The Annals of Applied Probability
1998, Vol. 8, No. 2, 375–396

PIECEWISE CONSTANT TRIANGULAR COOLING SCHEDULES FOR GENERALIZED SIMULATED ANNEALING ALGORITHMS

BY CÉCILE COT[1] AND OLIVIER CATONI

Ecole Normale Supérieure

We investigate how to tune a generalized simulated annealing algorithm with piecewise constant cooling schedule to get an optical convergence exponent. The optimal convergence exponent of generalized simulated annealing algorithms has been computed by Catoni and Trouvé. 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 tabulating 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.

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 ([22], [15]). 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.

Received July 1996; revised July 1997.

[1] Also at Université Paris Sud.

AMS 1991 *subject classifications*. Primary 82C80, 65C05; secondary 60J10, 90C42, 60F10.

Key words and phrases. Simulated annealing, Metropolis dynamic, Markov chains with rare transitions, triangular cooling schedules, optimal convergence exponent.

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. Good ways to cool down the temperature depend on some critical quantities related to the energy landscape (E, U) and the connectivity of the dynamic used.

Among the pioneering works on simulated annealing is the paper by Geman and Geman [17], 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 [19], [20], [23], [12], [21], [30], [3] and [4]).

Rates of convergence for simulated annealing algorithms can be established in different ways. Two main approaches dominate: the semigroup approach initiated in [20] and the large deviation approach initiated by Freidlin and Wentzell in their book about random perturbations of dynamical systems [16].

The optimal exponent of the convergence rate of decreasing cooling schedules is established in [6] for reversible simulated annealing processes and in [27] in the general nonreversible case. It is shown in [5] that the cooling schedules optimizing the marginal distribution after N iterations are triangular sequences of temperatures $(T_n^N)_{n=1,\dots,N}$, where all the temperatures are chosen as a function of the horizon N. Exponential cooling schedules of the form $(T_n^N) = A(\rho(N))^n$ are proved to be almost optimal for a proper choice of the function $\rho(N)$ in [6], [23] (see also [10]). We propose here two other types of cooling schedules. They are both piecewise constant triangular sequences of the exponential form

$$T_n^N = T_{\max}(N)\left(\frac{T_{\min}(N)}{T_{\max}(N)}\right)^{(k-1)/(r-1)} \quad \text{for } (k-1)\frac{N}{r} < n \le k\frac{N}{r}, \quad k = 1, \dots, r,$$

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 $T_{\max}(N)$ and $T_{\min}(N)$, 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 $T_{\max}(N)$ and $T_{\min}(N)$ 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 $T_{\max}(N) = T_{\max}$ be a constant independent of N and get, for a proper but fixed choice of $T_{\min}(N)$, 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 [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 consist ideally 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 nonrobust choices of the temperature sequence. In order to get a robust cooling schedule, needed for practical applications, we proceed 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 [16]) 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 proceed as if all the cycles had the same ratio. This leads naturally to considering a geometrical sequence of target energy levels, corresponding to a geometrical sequence of temperature. 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. *The generalized Metropolis chain.* Let E be a finite space. We consider a family of time homogeneous Markov chains on E, $\mathscr{F} = (E^{\mathbb{N}}, (X_n)_{n \in \mathbb{N}}, \mathscr{B}, P_\beta)_{\beta \in \mathbb{R}_+}$ indexed by a positive parameter β, 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 $\mathscr{B}=\sigma(X_n^{-1}(\mathscr{P}(E)),\ n\in\mathbb{N})$, the σ field generated by (X_n). We consider a family of probability distributions (P_β) on $(E^{\mathbb{N}},\mathscr{B})$ indexed by β. Under each distribution P_β the coordinate process is a Markov chain. The Markov chain $((X_n),P_\beta)$ is entirely characterized once we give its initial law $P_\beta\circ X_0^{-1}\in\mathscr{M}_1(E)$ 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_β are "rare," meaning thereby that they obey some large deviation principle with speed β, and rate function $V\colon E\times E\to\mathbb{R}_+\cup\{+\infty\}$, namely that the following hypothesis LDP (a) holds for some positive constant a:

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

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

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

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

It is a family of irreducible Markov chains indexed by $\beta\in\mathbb{R}_+$. To characterize the corresponding family $(\mu_\beta)_{\beta\in\mathbb{R}_+}$ of invariant probability measures, let us introduce, following [16], for any subset W of E, the set $G(W)$ of all oriented graphs over E such that the following hold.

1. There is no arrow starting from the points of W.
2. Each point of $E\setminus W$ is the initial point of exactly one arrow.
3. There is no cycle in the graph or equivalently, for each point x in $E\setminus W$, there is a path in the graph leading from x to W.

Then, if we put $p_\beta(g)=\prod_{(u,v)\in g}p_\beta(u,v)$, $g\in G(W)$, the invariant probability measure μ_β can be expressed as

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

which proves the following lemma where $V(g)=\sum_{(u,v)\in g}V(u,v)$. (It is a special case of Lemma 3.2, page 178 of [16].)

Lemma 2.1. *There is a positive constant b depending only on a and $|E|$ such that for any $x\in E$, any $\beta\in\mathbb{R}_+$,*

$$(2)\qquad be^{-\beta U(x)}\le\mu_\beta\le b^{-1}e^{-\beta U(x)},$$

where

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

Lemma 2.1 is an approximate Gibbs formula in which U plays the role of a "virtual energy function."

It is also possible, and often easier, to characterize U in terms of paths. This is the weak reversibility condition of Hajek [19] and Trouvé [28]. Let $\tilde{U}: E \to \mathbb{R}$ be any candidate energy function. For any $x, y \in E$, let

$$\Gamma_{x,y} = \left\{ (\gamma_i)_{i=0}^{r},\, r \in \mathbb{N}^*,\, \gamma_0 = x,\, \gamma_r = y \right\},$$

be the set of all paths joining x to y. Let

$$H_{\tilde{U}}(x, y) = \min_{\gamma \in \Gamma_{x,y}} \max_{i=1,\dots,r} \tilde{U}(\gamma_{i-1}) + V(\gamma_{i-1}, \gamma_i)$$

be the elevation function corresponding to $\tilde{U}$. It is proved in [27] that $U = \tilde{U} - \tilde{U}_{\min}$ if and only if the elevation function corresponding to $\tilde{U}$ is symmetric:

$$(3) \qquad H_{\tilde{U}}(x, y) = H_{\tilde{U}}(y, x), \qquad x, y \in E.$$

REMARK. A sufficient condition to have (3) is that $\forall\ x, y \in E$, $V(x, y) < \infty \Rightarrow V(y, x) < \infty$, and that $V(x, y) = (U(y) - U(x))^+$ whenever $V(x, y) < \infty$. This is the case for the "classical" Metropolis and simulated annealing algorithms, the first ones which have been studied.

2.2. *The generalized simulated annealing chain.* Let $(\beta.) = (\beta_n)_{n \in \mathbb{N}}$ be an arbitrary nondecreasing sequence of inverse temperatures. We consider the corresponding nonhomogeneous Markov chain $\mathscr{R} = (E^{\mathbb{N}}, (X_n)_{n \in \mathbb{N}}, \mathscr{B}, P_{(\beta.)})$ on E with transitions defined by

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

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

We call $\mathscr{R}$ a generalized simulated annealing chain.

3. An upper bound for the marginal distributions of the generalized Metropolis chain. 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.

For any subset A of E, we let $\tau(A)$ be the first hitting time of A:

$$\tau(A) = \inf\{n \in \mathbb{N} \mid X_n \in A\}.$$

We let

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

be the depth of A (see [27], [29], and [6], [8] or [11] for further explanations). It is clear from this definition that the depth of a domain is a nonnegative quantity.

We recall (see [16], page 182) that

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

where $G_{x,y}(W)$ is the set of graphs $g\in G(W)$ which lead from x to y, when $x\in E\setminus W$ and $y\in W$, and where, by convection, $G_{y,y}(W) = G(W)$.

Lemma 3.1 (Estimate for the entrance time in an arbitrary subdomain of E). *There exists a positive constant c, depending only on a and $|E|$, such that, for any proper subdomain A of E, $A\neq E$ and $A\neq\varnothing$, any $n\in\mathbb{N}$, any inverse temperature $\beta>0$,*

$$\max_{x\in E\setminus A} P_\beta(\tau(A) > n\mid X_0 = x) \le \exp(-\lfloor cn\exp(-\beta H(E\setminus A))\rfloor),$$

where $\lfloor r\rfloor = \max\{n\in\mathbb{Z}\mid n\le r\}$ is the integer part of the real number r.

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

$$\begin{aligned} \max_{x\in E\setminus A} P_\beta(\tau(A) > n\mid X_0 = x) &\le \Bigl(\max_{y\in E\setminus A} P_\beta(\tau(A) > k\mid X_0 = y)\Bigr)^{\lfloor n/k\rfloor} \\ &\le \Bigl(\frac{1}{k}\max_{y\in E\setminus A} E_\beta(\tau(A)\mid X_0 = y)\Bigr)^{\lfloor n/k\rfloor}. \end{aligned}$$

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 $|E|$,

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

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

$$\max_{x\in E\setminus A} P_\beta(\tau(A) > n\mid X_0 = x) \le \exp\biggl(\Bigl\lfloor\frac{n}{k}\Bigr\rfloor\log\Bigl(\frac{C}{k}\exp(\beta H(E\setminus A))\Bigr)\biggr).$$

Choosing the least integer k such that $k \geq C \exp(\beta H(E \setminus A) + 1)$, we get

$$\max_{x \in E \setminus A} P_\beta(\tau(A) > n \mid X_0 = x) \leq \exp(-\lfloor nc \exp(-\beta H(E \setminus A)) \rfloor)$$

where the constant $c = 1/(eC + 1)$ 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 3.2 (Upper bound for the marginal distributions of the generalized Metropolis chain). *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_β satisfies the following condition: for all $n \in \mathbb{N}$, and all $y \in E$,*

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

COROLLARY 3.3 (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 $n \in \mathbb{N}$,*

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

REMARK. A similar result could have been obtained using spectral gap estimates. For instance, it is well known that in the reversible case, that is, when

$$\mu_\beta(x) p_\beta(x, y) = \mu_\beta(y) p_\beta(y, x), \qquad x, y \in E,$$

we have

$$\left| P_\beta(X_n \in A \mid X_0 = x) - \mu_\beta(A) \right| \leq \left(\frac{\mu(A)}{\mu(x)} \right)^{1/2} \exp(-cn \exp(-\beta\Lambda)),$$

where c depends only on a and $|E|$ and where

$$\Lambda = \max_{x, y} H_U(x, y) - U(x) - U(y) \geq H_1(V).$$

In order to replace Λ by $H_1(V)$ and to cover the nonreversible case, it is necessary to use some technical tricks, such as pasting some states together and considering a new process based on a symmetrized transition matrix (see [14], [13], [18], [24], [25], [26] or [8]). There are also some minor differences between the results we give here and those obtained by the spectral gap approach, since we avoid here an extra $(1/\mu(x))^{1/2}$ term. On the other hand, in some situations, Poincaré estimates for the spectral gap give a better control on the dependence of the constants with the size of the state space. The dependence in $|E|$ given by the Freidlin and Wentzell approach is crude; this is why we did not make it explicit in this paper.

PROOF OF LEMMA 3.2. Let $B = \arg\min U$ be the "bottom" of E, let $n \in \mathbb{N}$, $\beta > 0$ and $y \in E \setminus B$. We have

$$\begin{aligned}\max_{x\in E} & P_\beta(X_n = y \mid X_0 = x)\\ &\le \max_{x\in E\setminus B} P_\beta\big(\tau(B) > n \mid X_0 = x\big)\\ &\quad + \sum_{z\in B} \max_{x\in E} P_\beta\big(X_n = y,\, X_{\tau(B)} = z, \tau(B) \le n \mid X_0 = x\big).\end{aligned}$$

Using Lemma 3.1 to estimate from above the law of the entrance time in the bottom of E, we find

$$\begin{aligned}\max_{x\in X} P_\beta(X_n = y \mid X_0 = x) &\le \exp(-\lfloor nc\exp(-\beta H_1(V))\rfloor)\\ &\quad + \max_{z\in B,\, k\in\mathbb{N}} P_\beta(X_k = y \mid X_0 = z).\end{aligned}$$

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

$$f_n(t) = \frac{P_\beta(X_n = t \mid X_0 = z)}{\mu_\beta(t)}, \qquad t \in E.$$

Each $f_{n+1}(u)$ is a convex combination of $f_n(t)$, $t \in E$. Indeed, we have

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

Thus we find

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

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

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

After using the lower bound and upper bound of the invariant measure (2) stated in Lemma 2.1, we finally find the expected upper bound of the probability of failure:

$$\begin{aligned}\max_{x\in E} P_\beta(X_n = y \mid X_0 = x) &\le \exp(-\lfloor nc\exp(-\beta H_1(V))\rfloor)\\ &\quad + b^{-2}\max_{z\in B}\exp(-\beta(U(y) - U(z)))\\ &\le \exp\left(-\left\lfloor \frac{n}{d}\exp(-\beta H_1(V))\right\rfloor\right) + d\exp(-\beta U(y)),\end{aligned}$$

where the constant d only depends on the cardinality of the state space and on the constant a. The corollary is straightforward. □

4. A lower bound for the marginal distributions of the generalized Metropolis chain.

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

$$\mathscr{R}_\lambda = \{(x, y) \in E^2 : H_U(x, y) \le \lambda\} \cup \{(x, x) : x \in E\}, \qquad \lambda \in \mathbb{R}.$$

For more details on the cycle decomposition, see [8], [27–29] or [11].

LEMMA 4.1. *There exists a constant $C > 0$ depending only on a and $|E|$, such that for any cycle Π of E, there exists an inverse temperature β_0 such that for any $\beta \ge \beta_0$, any n, and any $x \in \Pi$,*

$$P_\beta(\tau(E \setminus \Pi) > n \mid X_0 = x) \ge \frac{1}{C} \exp(-Cn \exp(-\beta H(\Pi))),$$

where, as in the previous section, $H(\Pi)$ is the depth of Π.

COROLLARY 4.2. *There exists a positive constant K depending only on a and $|E|$, there exist an inverse temperature β_0 and $\eta_0 \in (U(E) \setminus \{0\})$ such that for any $\beta \ge \beta_0$, any $\eta \in (U(E) \cap\,]0, \eta_0])$, any n,*

$$\begin{aligned}\max_{x \in E} P_\beta(U(X_n) \ge \eta \mid X_0 = x) \\ \ge \frac{1}{K}\big(\exp(-\lfloor Kn \exp(-\beta H_1(V))\rfloor) \vee \exp(-\beta\eta)\big).\end{aligned}$$

PROOF. Let Π be a cycle in E such that $\Pi \cap B = \varnothing$ and $H(\Pi) = H_1(V)$. Let $\eta_0 = \min_{x \in \Pi} U(x)$ and let $\overline{\Pi} = E \setminus \Pi$.

We have for any $x \in \Pi$, any $\eta \in (U(E) \cap\,]0, \eta_0])$,

$$P_\beta(U(X_n) \ge \eta \mid X_0 = x) \ge P_\beta\big(\tau(\overline{\Pi}) > n \mid X_0 = x\big).$$

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

$$P_\beta(U(X_n) \ge \eta \mid X_0 = x) \ge \frac{1}{C} \exp\big(-Cn \exp(-\beta H_1(V))\big).$$

Another lower bound can be obtained, using the fact that

$$\begin{aligned}\mu_\beta(y) &= \sum_{x \in E} \mu_\beta(x) P_\beta(X_n = y \mid X_0 = x) \\ &\le \max_{x \in E} P_\beta(X_n = y \mid X_0 = x)\end{aligned}$$

and the lower bound (2) for the invariant probability at temperature β. Indeed, we get that for some constant $b > 0$ depending only on a and $|E|$,

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

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

Proof of Lemma 4.1. Let Π be a cycle in E and x a point in Π; let $\overline{\Pi} = E \setminus \Pi$. We have

$$P_\beta\bigl(\tau(\overline{\Pi}) \leq n \mid X_0 = x\bigr) \leq \min_{y \in \Pi} P_\beta\bigl(\tau(\overline{\Pi}) \leq n \mid X_0 = y\bigr)$$
$$+ \max_{y \in \Pi} P_\beta\bigl(X_{\tau(\overline{\Pi} \cup \{y\})} \neq y \mid X_0 = x\bigr).$$

Then we use the property of cycles, which states that we visit with a large probability any point before leaving; this proves that, for β large enough, we have

$$\min_{y \in \Pi} P_\beta\bigl(X_{\tau(\overline{\Pi} \cup \{y\})} = y \mid X_0 = x\bigr) \geq \tfrac{1}{2}.$$

We use the remark that

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

and that

$$\sum_{k=0}^{+\infty} P_\beta\bigl(\tau(\overline{\Pi}) > kn \mid X_0 = x\bigr) \leq \sum_{k=0}^{+\infty} \Bigl(\max_{y \in \Pi} P_\beta\bigl(\tau(\overline{\Pi}) > n \mid X_0 = y\bigr)\Bigr)^k$$
$$\leq \frac{1}{\min_{y \in \Pi} P_\beta\bigl(\tau(\overline{\Pi}) \leq n \mid X_0 = y\bigr)}$$

so that, using the expression of the expectation of the exit time out of Π 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

$$\min_{y \in \Pi} P_\beta\bigl(\tau(\overline{\Pi}) \leq n \mid X_0 = y\bigr) \leq \frac{n}{\max_{x \in \Pi} E_\beta\bigl(\tau(\overline{\Pi}) \mid X_0 = x\bigr)}$$
$$\leq nc \exp(-\beta H(\Pi)).$$

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

Using the Markov property, we have for any integer k,

$$P_\beta\big(\tau(\overline{\Pi}) > n \mid X_0 = x\big) \geq \Big(\min_{x \in \Pi} P_\beta\big(\tau(\overline{\Pi}) > k \mid X_0 = x\big)\Big)^{\lceil n/k \rceil}.$$

We end the proof by choosing $k = \lfloor (1/4c)\exp(\beta H(\Pi)) \rfloor$, so that for β large enough we have

$$\left\lceil \frac{n}{k} \right\rceil \log\left(\frac{1}{2} - kc\exp(-\beta H(\Pi))\right) \geq -(\log 4)(1 + 4cn\exp(-\beta H(\Pi))). \quad \square$$

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*

$$\beta(N) = \frac{1}{H_1(V)}\left(\log\frac{N}{d} - \log\left(\frac{\eta}{H_1(V)}\log\frac{N}{d} + 1\right)\right),$$

we have

$$\max_{x \in E} P_{\beta(N)}\big(U(X_N) \geq \eta \mid X_0 = x\big) \leq d\left(\frac{d\eta}{NH_1(V)}\log N\right)^{\eta/H_1(V)}.$$

An immediate consequence is that

$$\inf_{\beta > 0}\max_{x \in E} P_\beta\big(U(X_N) \geq \eta \mid X_0 = x\big) \leq d\left(\frac{d\eta}{NH_1(V)}\log N\right)^{\eta/H_1(V)}.$$

Furthermore, there exists another constant $d' > 0$, depending only on a and E, there exists $\eta_0 \in (U(E) \setminus \{0\})$ (depending also on V), such that for any $0 < \eta \leq \eta_0$, $\eta \in U(E)$, for N large enough, we have

$$\inf_{\beta > 0}\max_{x \in E} P_\beta\big(U(X_N) \geq \eta \mid X_0 = x\big) \geq d'\left(\frac{d'\eta}{NH_1(V)}\log N\right)^{\eta/H_1(V)}.$$

So, for the best temperature and the worst initial point of the Metropolis dynamic, the speed of convergence is at best of order $((1/N)\log N)^{\eta/H_1(V)}$. When $\eta \in U(E)$ is small enough, the convergence speed obtained with $\beta(N)$ is almost optimal for the large deviation criterion

$$\lim_{N \to +\infty} -\frac{1}{\log N}\log\max_{x \in E} P_{\beta(N)}\big(U(X_N) \geq \eta \mid X_0 = x\big) = \frac{\eta}{H_1(V)}.$$

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 β. □

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 $(\lambda_k)_{k=0,\dots,r-1}$, $(\eta_k)_{k=0,\dots,r-1}$, any increasing sequence $(\gamma_k)_{k=0,\dots,r-1}$, such that*:

(i) $\eta_{r-1} \le \eta$;
(ii) $\lambda_k \ge (1+D)\eta_{k-1}$ *for all $k = 1,\dots,r$, 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

$$\beta_n^N = \gamma_k, \qquad k\frac{N}{r} < n \le (k+1)\frac{N}{r}, \qquad k = 0,\dots,r-1,$$

satisfies

$$\begin{aligned}
&P_{(\beta^N)}(U(X_N) \ge \eta \mid X_0 = x)\\
&\quad\le \left(1 + \frac{cN}{r}\exp(-\gamma_1(\lambda_1 - \eta_0))\right)\\
&\qquad\times\left(\exp\left(-\left\lfloor \frac{N}{rc}\exp(-\gamma_0 H_1(V))\right\rfloor\right) + c\exp(-\gamma_0\eta_0)\right)\\
&\qquad+ \sum_{k=2}^{r-1}\left(1 + c\frac{N}{r}\exp(-\gamma_k(\gamma_k - \eta_{k-1}))\right)\\
&\qquad\qquad\times\Bigg(\exp\left(-\left\lfloor \frac{N}{rc}\exp\left(-\gamma_{k-1}\lambda_{k-1}(1+1/D)^{-1}\right)\right\rfloor\right)\\
&\qquad\qquad\qquad + c\exp(-\gamma_{k-1}\eta_{k-1})\Bigg)\\
&\qquad+ \exp\left(-\left\lfloor \frac{N}{rc}\exp\left(-\gamma_{r-1}\lambda_{r-1}(1+1/D)^{-1}\right)\right\rfloor\right) + c\exp(-\gamma_{r-1}\eta_{r-1}).
\end{aligned}$$

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, $\overline{H}$ and $\underline{D}$ satisfying $d \geq c$, $\overline{H} \geq H_1(V)$ and $\underline{D} \leq D(E, V)$, for any constant $0 < \eta \leq H_1(V)/D$, for any integers r, N such that r divides N and $N/r \geq d$, the generalized simulated annealing chain stopped at time N with cooling schedule*

$$\beta_n^N = \gamma_k, \qquad k\frac{N}{r} < n \leq (k+1)\frac{N}{r}, \qquad k = 0, \dots, r-1,$$

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

satisfies

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

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

(i) *If we choose $r = \underline{D}^{-1}\log(\overline{H}/\eta\underline{D})/\underline{D}\log N$, 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 $N/\log N$.*

(ii) *If we choose $r = \lceil\log(\overline{H}/\eta\underline{D})/\log(1+\alpha)\rceil$ independently of N* (*where $\alpha > 0$ is a small parameter*), *the probability of failure is at most of order* $(1/N)^{1/D(1+\alpha)}(\log N)^{(1+1/D)/(1+\alpha)}$.

REMARKS. (i) In order to get the optimal exponent of convergence, we need not know the exact value of D, $H_1(V)$ and c, but only some bounds for these quantities. This shows that this type of cooling schedule is robust. However, this is still only a theoretical result, since it says absolutely nothing about the way to estimate the constant c, and therefore to choose the constant d which is a parameter of the schedule and plays an important role in the upper bound given for the probability of failure. The equations show that the estimation of $\overline{H}$ can be very rough, since it is "killed" by a $1/r$ exponent. As for $\underline{D}$, since $D(E, V)$ is a supremum, it is easy to see that, in the case when $\underline{D} > D$, the theorem still holds with D replaced by $\underline{D}$.

(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 $1/D$ or $1/D(1+\alpha)$) does not depend on the precision η with which we want to get close to the minima of the energy U, whereas in the case of the Metropolis algorithm, this exponent is equal to $\eta/H_1(V)$ and tends to zero when η does.

In the case when we want to find out an exact ground state, simulated annealing will be asymptotically faster then the Metropolis algorithm when the level η being next to the ground state energy level (which we set to zero by convention) is such that $\eta < H_1(V)/D$. Indeed in this case, $\eta/H_1(V)$ is the optimal convergence exponent for Metropolis and

$$\left(\frac{1}{N}\right)^{1/D} \ll \left(\frac{1}{N}\right)^{\eta/H_1(V)}$$

when N is large.

THEOREM 6.3 (Universally robust cooling schedule). *With the same notations, 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*

$$\begin{aligned} \gamma_k &= \gamma_0\left(1 + (\log M)^{-1-\varepsilon}\right)^k, \\ r &= \left\lfloor (\log M)^{1+2\varepsilon} \right\rfloor, \\ N &= Mr, \end{aligned}$$

we have

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

and therefore

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

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 a somewhat theoretical result, since the minimal 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 PROPOSITION 6.1. Let us fix N and $d > 0$.
Let us put $\lambda_0 = +\infty$ and introduce the events

$$\begin{aligned} \mathscr{B}_k &= \left\{ U(X_n) + V(X_n, X_{n+1}) \leq \lambda_k,\ n \in \mathbb{N},\ k\frac{N}{r} \leq n < (k+1)\frac{N}{r} \right\}, \\ \mathscr{A}_k &= \mathscr{B}_k \cap \left\{ U\big(X_{(k+1)(N/r)}\big) < \eta_k \right\}. \end{aligned}$$

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

For any initial point x in E,

$$\begin{aligned}P_N\big(U(X_N) \geq \eta \mid X_0 = x\big) &\leq P_N\big(U(X_N) \geq \eta_{r-1} \mid X_0 = x\big)\\ &\leq 1 - P_N\left(\bigcap_{k=0}^{r-1} \mathscr{A}_k \mid X_0 = x\right)\\ &\leq \sum_{k=1}^{r-1} P_N\left(\bar{\mathscr{A}}_k \cap \bigcap_{l=0}^{k-1} \mathscr{A}_l \mid X_0 = x\right) + P_{\gamma_0}\big(\bar{\mathscr{A}}_0 \mid X_0 = x\big).\end{aligned}$$

Applying Corollary 3.4 to the finite generalized Metropolis chain $(X_n)_{0 \leq n \leq N/r}$, we know that there exists a constant c depending only on $|E|$ and a such that

$$P_{\gamma_0}\big(U(X_{N/r}) > \eta_0 \mid X_0 = x\big) \leq \exp\left(-\left\lfloor \frac{N}{rc} \exp(-\gamma_0 H_1(V))\right\rfloor\right) + ce^{-\gamma_0 \eta_0}.$$

Furthermore, for $k \geq 1$,

$$\begin{aligned}&P_N\big(\bar{\mathscr{A}}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\big)\\ &\quad = P_N\big(\bar{\mathscr{B}}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\big)\\ &\qquad + P_N\big(U(X_{(k+1)(N/r)}) \geq \eta_k, \mathscr{B}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\big).\end{aligned}$$

And for all $y \in E$,

$$\begin{aligned}&P_N\big(X_{(k+1)(N/r)} = y, \mathscr{B}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\big)\\ &\quad \leq \max_{z, U(z) < \eta_{k-1}} P_N\big(X_{(k+1)(N/r)} = y, \mathscr{B}_k \mid X_{k(N/r)} = z\big)\\ &\quad \leq \max_{z, U(z) < \eta_{k-1}} P_{\gamma_k}\left(X_{N/r} = y, \tau(\bar{C}_z) > \frac{N}{r} \mid X_0 = z\right),\end{aligned}$$

where C_z denotes the smallest cycle containing z such that $\lambda_k < H(C_z) + U(C_z)$, and $\tau(\bar{C}_z)$ 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 λ_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

$$P_{\gamma_k}\left(X_{N/r} = y, \tau(\overline{C}_z) > \frac{N}{r} \mid X_0 = z\right) \leq P(Y_{N/r} = y \mid Y_0 = z).$$

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

$$\begin{aligned}&P_N\left(X_{(k+1)(N/r)} = y, \mathscr{B}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\right)\\ &\quad\leq \max_{z, \tilde{U}(z) < \eta_{k-1}} \exp\left(-\left\lfloor \frac{N}{rc} \exp(-\gamma_k H_1(C_z, V))\right\rfloor\right) + c\exp(-\gamma_k U(y)).\end{aligned}$$

Furthermore, we see that $H_1(C_z, V) \leq (1 + 1/D)^{-1}\lambda_k$, 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 $H(C)(1 + 1/D) \leq H(C) + U(C) \leq \lambda_k$. Thus, bounding $H_1(C_z, V)$ in the previous inequality, we find

$$\begin{aligned}&P_N\left(X_{(k+1)(N/r)} = y, \mathscr{B}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\right)\\ &\quad\leq \exp\left(-\left\lfloor \frac{N}{rc} \exp\left(-\gamma_k \lambda_k \left(1 + \frac{1}{D}\right)^{-1}\right)\right\rfloor\right)\\ &\qquad + c\exp(-\gamma_k U(y)).\end{aligned}$$

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

$$\begin{aligned}&P_N\left(U(X_{(k+1)(N/r)}) \geq \eta_k, \mathscr{B}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\right)\\ &\quad\leq \exp\left(-\left\lfloor \frac{N}{rc} \exp\left(-\gamma_k \lambda_k \left(1 + \frac{1}{D}\right)^{-1}\right)\right\rfloor\right)\\ &\qquad + c\exp(-\gamma_k \eta_k).\end{aligned}$$

Furthermore,

$$\begin{aligned}&P_N\left(\overline{\mathscr{B}}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\right)\\ &\quad= \sum_{z, U(z) < \eta_{k-1}} P_N\left(\overline{\mathscr{B}}_k \mid X_{k(N/r)} = z\right)\\ &\qquad\qquad \times P_N\left(X_{k(N/r)} = z, \mathscr{A}_0, \dots, \mathscr{A}_{k-2}, \mathscr{B}_{k-1} \mid X_0 = x\right)\end{aligned}$$

and

$$
\begin{aligned}
&P_N\left(\overline{\mathscr{B}}_k \mid X_{k(N/r)} = z\right) \\
&\quad \le \sum_{n=k(N/r)}^{(k+1)(N/r)} P_{\gamma_k}\left(U(X_n) + V(X_n, X_{n+1}) > \lambda_k \mid X_{k(N/r)} = z\right) \\
&\quad = \sum_{n=k(N/r)}^{(k+1)(N/r)} \sum_{U(u)+V(u,v)>\lambda_k} P_{\gamma_k}\left(X_n = u \mid X_{k(N/r)} = z\right) p_{\gamma_k}(u,v) \\
&\quad \le \sum_{n=k(N/r)}^{(k+1)(N/r)} \sum_{U(u)+V(u,v)>\lambda_k} \frac{\mu_{\gamma_k}(u)}{\mu_{\gamma_k}(z)} p_{\gamma_k}(u,v).
\end{aligned}
$$

Thus, applying inequalities (1) and (2) 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

$$P_{\gamma_k}\left(\overline{\mathscr{B}}_k \mid X_{k(N/r)} = z\right) \le c\frac{N}{r}\exp\left(-\gamma_k\left(\lambda_k - U(z)\right)\right).$$

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

$$
\begin{aligned}
&P_{\gamma_k}\left(\overline{\mathscr{B}}_k \mid X_{k(N/r)} = z\right) P_N\left(X_{k(N/r)} = z, \mathscr{A}_0, \dots, \mathscr{A}_{k-2}, \mathscr{B}_{k-1} \mid X_0 = x\right) \\
&\quad \le c\frac{N}{r}\exp\left(-\gamma_k\left(\lambda_k - U(z)\right)\right)\left\{\exp\left(-\left[\frac{N}{rc}\exp\left(-\gamma_{k-1}\lambda_{k-1}\left(1+\frac{1}{D}\right)^{-1}\right)\right]\right)\right. \\
&\qquad\qquad \left. + c\exp\left(-\gamma_{k-1}U(z)\right)\right\} \\
&\quad \le c\frac{N}{r}\exp\left(-\gamma_k\left(\lambda_k - \eta_{k-1}\right)\right)\left\{\exp\left(-\left[\frac{N}{rc}\exp\left(-\gamma_{k-1}\lambda_{k-1}\left(1+\frac{1}{D}\right)^{-1}\right)\right]\right)\right. \\
&\qquad\qquad \left. + c\exp\left(-\gamma_{k-1}\eta_{k-1}\right)\right\}.
\end{aligned}
$$

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

$$
\begin{aligned}
&P_{\gamma_k}\left(\overline{\mathscr{B}}_k, \mathscr{A}_0, \dots, \mathscr{A}_{k-1} \mid X_0 = x\right) \\
&\quad \le c\frac{N}{r}\exp\left(-\gamma_k\left(\lambda_k - \eta_{k-1}\right)\right)\left(\exp\left(-\left[\frac{N}{rc}\exp\left(-\gamma_{k-1}\lambda_{k-1}\left(1+\frac{1}{D}\right)^{-1}\right)\right]\right)\right. \\
&\qquad\qquad \left. + c\exp\left(-\gamma_{k-1}\eta_{k-1}\right)\right).
\end{aligned}
$$

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

$$\begin{aligned}
&P_N(U(X_N) \geq \eta \mid X_0 = x)\\
&\quad \leq \sum_{k=1}^{r-1} \Biggl\{ c\,\frac{N}{r} \exp(-\gamma_k(\lambda_k - \eta_{k-1}))\\
&\qquad\qquad \times \Biggl(\exp\Biggl(- \biggl\lfloor \frac{N}{rc} \exp\biggl(-\gamma_{k-1}\lambda_{k-1}\Bigl(1 + \frac{1}{D}\Bigr)^{-1}\biggr)\biggr\rfloor\Biggr)\\
&\qquad\qquad\quad + c\exp(-\gamma_{k-1}\eta_{k-1})\Biggr)\\
&\qquad\qquad + \exp\Biggl(- \biggl\lfloor \frac{N}{rc} \exp\biggl(-\gamma_k\,\lambda_k\Bigl(1 + \frac{1}{D}\Bigr)^{-1}\biggr)\biggr\rfloor\Biggr) + c\exp(-\gamma_k\eta_k)\Biggr\}\\
&\qquad + \exp\biggl(- \biggl\lfloor \frac{N}{rc} \exp(-\gamma_0 H_1(V))\biggr\rfloor\biggr) + c\exp(-\gamma_0\eta_0). \qquad \square
\end{aligned}$$

Proof of Theorem 6.2. Let us define, in order to simplify formulas,

$$\begin{aligned}
\alpha &= \frac{1}{\rho D}\biggl(\log\frac{N}{rd} - (1 + D)\log\biggl(1 + \frac{1}{\underline{D}}\log\frac{N}{rd}\biggr)\biggr),\\
\rho &= \biggl(\frac{\overline{H}}{\eta\underline{D}}\biggr)^{1/r}
\end{aligned}$$

and put

$$\begin{aligned}
\eta_k &= \frac{\alpha}{\gamma_k},\\
\lambda_0 &= +\infty,\\
\lambda_k &= \frac{1}{\gamma_k}\biggl(1 + \frac{1}{D}\biggr)\biggl(\log\frac{N}{rd} - \log(1 + \alpha)\biggr).
\end{aligned}$$

Note that $\rho > 1$ and that $\alpha \leq (1/\underline{D})\log(N/rd)$ since we chose η small enough such that $H_1(V)/\eta \geq D$. Indeed, we have

$$\alpha \leq \frac{1}{\rho D}\log\frac{N}{rd} \leq \frac{1}{\underline{D}}\log\frac{N}{rd}.$$

Note also that $\lambda_k \geq (1 + D)\eta_{k-1} \geq \eta_{k-1}$, since $\alpha \leq (1/\underline{D})\log(N/rd)$ and that $(\eta_k)_{k \leq r-1}$ and $(\lambda_k)_{k \leq r-1}$ are geometrical decreasing sequences.

Moreover, $\eta_{r-1} \le \eta$, since

$$\eta_{r-1} = \frac{\alpha}{\gamma_0}\rho^{1-r} = \rho^{-r}\frac{1/D\big(\log(N/rd) - (1+D)\log(1+(1/\underline{D})\log(N/rd))\big)}{(1/\overline{H})\big(\log(N/rd) - \log(1+(1/\underline{D})\log(N/rd))\big)}$$
$$\le \frac{\overline{H}}{\underline{D}}\rho^{-r} \le \eta,$$

thus conditions (i) and (ii) given in Proposition 6.1 are satisfied.

Coming back to the definitions of η_k and λ_k, we check that for any $d > c$,

$$\frac{N}{r}\exp(-\gamma_k(\lambda_k - \eta_{k-1})) = d(1+\alpha)^{1+1/D}\left(\frac{N}{rd}\right)^{-1/D}\exp(\alpha\rho)$$
$$= d\left(\frac{1+\alpha}{1+(1/\underline{D})\log(N/rd)}\right)^{1+1/D}.$$

Using the inequality $\alpha \le (1/\underline{D})\log(N/rd)$, we find that

$$\frac{N}{r}\exp(-\gamma_k(\lambda_k - \eta_{k-1})) \le d.$$

We also have

$$\exp\left(-\left\lfloor\frac{N}{rc}\exp\left(-\gamma_k\lambda_k\left(1+\frac{1}{D}\right)^{-1}\right)\right\rfloor\right)$$
$$\le \exp\left(-\frac{N}{rd}\exp\left(-\gamma_k\lambda_k(1+1/D)^{-1}\right)+1\right)$$
$$\le \exp(-\alpha)$$

and

$$\exp\left(-\left\lfloor\frac{N}{rc}\exp(-\gamma_0 H_1(V))\right\rfloor\right)$$
$$\le \exp\left(-\frac{N}{rd}\exp(-\gamma_0 H_1(V))+1\right)$$
$$\le \exp\left[-\frac{N}{rd}\exp\left(-\log\frac{N}{rd} + \log\left(1+\frac{1}{\underline{D}}\log\frac{N}{rd}\right)\right)+1\right]$$
$$\le \exp\left[-\frac{1}{\underline{D}}\log\frac{N}{rd}\right] \le \exp(-\alpha).$$

Finally, $\exp(-\gamma_k\eta_k) = \exp(-\alpha)$. Thus, $P_N(U(X_N) \ge \eta \mid X_0 = x) \le r(1 + cd)(1+c)e^{-\alpha}$, 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 γ_0. Put

$$\begin{aligned} \rho &= 1 + (\log M)^{-1-\varepsilon}, \\ \gamma_k &= \gamma_0 \rho^k, \\ r &= \lfloor (\log M)^{1+2\varepsilon} \rfloor, \\ N &= Mr, \end{aligned}$$

with parameter $\varepsilon > 0$, and put also

$$\begin{aligned} \lambda_0 &= +\infty, \\ \lambda_k &= \lambda / \rho^k, \\ \eta_k &= \alpha / \rho^k, \end{aligned}$$

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 conditions (i) and (ii) of Proposition 6.2. Condition (i), $\eta_{r-1} \le \eta$, which can be also written as $\alpha \le \eta\rho^{r-1}$, is satisfied for large enough M since we have $\rho^r = O(M)$. Condition (ii), $\lambda_k \ge (1 + D)\eta_{k-1}$, which can also be written as $\lambda \ge (1 + D)\rho\alpha$, is also satisfied. We end the proof by substituting the values of γ_k, λ_k and η_k into Proposition 6.1. □

7. Conclusion. We showed in this paper how to tune piecewise constant cooling schedules to make the logarithm of the probability of failure after N iterations have the optimal asymptotic equivalent when N tends to infinity. Although this asymptotic performance can be achieved by a fixed "robust" triangular cooling schedule chosen independently of the minimization problem to be solved, our results should not disguise the fact that in practice and for a given N and a given energy landscape, a suitable choice of the three parameters $T_{\min}(N)$, $T_{\max}(N)$ and $r(N)$ in the general formula

$$T_n^N = T_{\max}(N)\left(\frac{T_{\min}(N)}{T_{\max}(N)}\right)^{(k-1)/(r(N)-1)}, \qquad (k-1)\frac{N}{r(N)} \le n < k\frac{N}{r(N)},$$

can save a lot of computer time.

Beyond the theoretical properties we proved here, the strength of this widely used type of schedule comes from the fact that it has a simple parametric form depending on three parameters only. Therefore an interesting direction of research would be to estimate efficient values for $T_{\min}(N)$, $T_{\max}(N)$ and $r(N)$ from independent trials of length N, using some kind of stochastic gradient update rule. This makes sense in practice, since the use of multiple independent trials is known to be efficient in this context, even when an optimal choice of the temperature schedule is known from the beginning (see [1], and also [7] and [9] for a discussion of repeated optimization schemes). Another approach to adaptive tuning of the temperature, interpreted as an energy transformation, can be found in [7].

Of interest also is the fact that good performances can be achieved with a choice of $r(N)$ independent of N (Theorem 6.2), and that nearly optimal ones can be reached with a slowly increasing choice of $r(N)$. This means that each constant temperature step will be large when N is large, and is of practical importance when the number of values taken by the rate function $V(x, y)$ is small. Indeed, in this case, the values of $e^{-\beta V(x,y)}$ can be tabulated once for each step, and this can save a lot of time in the computation of the transition probabilities. This favourable situation is encountered with the noisy Ising model (see [2]), and more generally with Potts' models, widely used in image analysis applications.

Acknowledgment. We are glad to acknowledge the useful remarks of one of our referees on a first draft of this paper.

REFERENCES

[1] AZENCOTT, R. (1992). Sequential simulated annealing speed of convergence and acceleration techniques. In *Simulated Annealing: Parallelization Techniques* (R. Azencott, ed.) 1–10. Wiley, New York.

[2] CATONI, O. (1990). Image restoration by stochastic dichotomic reconstruction of contour lines. *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis. Lecture Notes in Statist.* **74** 101–116. Springer, Berlin.

[3] CATONI, O. (1991). Applications of sharp large deviation estimates to optimal cooling schedules. *Ann. Inst. H. Poincaré* **27** 463–518.

[4] CATONI, O. (1991). Sharp large deviation estimates for simulated annealing algorithms. *Ann. Inst. H. Poincaré* **27** 291–383.

[5] CATONI, O. (1991). Exponential triangular cooling schedules for simulated annealing algorithms: a case study. *Applied Stochastic Analysis, Proceedings of a US-French Workshop. Lecture Notes in Control and Inform. Sci.* **177**. Springer, Berlin.

[6] CATONI, O. (1992). Rough large deviation estimates for simulated annealing: application to exponential schedules. *Ann. Probab.* **20** 1109–1146.

[7] CATONI, O. (1994). The energy transformation method for the Metropolis algorithm compared with simulated annealing. *Probab. Theory Related Fields* **110** 69–89.

[8] CATONI, O. (1995). Algorithmes de recuit simulé et chaînes de Markov á transitions rares. Notes de cours de DEA, Univ. Paris XI, Orsay. (English translation: Simulated annealing algorithms and Markov chains with rare transitions. Preprint. LMENS 97-09. *Séminaire de Probabilities*. Available at http://www.dmi.ens.fr/preprints.)

[9] CATONI, O. (1996). Solving scheduling problems by simulated annealing. Preprint. LMENS 96-20. *SIAM J. Control Optim.* To appear. Available at http://www.dmi.ens.fr/dmi/preprints.

[10] CATONI, O. (1996). Metropolis, simulated annealing and I.E.T. algorithms: theory and experiments. *J. Complexity* **12** 595–623 [correction (1997) **13** 384].

[11] CATONI, O. and CERF, R. (1997). The exit path of a Markov chain with rare transitions. *ESAIM: Probab. Statist.* **1** 95–144. Available at http://www.emath.fr/Math/Ps/ps.html.

[12] CHIANG, T. S. and CHOW, Y. (1989). A limit theorem for a class of inhomogeneous Markov processes. *Ann. Probab.* **17** 1483–1502.

[13] DEUSCHEL, J. D. and MAZZA, C. (1994). L^2 convergence of time nonhomogeneous Markov processes I. Spectral estimates. *Ann. Appl. Probab.* **4** 1012–1056.

[14] DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61.

[15] DUFLO, M. (1996). *Algorithmes Stochastiques*. Springer, Berlin.

[16] FREIDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbation of Dynamical Systems*. Springer, New York.

[17] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 721–741.

[18] GÖTZE, F. (1991). Rate of convergence of simulated annealing processes. Preprint.

[19] HAJEK, B. (1988). Cooling schedule for optimal annealing. *Math. Oper. Res.* **13** 311–329.

[20] HOLLEY, R. and STROOCK, D. (1988). Simulated annealing via Sobolev inequalities. *Comm. Math. Phys.* **115** 553–569.

[21] HWANG, C. R. and SHEU, S. J. (1992). Singular perturbed Markov chains and exact behaviors of simulated annealing process. *J. Theoret. Probab.* **5** 223–249.

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

[23] MICLO, L. (1991). Evolution de l'énergie libre. Applications à l'étude de la convergence des algorithmes de recuit simulé. Ph.D. dissertation, Univ. Paris 6.

[24] MICLO, L. (1995). Sur les temps d'occupations des processus de Markov finis inhomogènes à basse température. Preprint.

[25] MICLO, L. (1996). Sur les problèmes de sortie discrets inhomogènes. *Ann. Appl. Probab.* **6** 1112–1156.

[26] SALOFF-COSTE, L. (1996). Lectures on finite Markov chains. *Ecole d'été de Probabilités de Saint-Flour XXVI. Lecture Notes in Math.* **1665** 301–413. Springer, Berlin.

[27] TROUVÉ, A. (1993). Parallelisation massive du recuit simulé. Thèse de doctorat, Univ. Paris 11.

[28] TROUVÉ, A. (1996). Cycle decompositions and simulated annealing. *SIAM J. Control Optim.* **34** 966–986.

[29] TROUVÉ, A. (1996). Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms. *Ann. Inst. H. Poincaré* **32** 299–348.

[30] TSITSIKLIS, J. N. (1989). Markov chains with rare transitions and simulated annealing. *Math. Oper. Res.* **14** 70–90.

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
BATÎMENT 425
91405 ORSAY CEDEX
FRANCE
E-MAIL: cot@dmi.ens.fr
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: catoni@dmi.ens.fr.