Astérisque

ROBERT AZENCOTT

Simulated annealing

Astérisque, tome 161-162 (1988), Séminaire Bourbaki,
exp. n° 697, p. 223-237

<http://www.numdam.org/item?id=SB_1987-1988__30__223_0>

Séminaire BOURBAKI Juin 1988
40ème année, 1987-88, n° 697

SIMULATED ANNEALING

par Robert AZENCOTT

1. LARGE SCALE OPTIMIZATION PROBLEMS AND SPIN-GLASS MODELS

Consider a finite set S of indices and for each $s \in S$ a variable x_s taking its values in a finite set L. Call $E = L^S$ the set of all "configurations" $x = (x_s)_{s \in S}$. Now let $H : E \longrightarrow \mathbb{R}$ be an arbitrary function ; the problem we consider here is to evaluate $H_{min} = \min_{x \in E} H(x)$, and to find at least one configuration x minimizing $H(x)$.

When the cardinal of S is small, a simple enumeration of the $x \in E$ would be a feasible algorithm, but minimization problems for which the set S of variables has *very large cardinal* are quite common in statistical mechanics, combinatorial optimization, image analysis, etc.

One instance of such large scale optimization problems was provided by the *spin-glass models in statistical mechanics* ; in this context, x_s represents the physical state of the vertex s in a crystal lattice S of very large cardinal $N = \mathrm{car}(S)$. The lattice is imbedded in a two or three dimensional euclidean space, and $H(x)$ represents the energy of the configuration x.

For the spin-glass model, the energy will typically be of the form

$$H(x) = \sum_{K \in C} U_K(x) ,$$

where C, the set of "*cliques*", is the family of all subsets K in the lattice S such that diameter $(K) \leq \rho$, and where each action potential $U_K(x)$ depends only on the $(x_s)_{s \in K}$.

No direct evaluation of H_{min} is feasible since, in the spin-glass situation, which is supposed to modelize crystals mixed with randomly scattered impurities, the map $K \longrightarrow U_K$ is assumed to assign "at random" to each clique K an action potential U_K belonging to a fixed vector space of real valued functions.

For particularly simple interactions, the asymptotic behaviour of Average(H_{min}) as $N \longrightarrow \infty$ has been obtained by *the replica method* (Parisi, Mezard) in the physics literature, which gave also rough descriptions of the "ground states"

(minimizing configurations) for large N. Due to the large part of heuristics in these computations, the empirical verification of the results was crucial, so that spin-glass specialists such as Kirkpatrick, Toulouse, Mezard, Gutfreund, Amit and many others have been natural customers for feasible minimization algorithms such as the now celebrated "simulated annealing" method.

One of the tricky features in typical spin-glass models is that the number of "local minima" for $H(x)$ is huge for large N, so that any kind of deterministic gradient algorithm is useless in such situations. Here the notion of local minima refers to the so-called *Hamming distance*, two configurations x and y being neighbours if and only if they differ at most at a single site.

2. THE GIBBS DISTRIBUTION

In the context of statistical mechanics, it makes sense to consider on any given configuration space $E = L^S$, the set M_a of probability distributions P for which the average energy is constrained by

$$\sum_{x \in E} P(x)\, H(x) = a \,. \tag{2.1}$$

Within the set M_a, the so-called "most disorderly" distribution P must maximize the *entropy*

$$- \sum_{x \in E} P(x) \log P(x)$$

and hence, using Lagrange multipliers, one readily sees that the most disorderly distributions in M_a are *the Gibbs distributions*

$$G_T(x) = \frac{1}{Z_T} \exp\left[-\frac{H(x)}{T}\right]$$

where $Z_T = \sum_{x \in E} \exp\left[-\frac{H(x)}{T}\right]$ is the "*partition function*" and the positive parameter T, determined by (2.1) is usually identified with a "*temperature*".

An obvious feature of G_T which is crucial here is that, as $T \longrightarrow 0$, the distribution G_T becomes more and more concentrated on the set

$$E_{MIN} = \{x \in E \mid H(x) = \min_{y \in E} H(y)\} \,. \tag{2.2}$$

More precisely, we have

$$\lim_{T \to 0} G_T(x) = G_O(x) \quad \text{for all } x \in E \,,$$

where G_O is *the uniform probability distribution on* E_{MIN}. Thus, if one could select *effectively* a random configuration $x \in E$ with probability distribution G_T, *and* T *small enough*, such a configuration should be, in an overwhelming proportion of cases, an almost minimizing configuration for the energy H.

However, for large cardinal(S), building up an effective method of random sampling in E with respect to the probability G_T turns out to be a serious

computational problem, which was initially solved by Glauber, Metropolis and altri.

3. THE GLAUBER DYNAMICS

We sketch the now classical stochastic algorithm suggested by Glauber, to construct a random variable X with values in the space of configurations E and with probability distribution close to G_T. Note that here the temperature T is *fixed*.

First we fix an *arbitrary symetric* Markov transition matrix $Q = (q_{xy})$ where $x,y \in E$. For instance, a typical choice for the spin-glass model is to set

$$(3.1) \qquad \begin{cases} q_{xy} = 0 & \text{if } y \notin \{V_x - x\} \\ q_{xy} = \frac{1}{r} & \text{if } y \in \{V_x - x\} \end{cases}$$

where V_x is a set of neighbours of x in E for the Hamming distance (cf. § 1), and $(1+r)$ is the common cardinal of all the V_x, $x \in E$.

We want to construct a *random* sequence X_n of configurations. Assume the configuration X_n already obtained. Select then a random configuration Y_n such that

$$(3.2) \qquad P(Y_n = y \mid X_0,\ldots,X_n) = q_{X_n y}$$

We then impose, with probability one,

$$(3.3) \qquad \{\text{either } X_{n+1} = Y_n \text{ or } X_{n+1} = X_n\} .$$

This (random) choice is made according to the following rule

$$(3.4) \qquad P(X_{n+1} = Y_n \mid X_0 \ldots X_n, Y_n) = \exp - \frac{[H(Y_n) - H(X_n)]^+}{T} ,$$

where $[v]^+ = v$ if $v \geq 0$ and $[v]^+ = 0$ if $v \leq 0$.

It is quite easy to prove that *the Markov chain* X_n *has a unique equilibrium distribution, coinciding with the Gibbs distribution* G_T, provided the "selection" matrix Q is symmetric and irreducible. In other words, one has then

$$(3.5) \qquad \lim_{n \to +\infty} P(X_n = x) = G_T(x) \quad , \quad x \in E .$$

Here irreducibility of Q means that any two configurations can be connected by a finite chain of configurations x_i such that $q_{x_i x_{i+1}} > 0$. Of course, the actual computation of X_{n+1} given X_n becomes lengthier when the cardinal of $\{y \in E \mid q_{xy} > 0\}$ increases.

4. SIMULATED ANNEALING : HEURISTICS

The result (3.5) and the fact that, for low temperature T, the distribution G_T concentrates on minimizing configurations suggest the use of Glauber random dynamics where the temperature T is no longer constant but decreases to 0 as $n \longrightarrow +\infty$. If we fix a decreasing sequence T_n such that $\lim_{n \to +\infty} T_n = 0$, and if we replace T by T_n in the conditional distribution (3.4), we obtain a new Markov chain X_n, for which one should *hopefully* have

$$\lim_{n \to +\infty} P(X_n \in E_{MIN}) = 1 \ . \tag{3.6}$$

This new algorithm was called *simulated annealing* by Kirkpatrick, Gelatt, Vecchi who introduced the idea.

Their (heuristic) arguments were based on a formal analogy with progressive and *very slow* physical cooling, which had long been used to bring actual physical systems into stable low-energy states, while fast cooling would freeze the system in undesirable high energy metastable states.

Actually not all *cooling schedules* (T_n) will exhibit the crucial minimizing feature (3.6). The first sufficient condition for (3.6) was obtained by D. and S. Geman, who proved rigorously that if

$$\lim_{n \to +\infty} T_n \log n > R \ , \tag{3.7}$$

with $R > 0$ *large enough*, then the minimizing property (3.6) would hold for the simulated annealing algorithm.

Easily built counterexamples were quickly exhibited (by Brétagnolle for instance) to show that when $\lim_{n \to +\infty} T_n \log n = 0$, the minimizing property (3.6) cannot hold in general. Then Hajek computed the value of the best constant R in (3.7), and several mathematical papers have since refined the asymptotic study of these algorithms, and of their continuous time analogues. Let us mention a few names : Holley - Stroock, Föllmer, Gidas, Hwang - Sheu, Chiang - Chow, Catoni, Trouvé and many more.

Simultaneously, practical uses of simulated annealing for large scale minimization problems have been explored by a very large community of physicists and/or computer science specialists such as Sherrington, Toulouse, Dreyfus, Aarts - Laarhoven, Bonomi - Lutton, Uhry, D. and S. Geman, etc.

5. SIMULATED ANNEALING : THE ABSTRACT SETUP

Consider an abstract finite set E, which will still be called the *configuration space*. Let $H : E \longrightarrow \mathbb{R}$ be an arbitrary function, still called the *energy function*.

Fix a symmetric stochastic matrix $Q = (q_{xy})$, $x \in E$, $y \in E$ such that any two configurations x and y can be connected by a finite chain $x_k \in E$ with $q_{x_k x_{k+1}} > 0$. The matrix Q will be called the *exploration matrix*.

Fix a *decreasing sequence* T_n *of "temperatures"* tending to zero as $n \longrightarrow +\infty$. This sequence will be called a *cooling schedule*.

On the state space E define a (non homogeneous) Markov chain X_n , with arbitrary initial state X_n , and transition matrix

$$p_n(x,y) = P(X_{n+1} = y \mid X_n = x)$$

given by

$$(5.1) \qquad \begin{cases} p_n(x,y) = q_{xy} \exp - \dfrac{[H(y) - H(x)]^+}{T_n} & \text{for } y \neq x \\ p_n(x,x) = 1 - \displaystyle\sum_{y \neq x} p_n(x,y) \ . \end{cases}$$

Before stating the main asymptotic results concerning the simulated annealing algorithm (X_n) , we need a few more definitions.

For each $x \in E$, let V_x be the set of all points $y \in E$ such that $q_{xy} > 0$, which will be called *the set of neighbours* of x .

A point $x \in E$ is said to be *a local minimum of* H if $H(x) \leq H(y)$ for all $y \in V_x$. It is called *a global minimum of* H if $H(x) \leq H(y)$ for all $y \in E$. We denote by E_{MIN} the set of global minima and E_{LOCMIN} the set of local minima of H .

Introduce now several notions used by Hajek.

Two states x and y in E are said to *communicate at height* h if either $y = x$ and $H(x) \leq h$, or if there is a sequence $x_1 \ldots x_k$, $k \geq 2$ with $x_1 = x$, $x_k = y$ and such that $\{H(x_j) \leq h \ , \ x_{j+1} \in V_{x_j}\}$ for all $j = 1 \ldots k$. Note that this property is *symmetric* in (x,y) .

The *depth* d_x of a local minimum $x \in E$ will be the smallest number $D > 0$ for which there exists a $y \in E$ such that x and y communicate at height $H(x) + D$, and $H(y) < H(x)$.

Note that $d_x = +\infty$ whenever x is a global minimum.

5.2. THEOREM (Hajek).- *Consider an arbitrary energy function* $H : E \longrightarrow \mathbb{R}$ *and the simulated annealing algorithm* (5.1). *Then one has*

$$(5.3) \qquad \lim_{n \to +\infty} P(X_n \in E_{MIN}) = 1$$

if and only if

$$(5.4) \qquad \sum_{n=1}^{+\infty} \exp\left(-\frac{D}{T_n}\right) = +\infty$$

where the constant D *is given by*

$$(5.5) \qquad D = \sup \{d_x \mid x \in [E_{LOCMIN} - E_{MIN}]\} \ .$$

This elegant result clearly implies that sequences T_n such that $\lim_{n \to +\infty} T_n \log n = c$ will be "minimizing" cooling schedules if and only if $c \geq D$, a result which seriously improved on previous sufficient conditions given by D. and S. Geman, as well as Gidas.

A recent theorem of Chiang and Chow completes nicely Hajek's result.

5.6. THEOREM (Chiang and Chow).- *Consider the annealing algorithm* (5.1). *For any two distinct global minima* $x,y \in E_{MIN}$ *let* h_{xy} *be the smallest height at which* x *and* y *communicate. Define the constants*

$$\bar{R} = \sup \{h_{xy} \mid x,y \in E_{MIN} , x \neq y\} \tag{5.7}$$

$$R = \max(\bar{R},D) ,$$

where D *is given by* (5.5).

Then the property

$$\begin{cases} \lim_{n \to +\infty} P(X_n = x) = 0 & \text{for all } x \notin E_{MIN} \\ \lim_{n \to +\infty} P(X_n = x) > 0 & \text{for all } x \in E_{MIN} \end{cases} \tag{5.8}$$

holds if and only if

$$\sum_{n=1}^{+\infty} \exp\left(-\frac{R}{T_n}\right) = +\infty . \tag{5.9}$$

The methods of proof used by Hajek, Chiang and Chow, and more recently by Catoni who obtained technically better estimates, all rely implicitly or explicitly on large deviations ideas introduced by Freidlin and Wentzell in the study of invariant measures for small diffusions. The same source of inspiration underlies the approach of Hwang and Sheu for the asymptotics of the so-called Langevin equation.

Due to lack of space, instead of reporting on the quite technical proofs of all these authors, we prefer to propose a much quicker approach *which will be less general* but gives pertinent and easily reached clues ; these computations will be sketched informally but can be formalized at very low cost and do provide a useful tool to understand quickly new variants of the simulated annealing algorithms.

6. ASYMPTOTIC RESULTS OF FREIDLIN - WENTZELL

On a finite state space E , consider a stochastic transition matrix, depending on the parameter $T > 0$,

$$P_T = [p_{xy}(T)]_{x \in E, y \in E} \tag{6.0}$$

where for $x \neq y$

$$p_{xy}(T) = a_{xy} \exp\left[-\frac{1}{T} U_{xy}\right] . \tag{6.1}$$

Here the U_{xy} , a_{xy} are arbitrary numbers in $\mathbb{R}^+$, and the parameter T will tend to zero.

Assume P_T to be irreducible, so that there is a unique invariant probability measure μ_T on E verifying $\mu_T P_T = \mu_T$.

Let λ_T be the eigenvalue of P_T which has the largest modulus, among all eigenvalues distinct from 1 . Wentzell and Freidlin have proved two interesting results concerning the asymptotic behaviour of μ_T and λ_T as $T \longrightarrow 0$.

To state these results, introduce for any subset F of E a particular set of graphs $S(F)$ with vertices in E . By definition, a graph $G \in S(F)$ will be a set of arrows $f : x \longrightarrow y$ where $x,y \in E$, $x \neq y$, and such that

(6.2)
- G contains no cycle ;
- for each $x \in E-F$, G contains a unique arrow starting at x ;
- for all $x \in F$, the graph G contains no arrow starting at x .

For any arrow $f : x \longrightarrow y$, we let $U(f) = U_{xy}$, and one defines then *the cost* $U(G)$ *of any graph* G in $S(F)$ by

$$U(G) = \sum_{f \in G} U(f) . \tag{6.3}$$

Following Wentzell, we now define, for $k = 1,2,\ldots,\mathrm{card}(E)$, the numbers

$$C_k = \inf \{U(G) \mid G \in S(F) , F \subset E , \mathrm{card}(F) = k\} . \tag{6.4}$$

Wentzell proved that, if $N = \mathrm{card}(E)$

$$C_1 \geq C_2 \geq \ldots \geq C_N = 0 \tag{6.5}$$

$$C_1 - C_2 \geq C_2 - C_3 \geq \ldots \geq C_{N-1} - C_N , \tag{6.6}$$

and that *in the generic situation where the inequalities* (6.6) *are strict,* then for T small, the eigenvalues of P_T are distinct and real, and the $(1+k)^{\text{th}}$ eigenvalue (in decreasing order) $\theta_k(T)$ verifies

$$\lim_{T \to 0} T \log [1 - \theta_k(T)] = - [C_k - C_{k+1}] . \tag{6.7}$$

In particular, one can improve slightly Wentzell's result to show that the 2^{nd} eigenvalue λ_T verifies then

$$1 - \lambda_T \sim c \exp\left(-\frac{A}{T}\right) \tag{6.7}$$

where $A = C_1 - C_2 > 0$, and $c > 0$.

7. STEPWISE COOLING

To gain some heuristic insights, we are going to study cooling schedules where one keeps the temperature T_n fixed during a series of K_n consecutive steps, before lowering it to T_{n+1}. Such cooling schedules are quite common in applications.

Let P_n be one-step transition matrix at temperature T_n defined by the annealing scheme (5.1) associated with the energy function $H(x)$. Of course, with the notations (6.0)-(6.1), we have $P_n = P_{T_n}$ with $U_{xy} = [H(y) - H(x)]^+$.

Let ν_0 be the probability distribution of X_0. The probability distribution ν_n of X_{L_n} where

$$L_n = K_1 + \ldots + K_n \tag{7.1}$$

is given by the recurrence relation

$$\nu_n = \nu_{n-1} P_n^{K_n} . \tag{7.2}$$

Call Λ_n the set of distinct eigenvalues of P_n which are not equal to 1. Then the standard Jordan decomposition of P_n yields

$$P_n = M_n + \sum_{\lambda \in \Lambda_n} \lambda \, Q_{\lambda,n} \tag{7.3}$$

where, since P_n is an irreducible stochastic matrix, *all the rows of* M_n *must coincide with the invariant measure* μ_n of P_n, and

$$\mu_n \, Q_{\lambda,n} = 0 \quad \text{for all} \quad \lambda \in \Lambda_n . \tag{7.4}$$

In particular, the rows of M_n being identical, for any measure μ on E, we have the implication

$$\{ \sum_{x \in E} \mu(x) = 0 \quad \text{implies} \quad \mu M_n = 0 \} . \tag{7.5}$$

Let λ_{T_n} be the eigenvalue with largest modulus in the set Λ_n ; the estimates of § 6 show that, at least in the generic case,

$$|\lambda| \leq 1 - c \exp\left(-\frac{A}{T_n}\right) \quad \text{for all} \quad \lambda \in \Lambda_n , \tag{7.6}$$

where $A = C_1 - C_2 > 0$, $c > 0$.

One can prove also that

$$\| Q_{\lambda,n} \| \leq a \quad \text{for all} \quad \lambda \in \Lambda_n , \text{ all } T_n , \tag{7.7}$$

where a is a fixed constant. Assume P_n diagonalisable, to write a shorter proof.

Hence, for any measure μ on E *such that* $\sum_{x \in E} \mu(x) = 0$, we have

$$\mu \, P_n^{K_n} = \mu \left[M_n + \sum_{\lambda \in \Lambda_n} \lambda^{K_n} Q_{\lambda,n} \right] ,$$

so that by (7.5)-(7.6)-(7.7), we get

$$\| \mu \, P_n^{K_n} \|_\infty \leq a \left[1 - c \exp\left(-\frac{A}{T_n}\right) \right]^{K_n} \| \mu \|_\infty , \tag{7.8}$$

where a is a (new) constant).

To simplify the notations, let

$$(7.9)\qquad \begin{cases} t_n = \exp\left(-\frac{1}{T_n}\right) & \inf_{x\in E} H(x) = h \\ \inf_{x\in E-E_{MIN}} H(x) = B+h & \text{where } B > 0 \ . \end{cases}$$

Recall that the invariant measure μ_n of P_n is the Gibbs measure

$$(7.10)\qquad \mu_n(x) = \frac{1}{Z_n}\, t_n^{H(x)} \ ,$$

where $Z_n = \sum_{y\in E} t_n^{H(y)}$.

An elementary computation based on (7.9)-(7.10) shows that

$$(7.11)\qquad \| \mu_n - \mu_{n+1} \|_\infty \leq a\, t_n^B \ ,$$

where a is a (new) constant.

Now let

$$(7.12)\qquad \varepsilon_n = \| \nu_n - \mu_n \|_\infty \ .$$

Using (7.2) and the invariance of μ_n yields

$$\nu_n - \mu_n = \nu_{n-1}\, P_n^{K_n} - \mu_n\, P_n^{K_n}$$

$$= (\nu_{n-1} - \mu_{n-1})\, P_n^{K_n} + (\mu_{n-1} - \mu_n)\, P_n^{K_n} \ ,$$

which in view of (7.8) and (7.11) yields

$$(7.13)\qquad \varepsilon_n \leq a(1 - c\, t^A)^{K_n}\, \varepsilon_{n-1} + a\, t_n^B \ .$$

Of course, (7.13) implies immediately

$$(7.14)\qquad \varepsilon_n \leq u_n \left[\sum_{k=0}^{n} \frac{\omega_k}{u_k} \right] ,$$

with the notations

$$(7.15)\qquad \omega_n = a\, t_n^B \quad \text{for } n \geq 1 \ , \ \omega_0 = \varepsilon_0 \ , \ u_0 = 1 \ ,$$

$$(7.16)\qquad u_n = a^n \prod_{k=1}^{n} (1 - c\, t_k^A)^{K_k} \quad \text{for } n \geq 1 \ .$$

We can now play around with the K_n and $t_n = \exp -\frac{1}{T_n}$ *to make sure that* ε_n *tends to zero as* $n \longrightarrow +\infty$. Indeed, as soon as this is the case, the law ν_n of X_n must have *the same* limit as μ_n for $n \longrightarrow +\infty$, and this last limit is obviously *the uniform distribution* μ_{MIN} *over the set* E_{MIN} *of global minima.*

The expression (7.14) shows that there is a wide choice of such minimizing schedules, and that the bound obtained for ε_n cannot tend to zero as $n \longrightarrow +\infty$ unless one has $\lim_{n\to+\infty} u_n = 0$, which is equivalent to

$$(7.17)\qquad \lim_{n\to+\infty} \left[-\sum_{k=1}^{n} K_k\, t_k^A + n \log a \right] = -\infty \ .$$

In particular, $\sum_{k=1}^{\infty} K_k \, t_k^A = +\infty$.

For instance, select K_k such that

$$K_k = (c \, t_k^A)^{-1} \, (\log a + bc) \qquad \text{where} \quad b > 0 \ .$$

Then (7.16) yields, up to multiplicative constants

$$u_n \sim e^{-bn} \ .$$

Now impose $\frac{w_n}{u_n} = \frac{1}{n^\alpha}$ with $\alpha > 1$, so that (7.15)-(7.14) yield successively

$$t_n \sim \frac{1}{n^{\alpha/B}} \, e^{-(b/B)n}$$

$$\varepsilon_n \sim u_n \sim e^{-bn}$$

$$K_n \sim n^{\alpha A/B} e^{b(A/B)n}$$

$$L_n = K_1 + \ldots + K_n \sim n^{\alpha A/B} e^{b(A/B)n} \ . \tag{7.18}$$

Coming back now to *the computation time* $L = L_n$, we can evaluate the error $\varepsilon_{[L]}$ and the temperature $T_{[L]}$ after L elementary steps of the algorithm

$$n \sim \frac{1}{b} \frac{B}{A} \log L$$

$$\varepsilon_{[L]} \sim e^{-bn} \sim \frac{1}{L^{B/A}} \tag{7.19}$$

$$T_{[L]} = -\frac{1}{\log t_n} \sim \frac{B}{bn} \sim \frac{A}{\text{Log } L} \ . \tag{7.20}$$

Thus we see that, for this particular minimizing cooling schedule, *the temperature should decrease* like $\frac{A}{\text{Log } L}$ where L is the *computation time*, and the distance between ν_L and μ_{MIN} is of the order of $\frac{1}{L^{B/A}}$. Note that B is generally smaller than A , and that *the computation time* K_n *at temperature* T_n *is of the order of* cte × exp $\frac{A}{T_n}$.

Moreover, (7.18) shows that $L_n \sim K_n$ so that the *essential part of the computation time is spent at the lowest temperature reached during that time*. This means that such annealing schedules actually are very close to Glauber dynamics at fixed low temperature.

Of course, the features exhibited by this particular example do not necessarily hold for all cooling schedules based on the bound (7.14) but the expressions (7.19) (7.20) for $\varepsilon_{[L]}$ and $T_{[L]}$ are likely to be the best "rough" lower bounds which can be achieved by schedules based on (7.14).

We also point out that there is quite a bit of freedom in the selection of the stagewise cooling schedule T_n ; as is easily checked, very fast cooling in this setup simply has to be paid for by much longer stages at each fixed temperature.

One interesting aspect of the computations given in sections 6 and 7 is that *the constant* $A = C_1 - C_2$ *introduced in* (7.6)-(6.6) *has been rigorously identified (by Chiang and Chow) with the constant* R *of* th. 5.6.

We may also interpret the other basic constant D introduced by Hajek (cf. 5.5). Indeed assume we are only interested in the speed at which, in the preceding annealing schedules, $\nu_n[E - E_{MIN}]$ tends to zero as $n \longrightarrow +\infty$. Then, letting v be the column vector corresponding to the indicator function of $E - E_{MIN}$, we will be concerned with the behaviour of $\mu P_n^{Kn} v$ where μ is a measure on E having a total mass equal to 0. Thus we write, using (7.3)-(7.5),

$$\mu P_n^{Kn} v = \sum_{\lambda \in W_n} \lambda^{Kn} \mu Q_{\lambda,n} v ,$$

where W_n is a (*generally strict*) *subset* of the set Λ_n of all eigenvalues of P_n which are not equal to 1. Namely, we may very well have $Q_{\lambda,n} v = 0$ for several $\lambda \in \Lambda_n$, and such λ will not appear in W_n. Hence if the eigenvalue with highest modulus in W_n is the $(1+k)^{th}$ in decreasing order, the constant $A = C_1 - C_2$ of the preceding computations will be replaced by the constant $A_k = C_k - C_{k+1}$ with the notations (6.4)-(6.6).

It is then natural to expect the following *conjecture* to hold :

(7.21) *The constant* D *introduced by Hajek to characterize the minimizing cooling schedules (see* 5.5) *coincides with* $C_k - C_{k+1}$, *for some integer* $k \geq 1$, *where the* C_k *are the "costs" introduced by Wentzell (see* (6.4)).

Now a few simple examples show that when $\text{card } E_{MIN} = r$, it seems quite possible to have $D = C_k - C_{k+1}$ with $k \geq r+1$.

8. PARALLELIZATION OF ANNEALING ALGORITHMS

As the preceding asymptotic results indicate, annealing algorithms tend to be fairly slow, and for large scale optimization problems, most of the applied work relies on "fast" cooling schedules of the type $T_n = T_o a^n$ where $a = .95$ or $a = .99$. Needless to say that the minimizing property does not mathematically hold for these cooling schedules, which does not prevent them to have useful performances *when suitably taylored* to fit a given application (see for instance Bonomi - Lutton, Dreyfus, Uhry among many other specialists).

Another approach, much more recent, is to try to use *parallel computing* to accelerate the algorithm. This raises fairly complicated mathematical questions, and opens a wide field of experimentations for parallel computing experts. I am currently involved in the investigation of these questions with the collaboration of Trouvé, Graffigne, Lutton, Bougé, Virot, Roussel, Tourangeau, Uhry and others, within the framework of a CNET - University Paris-Sud project.

In this brief survey, I will select *only one* aspect of the parallelization problem for simulated annealing.

We come back to the microscopic description of the configuration space E as $E = L^S$, where S is a finite set of indices or "sites"; configurations are noted $x = (x_s)_{s \in S}$ and the energy function $H(x)$ to be minimized is of the form (see § 1)

$$H(x) = \sum_{K \in C} U_K(x) \quad , \tag{8.1}$$

where C is the set of all *cliques* in S associated to a *neighborhood system* W_s, $s \in S$. Recall that the only restriction on the $W_s \in S$ is that $s' \in W_s$ if and only if $s \in W_{s'}$.

Cliques are precisely the subsets K of S such that any two sites in K are neighbours in S. The action potentials $U_K : E \longrightarrow \mathbb{R}$ depend only on the x_s, $s \in K$.

Call $P_T(x)$ the Gibbs measure (see 2.2) associated to the energy $H(x)$. Given the form (8.1) of $H(x)$, the conditional distribution of x_s given all the x_t, $t \in E - s$ depends only on the x_t, $t \in W_s - s$. This is the so-called *Markov field* property; if we write $W_{s\bullet} = W_s - s$ and note x_F the restriction of any configuration x to $F \subset S$, this conditional distribution

$$p_s(\lambda \mid x_t, t \in W_{s\bullet}) = P_T(x_s = \lambda \mid x_t, t \in W_{s\bullet}) \tag{8.2}$$

also called *local specification* of P_T can be computed easily.

Define first the function

$$U\left(\lambda, x_{W_{s\bullet}}\right) = \sum_{s \in K \in C} U_K(y) \quad , \tag{8.3}$$

where $\lambda \in L$ and y are any configurations such that $y_s = \lambda$, and $y_t = x_t$ for all $t \in W_{s\bullet}$. We then set

$$u\left(x_{W_{s\bullet}}\right) = \sum_{\lambda \in L} \exp\left[-\frac{1}{T} U\left(\lambda, x_{W_{s\bullet}}\right)\right] \tag{8.4}$$

to obtain directly

$$p_s\left(\lambda \mid x_{W_{s\bullet}}\right) = \frac{1}{u\left(x_{W_{s\bullet}}\right)} \exp\left[-\frac{1}{T} U\left(\lambda, x_{W_{s\bullet}}\right)\right] . \tag{8.5}$$

In this context, the elementary steps used to build a (sequential) simulated annealing algorithm consist in first selecting a sequence (s_n) of sites, which one generally constrains to be *periodic* and *to visit evey site* of S. One fixes a cooling schedule $T_n \searrow 0$. At time n, call $X(n)$ the current configuration, and let $\mathcal{F}_n$ be the σ-algebra generated by the $X(k)$, $k \leq n$.

In the sequential annealing algorithm, the $X(n)$ form a Markov chain, such that $X(n+1)$ coincides with $X(n)$ at all sites $s \neq s_n$, and such that $Z_{n+1} = X_{s_n}(n+1)$ verifies :

(8.6) $$P(Z_{n+1} = \lambda \mid \mathcal{F}_n) = p_{s_n}(\lambda \mid X_t(n), t \in W_{s_n} - s_n),$$

where in formulas (8.4)-(8.5), T has been replaced by T_n.

Asymptotic results similar to those quoted earlier in this text show that for $\lim_{n \to +\infty} T_n \log n \geq \bar{D} > 0$ where $\bar{D}$ is a suitable constant, then the law of $X(n)$ concentrates on the set E_{MIN} of configurations minimizing H.

Strictly speaking, these algorithms do not fit the abstract scheme described earlier. However, for small $T > 0$, we obviously have (cf. (8.4))

(8.7) $$p_s\left(\lambda \mid x_{W_{s\bullet}}\right) \sim \exp - \frac{1}{T}\left[U\left(\lambda, x_{W_{s\bullet}}\right) - U\left(\hat{x}_s, x_{W_{s\bullet}}\right)\right]$$

where $\hat{x}_s$ *is any point* $\lambda_0 \in L$ *minimizing the function* $\lambda \longrightarrow U\left(\lambda, x_{W_{s\bullet}}\right)$.

One can then use asymptotic results described earlier to analyze the optimal cooling rates, and show that indeed the constant $\bar{D}$ coincides with the Hajek constant D associated to $H(x)$. The key point here is that (8.7) shows that at temperature T, *and when the configuration* x *is a local minimum* of $H(x)$, then the transition matrix $p_{xy}(T)$ for a single annealing step in configuration space E still verifies

$$p_{xy}(T) \sim \exp - \frac{1}{T}\left[H(y) - H(x)\right]^+$$

although *this last statement does not hold* for arbitrary $x \in F$.

Now a natural idea, if one has access to computers allowing simultaneous parallel computations, is to *refresh the values of large sets of sites simultaneously, by synchronous random choices at all such sites* ; at time n, we may for instance decide to change simultaneously all the $X_s(n)$, the simultaneous choices being independent, while, for each $s \in S$, the choice of $X_s(n+1)$ given $X(n)$ is still made according to the conditional distribution (8.6).

It is easy to write explicitly the transition matrix $q_{xy}(T)$ of this new Markov chain in configuration space, at temperature T ; indeed one proves easily that

(8.8) $$q_{xy}(T) \sim \exp\left[-\frac{1}{T} U_{xy}\right],$$

where the *cost function* U_{xy} is given by

(8.9) $$U_{xy} = \sum_{s \in S}\left[U\left(y_s, x_{W_{s\bullet}}\right) - U\left(\hat{x}_s, x_{W_{s\bullet}}\right)\right]$$

(the notations are those of (8.7)).

Now the Freidlin-Wentzell theory of stochastic matrices with exponentially vanishing terms can be applied to obtain optimal cooling rates and to understand the nature of the limiting equilibrium distribution. Indeed, if μ_T is the equilibrium distribution for a stochastic matrix $q(T)$ given by (8.8), Freidlin-Wentzell showed that $\lim_{T \searrow 0} \mu_T = \mu_0$ *exists and is concentrated on a set* E_0 *of*

configurations which can be characterized as follows.

With the notations of (6.3), recall that the cost function U_{xy} associates to each graph G over E a cost $U(G)$. Let

$$u(x) = \inf \{U(G) \mid G \in S(\{x\})\} , \tag{8.10}$$

where $S(F)$ has been defined in (6.2) for $F \subset E$.
Then the set E_o of *terminal configurations* is given by

$$E_o = \{y \in E \mid u(y) = \inf_{x \in E} u(x)\} . \tag{8.11}$$

Given the explicit but two-links remote relation between $H(x)$ and U_{xy}, there is *a priori* no reason why E_o should coincide with E_{MIN}. It is in fact reasonably easy, using (8.10)-(8.11), to construct *many examples where* E_o *contains configurations* x *which are not global minima.*

For recent results in this direction, we refer to Trouvé's note, where he proves by direct methods the existence of limit distribution in synchronous simulated annealing. The same question can also be handled with the help of results of Föllmer and altri on general non homogeneous Markov chains, in the spirit of Dobrushin's ideas.

We will expand on the subject in forthcoming papers as well as in the extended version of this text.

BIBLIOGRAPHY

[1] V. CERNY - A *thermodynamical efficient simulation algorithm*, J. Opt. Theory App., vol. 45, p. 41-51, 1985.

[2] M. FREIDLIN - A. WENTZELL - *Random perturbation of dynamical systems*, Springer, Berlin, 1984.

[3] T. CHIANG and Y. CHOW - *Convergence rate of annealing processes*, preprint, 1987.

[4] O. CATONI - *Optimal cooling schedules for annealing*, C.R. Ac. Sci. Paris, and preprint, 1988

[5] H. FÖLLMER - *Remarks on simulated annealing*, Lectures University Paris-Sud, 1988.

[6] S. GEMAN - D. GEMAN - *Stochastic relaxation, Gibbs distribution, Bayesian restauration of images*, I.E.E.E. Trans. P.A.M.I., vol. 6, p. 721-741, 1984.

[7] S. GEMAN - C.R. HWANG - *Diffusions for global optimization*, preprint, 1985, Brown University.

[8] B. GIDAS - *Non stationary Markov chains and convergence of annealing algorithms*, J. Stat. Phys. 39, p. 73-131, 1985.

[9] B. HAJEK - *Cooling schedules for optimal annealing*, preprints Math. Op. Research, 1987.

[10] B. HAJEK - *Tutorial survey of simulated annealing*, Proc. 24th Conf. Decision/ Control, Fort Lauderdale, 1985.

[11] R. HOLLEY - D. STROOCK - *Simulated annealing via Sobolev inequalities*, preprint, 1987.

[12] C.R. HWANG - S.J. SHEU - *Large time behaviour for perturbed diffusions* I, II, III, Preprints, 1986-88.

[13] S. KIRKPATRICK - C. GELATT - M. VECCHI - *Optimization by simulated annealing*, Science 220 (1983), p. 671-680.

[14] N. METROPOLIS and altri - *Equation of state calculations*, J. Chem. Physics 21, p. 1087-1092, 1953.

[15] A. TROUVÉ - *Synchronous simulated annealing*, C.R. Ac. Sci. Paris and preprint, 1988.

Robert AZENCOTT

Université Paris-Sud
Département de Mathématiques
ERA 743 du CNRS
Bâtiment 425
F-91405 ORSAY

et

École Normale Supérieure
Département de Mathématiques et Informatique
45 rue d'Ulm
F-75231 PARIS CEDEX 05