollary 4.2** There exists a positive constant $K$ depending only on $a$ and $|E|$, there exists an inverse temperature $\beta_0$, and $\eta_0 > 0$ such that for any $\beta \geq \beta_0$, any $0 < \eta \leq \eta_0$, any $n$,

$$\max_{x \in E} P_\beta(U(X_n) \geq \eta \mid X_0 = x) \geq \frac{1}{K} \left( \exp \left( - \left\lfloor K n e^{-\beta H_1} \right\rfloor \right) \vee \exp \left( - \beta \min_{y \in U(y) \geq \eta} U(y) \right) \right).$$

**Proof of corollary:** Let $\Pi$ be a cycle in $E$ such that $\Pi \cap B = \emptyset$ and $H(\Pi) = H_1$. Let $\eta_0 = U(\Pi) > 0$.

We have for any $x \in \Pi$,

$$P_\beta(U(X_n) \geq \eta \mid X_0 = x) \geq P_\beta(\tau(\Pi) > n \mid X_0 = x)$$

So, with a direct application of lemma 4.1, we get a first lower bound:

$$P_\beta(U(X_n) \geq \eta \mid X_0 = x) \geq \frac{1}{C} \exp \left( - C n e^{-\beta H_1} \right)$$

Another lower bound can be obtained, using the fact that

$$\mu_\beta(y) = \sum_{x \in E} \mu_\beta(x) P_\beta(X_n = y \mid X_0 = x) \leq \max_{x \in E} P_\beta(X_n = y \mid X_0 = x),$$

10
and the lower bound (3) for the invariant probability at temperature \( \beta \). Indeed, we get that for some constant \( b > 0 \) depending only on \( a \) and \( |E| \),

\[
\max_{x \in E} P_\beta(U(X_n) \geq \eta \mid X_0 = x) = \sum_{y, U(y) \geq \eta} \max_{x \in E} P_\beta(X_n = y \mid X_0 = x) \\
\geq b \exp(-\beta \min_{y, U(y) \geq \eta} U(y)).
\]

We end the proof by choosing the constant in the corollary to be equal to \( K = C \vee b^{-1} \).

**Proof of lemma:** Let \( \Pi \) be a cycle in \( E \) and \( x \) a point in \( \Pi \). We have

\[
P_\beta(\tau(\Pi) \leq n \mid X_0 = x) \leq \min_{y \in \Pi} P_\beta(\tau(\Pi) \leq n \mid X_0 = y) \\
+ \max_{y \in \Pi} P_\beta(X_{\tau(\Pi \cup \{y\})} \not= y \mid X_0 = x).
\]

Then we use the property of cycles which states that we visit with a large probability any point before leaving, which proves that, for \( \beta \) large enough, we have

\[
\min_{y \in \Pi} P_\beta(X_{\tau(\Pi \cup \{y\})} = y \mid X_0 = x) \geq \frac{1}{2}.
\]

We use the remark that

\[
\sum_{k=0}^{+\infty} P_\beta(\tau(\Pi) > kn \mid X_0 = x) \geq E_\beta(\tau(\Pi) \mid X_0 = x) \frac{1}{n},
\]

and that

\[
\sum_{k=0}^{+\infty} P_\beta(\tau(\Pi) > kn \mid X_0 = x) \leq \sum_{k=0}^{+\infty} \left( \max_{y \in \Pi} P_\beta(\tau(\Pi) > n \mid X_0 = y) \right)^k \\
\leq \frac{1}{\min_{y \in \Pi} P_\beta(\tau(\Pi) \leq n \mid X_0 = y)}.
\]

so that using expression of the expectation of the exit time out of \( \Pi \) given by Freidlin and Wentzell and the hypothesis (1) about the lower bound and upper bound of the transition kernel, we find that there exists a constant \( c > 0 \) depending only on \( a \) and \( |E| \) such that

\[
\min_{y \in \Pi} P_\beta(\tau(\Pi) \leq n \mid X_0 = y) \leq \frac{n}{\max_{x \in \Pi} E_\beta(\tau(\Pi) \mid X_0 = x)} \\
\leq n c \exp(-\beta H(\Pi)).
\]

Finally we find a lower bound for \( P_\beta(\tau(\Pi) > n \mid X_0 = x) \) of the form \( \frac{1}{2} - n c \exp(-\beta H(\Pi)) \) when \( \beta \) is large enough.

Using the Markov property, we have for any integer \( k \),

\[
P_\beta(\tau(\Pi) > n \mid X_0 = x) \geq \left( \min_{x \in \Pi} P_\beta(\tau(\Pi) > k \mid X_0 = x) \right)^\frac{1}{2}.
\]
We end the proof by choosing \( k = \left\lfloor \frac{1}{4c} e^{\beta H(1)} \right\rfloor \), so that for \( \beta \) large enough we have
\[
\left\lceil \frac{n}{k} \right\rceil \log\left( \frac{1}{2} - k c e^{-\beta H(1)} \right) \geq -\left( \log 4 \right) \left( 1 + 4 c n e^{-\beta H(1)} \right).
\]

5 Optimal rate of convergence of Metropolis algorithms.

We deduce from these bounds the optimal convergence rate of the generalized Metropolis dynamic:

**Theorem 5.1 (Optimal convergence rate of the generalized Metropolis algorithm).**

For any state space \( E \), any irreducible function \( V \) defined on \( E \), for any transition kernel satisfying (1), the associated generalized Metropolis chain is such that there exists a constant \( d > 0 \), depending only on \( a \) and \( |E| \), such that for any \( \eta > 0 \) and any \( N \), putting
\[
\beta(N) = \frac{1}{H_1} \left( \log \frac{N}{d} - \log \left( \frac{\eta}{H_1} \log \frac{N}{d} + 1 \right) \right),
\]
we have
\[
\max_{x \in E} P_{\beta(N)}(U(X_N) \geq \eta \mid X_0 = x) \leq d \left( \frac{d \eta}{N H_1} \log N \right) \frac{\eta}{H_1}.
\]

An immediate consequence is that
\[
\inf_{\beta > 0} \max_{x \in E} P_{\beta}(U(X_N) \geq \eta \mid X_0 = x) \leq d \left( \frac{d \eta}{N H_1} \log N \right) \frac{\eta}{H_1}.
\]

Furthermore, there exists another constant \( d' > 0 \), depending only on \( a \) and \( E \), there exists \( \eta_0 > 0 \) (depending also on \( V \) ), such that for any \( 0 < \eta \leq \eta_0 \), such that \( \eta \in U(E) \), for \( N \) large enough, we have
\[
\inf_{\beta > 0} \max_{x \in E} P_{\beta}(U(X_N) \geq \eta \mid X_0 = x) \geq d' \left( \frac{d \eta}{N H_1} \log N \right) \frac{\eta}{H_1}.
\]

So, for the best temperature, and the worst initial point of the Metropolis dynamic, the speed of convergence is at best of order \( \left( \frac{1}{N} \log N \right)^{\frac{\eta}{H_1}} \). When \( \eta \in U(E) \) is small enough, the convergence speed obtained with \( \beta(N) \) is almost optimal in the sense that
\[
\lim_{N \to +\infty} \frac{1}{\log N} \log \max_{x \in E} P_{\beta(N)}(U(X_N) \geq \eta \mid X_0 = x) = \frac{\eta}{H_1}.
\]
Proof: The upper bound is deduced from corollary 3.4 and the lower bound from corollary 4.2. The inverse temperature $\beta(N)$ has been chosen such that both terms in the upper bound of corollary 3.4 are of the same order. In the case of the lower bound, to deal with the case when $\beta < \beta_0$ in corollary 4.2, we remark that the second term of the lower bound given in corollary 4.2 holds without condition on $\beta$.

6 Rate of convergence of simulated annealing with a piecewise constant triangular cooling schedule

The theorems of this paragraph will be derived from the following technical proposition:

**proposition 6.1** For any $a$ and $E$, there exists a constant $c > 0$ such that for any irreducible rate function $V$ defined on $E$, any transition kernel satisfying (1), any $\eta > 0$, any couple of integers $r, N$, such that $r$ divides $N$, any decreasing sequences $(\lambda_k)_{k=0,\ldots,r-1}$, $(\eta_k)_{k=0,\ldots,r-1}$, any increasing sequence $(\gamma_k)_{k=0,\ldots,r-1}$, such that

1) $\eta_{r-1} \leq \eta$,
2) $\lambda_k \geq (1 + D)\eta_{k-1}$, for all $k = 1, \ldots, r$,

where $D = D(E, V) = \max_{x \in E \setminus B} \min_{y \in B} \frac{H_U(x, y) - U(x)}{U(x)}$, the generalized simulated annealing chain stopped at iteration number $N$ with cooling schedule $\beta_n^N = \gamma_k$, $k \frac{N}{r} < n \leq (k + 1) \frac{N}{r}$, $k = 0, \ldots, r - 1$,

satisfies

$$P_{\beta^N}(U(X_N) \geq \eta | X_0 = x) \leq \left( 1 + \frac{cN}{r} e^{-\gamma_1(\lambda_1 - \eta_0)} \right) \left( \exp \left( - \left[ \frac{N}{rc} e^{-\gamma_0 H_1} \right] \right) + ce^{-\gamma_0 \eta_0} \right)$$

$$+ \sum_{k=2}^{r-1} \left( 1 + \frac{cN}{r} e^{-\gamma_k(\lambda_k - \eta_{k-1})} \right) \left( \exp \left( - \left[ \frac{N}{rc} e^{-\gamma_{k-1}\lambda_{k-1}(1+\frac{1}{r})^{k-1}} \right] \right) + ce^{-\gamma_{k-1} \eta_{k-1}} \right)$$

$$+ \exp \left( - \left[ \frac{N}{rc} e^{-\gamma_{r-1}\lambda_{r-1}(1+\frac{1}{r})^{r-1}} \right] \right) + ce^{-\gamma_{r-1} \eta_{r-1}}.$$  

We will deduce from this proposition the two following theorems:
**Theorem 6.2** For any $a$ and $E$, there exists a constant $c > 0$ such that for any irreducible rate function $V$ defined on $E$, any transition kernel satisfying (1), any constants $d$, $H$ and $D$ satisfying $d \geq c$, $H \geq H_1(E, V)$ and $D \leq D(E, V)$, for any constant $0 < \eta \leq \frac{H}{D}$, for any integers $r, N$ such that $r$ divides $N$ and $\frac{N}{r} \geq d$, the generalized simulated annealing chain stopped at time $N$ with cooling schedule

$$\beta^N_n = \gamma_k, \quad k\frac{N}{r} < n \leq (k + 1)\frac{N}{r}, \quad k = 0, \ldots, r - 1$$

$$\gamma_k = \frac{1}{\eta D} \left( \log \frac{N}{rd} - \log \left( 1 + \frac{1}{D} \log \frac{N}{rd} \right) \right) \left( \frac{H}{\eta D} \right)^{\frac{1}{r}}$$

satisfies

$$\max_{x \in E} P_{(\beta^N_n)}(U(X_N) \geq \eta \mid X_0 = x) \leq cdr \left( \frac{rd}{N} \right)^{\frac{1}{r}} \left( 1 + \frac{1}{D} \log \left( \frac{N}{rd} \right) \right)^{(1 + \frac{1}{r})} \left( \frac{H}{\eta D} \right)^{-\frac{1}{r}}.$$

**Two noticeable choices for the number of steps $r$ are worth being mentioned:**

- If we choose $r = \frac{\log(H/\eta D)}{D} \log N$, then the probability of failure is at most of order $O((\log N)^2(1+1/D)N^{-1/D})$. For this first optimization, the number of steps increases like $\log N$ and their length increases like $N/\log N$.

- If we choose $r \sim \frac{\log(\frac{H}{\eta D})}{\log(1 + \alpha)}$ independently of $N$ (where $\alpha > 0$ is a small parameter), the probability of failure is at most of order $\left( \frac{1}{N} \right)^{\frac{1}{D(1+\alpha)}} (\log N)^{(1+1/D)/(1+\alpha)}$.

**Remarks:**

i) In order to get the optimal exponent of convergence, we need not know the exact value of $D$, $H_1$ and $c$, but only some bounds for these quantities.

ii) We can understand from this theorem why simulated annealing speeds up the Metropolis dynamic.

When the parameter $r$ is properly chosen, the main exponent of the probability of failure (namely $\frac{1}{D}$ or $\frac{1}{D(1+\alpha)}$) does not depend on the precision $\eta$ with which we want to get close to the minima of the energy $U$, whereas in the case of Metropolis algorithm, this exponent is equal to $\frac{1}{H}$ and tends to zero when $\eta$ does.

In the case when we want to find out an exact ground state, simulated annealing will be asymptotically faster than the Metropolis algorithm when the level $\eta$ being next to the ground state energy level (which we set to zero by convention) is such that $\eta < \frac{H}{D}$. Indeed in this case, $\frac{1}{H}$ is the optimal convergence exponent for Metropolis and

$$\left( \frac{1}{N} \right)^{\frac{1}{H}} \ll \left( \frac{1}{N} \right)^{\frac{1}{H}}$$
when $N$ is large.

**Theorem 6.3 (Universally Robust Cooling Schedule)**

With the same notation, for any $a > 0$ and $E$ there exists a constant $C > 0$ such that for any irreducible function $V$, any $\gamma_0 > 0$, any parameter $\varepsilon > 0$, any large enough $M$, putting

$$
\gamma_k = \gamma_0 (1 + (\log M)^{-1-\varepsilon})^k \\
r = \lceil (\log M)^{1+2\varepsilon} \rceil \\
N = Mr
$$

we have

$$
\max_{x \in E} P_{(\beta^N)}(U(X_N) \geq \eta \mid X_0 = x) \leq M^{-1/D} C^{(1+1/D)} (\log M)^{2(1+\varepsilon)+1/D},
$$

and therefore

$$
\max_{x \in E} P_{(\beta^N)}(U(X_N) \geq \eta \mid X_0 = x) \leq N^{-1/D} C^{(1+1/D)} (\log N)^{2(1+\varepsilon)(1+1/D)}.
$$

**Remark:**

We state this less precise theorem to show that it is possible to get the optimal convergence exponent with a fixed cooling schedule, independent of the energy landscape, when the number of steps $r$ is allowed to be a function of $N$. However, it should be noticed that it is somehow a theoretical result, since the value $M_0$ of $M$ for which the bound given in the theorem starts to hold depends on the energy landscape (that is on $a$, $E$, and $V$).

**Proof of the Proposition:**

Let us fix $N$ and $d > 0$.

Let us put $\lambda_0 = +\infty$ and introduce the events:

$$
\mathcal{B}_k = \left\{ U(X_n) + V(X_n, X_{n+1}) \leq \lambda_k, \ n \in \mathbb{N}, \ k\frac{N}{r} \leq n < (k+1)\frac{N}{r} \right\}
$$

$$
\mathcal{A}_k = \mathcal{B}_k \cap \left\{ U(X_{(k+1)\frac{N}{r}}) < \eta_k \right\}.
$$

and the short notation $P_N$ for $P_{(\beta^N)}$.

For any initial point $x$ in $E$:

$$
P_N(U(X_N) \geq \eta \mid X_0 = x) \leq P_N(U(X_N) \geq \eta_{r-1} \mid X_0 = x)
$$

$$
\leq 1 - P_N(\bigcap_{k=0}^{r-1} \mathcal{A}_k \mid X_0 = x)
$$

$$
\leq \sum_{k=1}^{r-1} P_N(\mathcal{A}_k \cap \bigcap_{l=0}^{k-1} \mathcal{A}_l \mid X_0 = x) + P_{\gamma_0}(\mathcal{A}_0 \mid X_0 = x).
$$
Applying corollary 3.4 to the finite generalized Metropolis chain \((X_n)_{0 \leq n \leq N}\), we know that there exists a constant \(c\) depending only on \(|E|\) and \(a\) such that:

\[
P_{\gamma_0}(U(X_{\frac{\xi}{r}}) > \eta_0 \mid X_0 = x) \leq \exp \left( - \frac{N}{r} e^{-\gamma_0 H_1(E, V)} \right) + ce^{-\gamma_0 \eta_0}.
\]

Furthermore, for \(k \geq 1\),

\[
P_N(\mathcal{A}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x) = P_N(\mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x) + P_N(U(X_{(k+1)}_{\xi} > \eta_k, \mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x).
\]

And for all \(y \in E\),

\[
P_N(X_{(k+1)}_{\xi} = y, \mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x)
\]

\[
\leq \max_{z \in U(\tau(z) < \eta_{k-1})} P_N(X_{(k+1)}_{\xi} = y, \mathcal{B}_k \mid X_{k} = z)
\]

\[
\leq \max_{z \in U(\tau(z) < \eta_{k-1})} P_N(X_{\xi} = y, \tau(C_{z}) > \frac{N}{r} \mid X_0 = z),
\]

where \(C_z\) denotes the smallest cycle containing \(z\) such that \(\lambda_k < H(C_z) + U(C_z)\), and \(\tau(C_z)\) denotes the exit time from \(C_z\).

Observe that \(U(C_z) = 0\), since if we had \(U(C_z) > 0\), then we would have also \(U(C_z) + H(C_z) \leq U(z)(1 + D) \leq \eta_{k-1}(1 + D) \leq \lambda_k\). This would contradict the definition of \(C_z\).

Note that \(C_z\) is a cycle of communication level at most equal to \(\lambda_k\). We need to control the probability of paths of length \(N/r\) which do not get out of \(C_z\).

Consider the Markov chain \((Y_n)_{n \in \mathbb{N}}\) defined on the restricted state space \(C_z\), whose transition matrix \(P(Y_n = y_2 \mid Y_{n-1} = y_1) = q(y_1, y_2)\) for any \(y_1, y_2 \in C_z\) is defined by:

\[
q(y_1, y_2) = \begin{cases} 
    p_{\gamma_k}(y_1, y_2) & \text{when } y_1 \neq y_2 \\
    1 - \sum_{y_3 \in C_z} q(y_1, y_3) & \text{otherwise}.
\end{cases}
\]

This new Markov chain is simply the previous chain \((X_n)_{n \in \mathbb{N}}\) reflected on the boundary of \(C_z\). Merely by the fact that for any \(y_1, y_2 \in C_z\)

\[
0 \leq p_{\gamma_k}|_{C_z \times C_z}(y_1, y_2) \leq q(y_1, y_2),
\]

we can write

\[
P_{\gamma_k} \left( X_{\frac{\xi}{r}} = y, \tau(C_z) > \frac{N}{r} \mid X_0 = z \right) \leq P \left( Y_{\frac{\xi}{r}} = y \mid Y_0 = z \right).
\]
Furthermore, the new transitions $q$ obey some Large Deviation Principle with rate function $V$ restricted to $C_z$ and the chain $(Y_n)$ is irreducible. Thus, since $C_z$ is a cycle of minimal energy, the energy of $(Y_n)$ is exactly $U$ restricted to $C_z$.

Applying lemma 3.3 to $(Y_n)$, we see that there exists a constant $c > 0$, depending only on $a$ and $|E|$ such that

$$P_N(X_{(k+1)}^{\frac{\lambda}{\gamma}} = y, \mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x)$$

$$\leq \max_{\gamma, U(z) < \gamma_{k-1}} \exp \left(- \left[ \frac{N}{\gamma_c} e^{-\gamma_k H_1(C, V)} \right] \right) + ce^{-\gamma_k U(y)}.$$

Furthermore we see that $H_1(C_z, V) \leq \left(1 + \frac{1}{D} \right)^{-1} \lambda_k$, coming back to the definition of $H_1(C_z, V) = \max \{ H(C), C \text{ cycle}, C \subset C_z, U(C) > 0 \}$, and observing that for any cycle $C$ strictly included in $C_z$, with positive energy, we have $H(C) \left(1 + \frac{1}{D} \right) \leq H(C) + U(C) \leq \lambda_k$.

Thus, substituting in the inequality, we find

$$P_N(X_{(k+1)}^{\frac{\lambda}{\gamma}} = y, \mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x)$$

$$\leq \exp \left(- \left[ \frac{N}{\gamma_c} e^{-\gamma_k \lambda_k (1 + \frac{1}{D})^{-1}} \right] \right) + ce^{-\gamma_k U(y)}.$$

For the same reason there exists a constant $c(a, |E|)$ such that we have also

$$P_N(U(X_{(k+1)}^{\frac{\lambda}{\gamma}}) \geq \lambda_k, \mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x)$$

$$\leq \exp \left(- \left[ \frac{N}{\gamma_c} e^{-\gamma_k \lambda_k (1 + \frac{1}{D})^{-1}} \right] \right) + ce^{-\gamma_k \lambda_k}.$$

Furthermore,

$$P_N(\mathcal{B}_k, \mathcal{A}_0, \ldots, \mathcal{A}_{k-1} \mid X_0 = x)$$

$$= \sum_{z, U(z) < \gamma_{k-1}} P_N(\mathcal{B}_k \mid X_k^{\frac{\lambda}{\gamma}} = z) P_N(X_k^{\frac{\lambda}{\gamma}} = z, \mathcal{A}_0, \ldots, \mathcal{A}_{k-2}, \mathcal{B}_{k-1} \mid X_0 = x),$$

and,

$$P_N(\mathcal{B}_k \mid X_k^{\frac{\lambda}{\gamma}} = z) \leq \sum_{n=k}^{(k+1)} P_{\gamma_k}(U(X_n) + V(X_n, X_{n+1}) > \lambda_k \mid X_k^{\frac{\lambda}{\gamma}} = z)$$

$$= \sum_{n=k}^{(k+1)} \sum_{U(u) + V(u, v) > \lambda_k} P_{\gamma_k}(X_n = u \mid X_k^{\frac{\lambda}{\gamma}} = z)p_{\gamma_k}(u, v)$$

$$\leq \sum_{n=k}^{(k+1)} \sum_{U(u) + V(u, v) > \lambda_k} \frac{\mu_{\gamma_k}(u)}{\mu_{\gamma_k}(v)} p_{\gamma_k}(u, v).$$
Thus, applying inequalities (1) and (3) on the transition matrix and the invariant measure, we finally find that there exists a constant $c$ only depending on $a$ and $|E|$ such that

$$P_{\gamma_k}(B_k \mid X_{k \infty} = z) \leq c \frac{N}{r} e^{-\gamma_k (\lambda_k - U(z))}.$$  

We then deduce that for a constant $c$ big enough, for any $z$ such that $U(z) < \eta_{k-1}$,

$$P_{\gamma_k}(B_k \mid X_{k \infty} = z)P_N(X_{k \infty} = z, A_0, \ldots, A_{k-2}, B_{k-1} \mid X_0 = x)$$

$$\leq c \frac{N}{r} e^{-\gamma_k (\lambda_k - U(z))} \left\{ \exp \left( - \left| \frac{N}{r} e^{-\gamma_{k-1} \lambda_{k-1} (1 + \frac{1}{\rho})^{-1}} \right| \right) + ce^{-\gamma_{k-1} U(z)} \right\}$$

$$\leq c \frac{N}{r} e^{-\gamma_k (\lambda_k - \eta_{k-1})} \left\{ \exp \left( - \left| \frac{N}{r} e^{-\gamma_{k-1} \lambda_{k-1} (1 + \frac{1}{\rho})^{-1}} \right| \right) + ce^{-\gamma_{k-1} \eta_{k-1}} \right\}.$$

Thus for a constant $c(a, |E|)$ big enough,

$$P_{\gamma_k}(B_k, A_0, \ldots, A_{k-1} \mid X_0 = x)$$

$$\leq c \frac{N}{r} e^{-\gamma_k (\lambda_k - \eta_{k-1})} \left( \exp \left( - \left| \frac{N}{r} e^{-\gamma_{k-1} \lambda_{k-1} (1 + \frac{1}{\rho})^{-1}} \right| \right) + ce^{-\gamma_{k-1} \eta_{k-1}} \right).$$

Finally, we find the following upper bound of the probability of failure: for any initial point $x$ in $E$

$$P_N(U(X_N) \geq \eta \mid X_0 = x)$$

$$\leq \sum_{k=1}^{r-1} \left\{ c \frac{N}{r} e^{-\gamma_k (\lambda_k - \eta_{k-1})} \left( \exp \left( - \left| \frac{N}{r} e^{-\gamma_{k-1} \lambda_{k-1} (1 + \frac{1}{\rho})^{-1}} \right| \right) + ce^{-\gamma_{k-1} \eta_{k-1}} \right) \right\}$$

$$+ \exp \left( - \left| \frac{N}{r} e^{-\gamma_k \lambda_k (1 + \frac{1}{\rho})^{-1}} \right| \right) + ce^{-\gamma_k \eta_k}$$

$$+ \exp \left( - \left| \frac{N}{r} e^{-\gamma_0 H_1} \right| \right) + ce^{-\gamma_0 \eta_0}. \quad \square$$

**proof of theorem 6.2:**
Let us define, in order to simplify formulas,

$$\left\{ \begin{array}{l}
\alpha = \frac{1}{\rho D} \left( \log \frac{N}{rD} - (1 + D) \log (1 + \frac{1}{D} \log \frac{N}{rD}) \right) \\
\rho = \left( \frac{H}{\eta D} \right)^{\frac{1}{\gamma_k}}
\end{array} \right.$$  

and put

$$\left\{ \begin{array}{l}
\eta_k = \frac{\alpha}{\gamma_k} \\
\lambda_0 = +\infty, \\
\lambda_k = \frac{1}{\gamma_k} \left( 1 + \frac{1}{D} \right) \left( \log \frac{N}{rD} - \log (1 + \alpha) \right)
\end{array} \right.$$
Note that \(\rho > 1\) and that \(\alpha \leq \frac{1}{D} \log \frac{N}{rd}\) since we chose \(\eta\) small enough such that 
\[
\frac{H}{\eta} \geq D.
\] Indeed, we have
\[
\alpha \leq \frac{1}{\rho D} \log \frac{N}{rd} \leq \frac{1}{D} \log \frac{N}{rd}.
\]
Note also that \(\lambda_k \geq (1 + D)\eta_{k-1} \geq \eta_{k-1}\), since \(\alpha \leq \frac{1}{D} \log \frac{N}{rd}\) and that \((\eta_k)_{k \leq r-1}\)
and \((\lambda_k)_{k \leq r-1}\) are geometrical decreasing sequences.
Moreover \(\eta_{r-1} \leq \eta\), since
\[
\eta_{r-1} = \frac{\alpha}{\gamma_0} \rho^{1-r} = \rho^{-r} \frac{\frac{1}{D} \left( \log \frac{N}{rd} - (1 + D) \log \left( 1 + \frac{1}{D} \log \frac{N}{rd} \right) \right)}{\left( \log \frac{N}{rd} - \log \left( 1 + \frac{1}{D} \log \frac{N}{rd} \right) \right)} \leq \frac{H}{D} \rho^{-r} \leq \eta,
\]
thus conditions \(i)\) and \(ii)\) are satisfied.

Coming back to the definitions of \(\eta_k\) and \(\lambda_k\), we check that for any \(d > c\),
\[
\frac{N}{r} e^{-\gamma_k (\lambda_k - \eta_{k-1})} = d (1 + \alpha)^{1 + \frac{1}{D}} \left( \frac{N}{rd} \right)^{-\frac{1}{D}} e^{\alpha \rho} \\
= d \left( \frac{1 + \alpha}{1 + \frac{1}{D} \log \frac{N}{rd}} \right)^{1 + \frac{1}{D}}.
\]
Using the inequality \(\alpha \leq \frac{1}{D} \log \frac{N}{rd}\), we find that
\[
\frac{N}{r} e^{-\gamma_k (\lambda_k - \eta_{k-1})} \leq d.
\]
We also have
\[
\exp \left( - \left[ \frac{N}{rd} e^{-\gamma_k \lambda_k (1 + \frac{1}{D})^{-1}} \right] \right) \leq \exp \left( - \frac{N}{rd} e^{-\gamma_k \lambda_k (1 + \frac{1}{D})^{-1} + 1} \right) \leq e^{-\alpha},
\]
and
\[
\exp \left( - \left[ \frac{N}{rd} e^{-\gamma_H} \right] \right) \leq \exp \left( - \frac{N}{rd} e^{-\gamma_H} + 1 \right) \leq \exp \left[ - \frac{N}{rd} \exp \left( - \log \frac{N}{rd} + \log \left( 1 + \frac{1}{D} \log \frac{N}{rd} \right) \right) + 1 \right] \leq \exp \left[ - \frac{1}{D} \log \frac{N}{rd} \right] \leq e^{-\alpha}.
\]
finally, \( e^{-\gamma_k \eta_k} = e^{-\alpha} \).

Thus, \( P_N(U(X_N) \geq \eta \mid X_0 = x) \leq r(1 + cd)(1 + c)e^{-\alpha} \), and we end the proof of the theorem by changing the value of \( c \) into \((1 + c)^2\).

**Proof of theorem 6.3:**

Let us fix \( M \) and \( \gamma_0 \). Put

\[
\rho = 1 + (\log M)^{-1-\varepsilon} \quad \gamma_k = \gamma_0 \rho^k \quad r = \left[ (\log M)^{1+2\varepsilon} \right] \quad N = Mr
\]

with parameter \( \varepsilon > 0 \), and put also

\[
\lambda_0 = +\infty \\
\lambda_k = \frac{\lambda}{\rho^k} \\
\eta_k = \frac{\alpha}{\rho^k}
\]

with

\[
\lambda = \frac{1}{\gamma_0} \left( 1 + \frac{1}{D} \right) \left( \log \frac{M}{c} - \log \left( \frac{1}{D} \log M \right) \right) \\
\alpha = \frac{1}{\gamma_0 \rho} \left( \frac{1}{D} \log \frac{M}{c} - \left( 1 + \frac{1}{D} \right) \log \left( \frac{1}{D} \log M \right) \right).
\]

To justify the tuning of parameters, we need to check up conditions \( i) \) and \( ii) \) of proposition 6.2.

Condition \( i) \eta_{r-1} \leq \eta \), which can be also written like \( \alpha \leq \eta \rho^{r-1} \), is satisfied for large enough \( M \) since we have \( \rho^r = O(M) \).

Condition \( ii) \lambda_k \geq (1 + D)\eta_{k-1} \), which can also be written like \( \lambda \geq (1 + D)\rho\alpha \), is also satisfied. We end the proof by substituting the values of \( \gamma_k, \lambda_k \) and \( \eta_k \) into proposition 6.1.

## 7 Conclusion.

We showed that properly tuned piecewise constant cooling schedules guaranteed the concentration of the marginal distributions of the chain around the global minima of \( U \) at almost optimal rate. These cooling schedules give the optimal convergence exponent uniformly for suitably bounded subsets of rate functions. One of them is even universally robust with respect to the convergence exponent, at the price of some loss in the subsequent terms of the convergence rate. Although these results are only concerned with the asymptotic of a large computer resource,
they indicate that exponential piecewise triangular schedules are suitable when little is known in advance about the energy landscape. However they should not disguise the practical fact that a suitable choice of $T_{\text{min}}(N)$ and $T_{\text{max}}(N)$ in the general formula $T_n^N = T_{\text{max}}(N) \left( \frac{T_{\text{min}}(N)}{T_{\text{max}}(N)} \right)^{(k-1)/(r-1)}$, $(k-1)/r < n < k N/r$, can save a lot of computer time. Therefore this paper should not be taken as an indication about how to choose the parameters of cooling schedules, but more as an indication that some kinds of cooling schedules are more promising than others. Their general form, temperature steps of equal length forming a geometrical sequence, is of a kind often used in applications and previously unjustified on rigorous mathematical ground. The proofs use Freidlin and Wentzell’s theory for homogeneous chains and are therefore more straightforward than those provided in Catoni [6] and Trouvé [25], [27] for more general temperature sequences.

A practical interest of piecewise constant schedules is that some computations can be saved by a proper tabulation of the exponential, when the rate function takes only a small number of values. This is for instance the case with the noisy Ising model (see Catoni [2]), and more generally with Pott’s model, which are often used in image analysis. In connection with this issue, we showed what exponent of convergence can be reached with a fixed number of steps $r$ independent of the computer resource $N$. We showed also that the proper choice for $r(N)$ in order to reach the optimal exponent is a slowly increasing function of $N$, which is a favourable situation, since the steps will be long and few, and therefore the tabulation procedure will still be efficient.

References


22


