August 1988 LIDS-P-1808

Convergence of the Simulated Annealing Algorithm[1]

B. Delyon

LIDS, MIT, Cambridge, 01239, MA, USA.

Abstract: We prove the convergence of the simulated anealing algorithm by estimating the second eigenvalue of the transition matrices (associated to each temperature).

[1] Work supported by the Army Research Office under grant DAAL03-86-K-0171,

I) Description of the algorithm; statement of the problem.

Simulated anealing is an algorithm used to minimize a any cost function J defined on a space E on which is defined a neighborhood structure (i.e.,a symetric binary relation on E; each point of E is assumed to have a finite number of neighbors). This algorithm creates a Markov chain X_n on E in the following way (the parameter θ is known as temperature): if X_n is given, then choose at random a neighbor x of X_n (usually with uniform probability) and an exponential variable Z_n, compute $\Delta J=J(x)-J(X_n)$, and if $\Delta J<\theta Z_n$, the transition is accepted and $X_{n+1}=x$, if not, $X_{n+1}=X_n$ (i.e. the transition is accepted with probablity $\exp(-(\Delta J)^+/\theta)$). Actually the temperature may vary (decreasing to zero) during the algorithm so that θ has to be replaced above by θ_n.

For a fixed θ, the invariant measure of the chain is ($\beta = \theta^{-1}$):

$$(1) \qquad \pi(x) = \pi(x,\beta) = Z^{-1} \exp(-\beta J(x)),$$

where Z is a normalization constant. When the temperature tends to zero (β tends to infinity), π tends to a uniform measure on the set of global minima of J; the idea of the algorithm is to decrease slowly enough the temperature to be close enough to π at each step and to get at a global minimum at the end. The problem is:

How fast has θ_n to decrease in order to keep the law of X_n close enough to $\pi(\beta_n)$?

This problem has been recently studied by B.Hajek in [1](and by many others); we propose here similar results proved by a quite different method which emphasize the key properties of the transition matrix.

Denoting by $\alpha_n = (\alpha_n(x))_{x\in E}$ the law of X_n and by $P(\beta)$ the transition matrix of the process at temperature β, we have

$$(2) \qquad \alpha_{n+1} = \alpha_n P(\beta_n).$$

We will also study the continuous case, where X_t is a jump process and the law α_t of X_t is solution of:

$$(3) \quad \frac{d\alpha_t}{dt} = \alpha_t \, (P(\beta_t)-I)$$

where $P(\beta)$ is the transition matrix associated to the temperature β^{-1}.

We will assume in the sequel that each point of E has the same number of neighbors $N_b \geq 2$. N_s will denote the number of points in E, and B the set of global minima of J.

Theorem1: *For any schedule of the form*

$$\beta_t = h \log(t+T) \qquad (\text{resp. } \beta_n = h \log(n+N))$$

where T (resp. N) is arbitrary and h is smaller than δ defined by

$$\delta = \max_{m \in B} \max_{y \in E} \min_{p \in P(y,m)} \max_{x \in p} J(x)-J(y)$$

and P(y,m) denotes the set of paths (sequence of neighbor points) leading from y to m,

α_t (resp. α_n) tends exponentially fast to the uniform measure on B as t (resp. n) tends to infinity.

Remark: The constant given by B.Hajek in [1] is

$$\delta' = \max_{m \in B} \max_{y \in E \backslash B} \min_{p \in P(y,m)} \max_{x \in p} J(x)-J(y)$$

Using this constant, he does not obtain that α_n tends to the uniform measure on B, but only that $\alpha_n(B) \to 1$.

II) Some properties of the transition matrices.

At each temperature θ, the transition matrix $P=P(\beta)$ will be determined by the relations($\beta = \theta^{-1}$):

$p(x,y) = 0$ if x and y are not neighbors and $x \neq y$

(4) $p(x,y) = N_b^{-1} \exp(-\beta(J(y)-J(x))^+)$ x and y neighbors

$p(x,x) = 1-\sum_{x \neq y} p(x,y)$.

The basic property of the matrix $P = P(\beta)$ is

(5) $p(x,y) = p(y,x)\ \pi(y)/\pi(x)$

so that, if we denote by $D=D(\beta)$ the diagonal matrix having $\pi(x)^{1/2}$ as (x,x) entry, we have:

(6) $S = DPD^{-1}$ is a symetric matrix.

All the eigenvalues of P are real.

Note that

$s(x,y) = N_b^{-1} \exp(-\beta|J(y)-J(x)|/2)$ if $x \neq y$ and x and y are neighbors,

$s(x,y) = p(x,y)$ elsewhere.

Once it is observed S is symetric, it will be easy, using a change of variables (section IV), to reduce the problem to the estimation of the second eigenvalue of $P(\beta)$ for any β.

The next section is devoted to the estimation of this eigenvalue.

III) Estimation of the second eigenvalue of P.

This section is devoted to the proof of the following

theorem2: *Denoting by $\lambda(\beta)$ is the eigenvalue of P which is closest to 1 and different from 1, the following is true (δ is defined in theorem1):*

$$\lim_{\beta \to \infty} \frac{\log(1-\lambda(\beta))}{\beta} = -\delta$$

The same property is true if P is replaced by P^2.

We begin by recall a result due to M.I.Friedlin and A.D.Wentzell, given in [2], which provides an expression for the characteristic polynomial of a stochastic matrix. It requires some notations:

Definition : *Let L be a finite set and let a subset W be selected in L. A graph on L is called a W-graph is it satisfies the following conditions:*

(1) every point $m \in L \backslash W$ is the initial point of exactly one arrow, and any arrow has its initial point in $L \backslash W$.

(2) there are no closed cycles in the graph.

Note that (2) may be replaced by

(2') every point $m \in L \backslash W$ is the initial point of a sequence of arrows leading to some point $n \in W$.

These W-graphs may be seen as disjoint unions of directed trees on L with roots in W.

Notations:

The set of W-graphs will be denoted by G(W).

Suppose that we are given a set of numbers p_{ij} $(i,j \in L)$, then for any graph g on L we define the number $\pi(g)$ by:

$$\pi(g) = \prod_{(m \to n) \in g} p_{m\,n}$$

$$\pi(\text{empty graph}) = 1.$$

For any subset W of L, we put;

$$\sigma(W) = \sum_{g \in G(W)} \pi(g) \tag{8}$$

In particular, $\sigma(L)=1$ and $\sigma(\emptyset)=0$.

We can now state

<u>theorem</u>3: The *characteristic polynomial of an nXn stochastic matrix* $P=(p_{ij})$, *has the form:*

$$(9) \quad P(\lambda) = \sum_{i=1}^{n} \sigma_i (\lambda-1)^i$$

where

$$(10) \quad \sigma_i = \sum_{|W|=i} \sigma(W).$$

An upper bound on the second characteristic value of P will be $1-\varepsilon$, for any ε such that all the roots of the polynomial

$$Q(x) = \sum_{i=1}^{n} (-1)^i \sigma_i x^{i-1}$$

are all larger than ε. Note that all the roots of Q are larger than σ_1/σ_2 (because they are all positive and σ_2/σ_1 is the sum of the inverse of the roots; in the case of a general Markov chain, when the roots are complex, this gives a bound on the real parts).

We are now going to study the W-graphs which have the largest contribution in the sums σ_1 and σ_2 (cf. eqs (8) and (10)). They will be denoted by g_0 and (g_1,g_2) (g_1 and g_2 are two connected graphs with no vertice in common) and do not depend on β. Because of (4), it is clear that, when β tends to infinity, we have

$$(11) \quad \begin{aligned} \sigma_1 &\sim N_1 \pi(g_0) \\ \sigma_2 &\sim N_2 \pi((g_1,g_2)) \end{aligned}$$

where N_1 (resp. N_2) is the number of graphs g in the sum σ_1 (resp. σ_2) such that $\pi(g) = \pi(g_0)$ (resp. $\pi((g_1,g_2))$). We will give a characterization of g_0 and prove that (g_1,g_2) may be obtained from g_0 by removing an arrow out of it.

The following lemma is basic for the estimation of σ_1 and σ_2.

Lemma1: *For any point x and y of E, there exist a one-to-one map* φ *between* $G(\{x\})$ *and* $G(\{y\})$ *such that, for any* $g \in G(\{x\})$,

$$\pi(\varphi(g)) = \exp(J(x)-J(y))\ \pi(g).$$

This map consists in changing, in g, the orientation of the sequence of arrows going from y to x.

This is an easy consequence of eq(5).

For simplicity, we will suppose that J has only one global minimum m_0.

It is then clear that $g_0 \in G(\{m_0\})$ and $(g_1,g_2) \in G(\{m_1,m_2\})$ where m_1 (resp. m_2) realizes the minimum of J over the set of vertices of g_1 (resp. g_2). From now on, we will only consider graphs having this last property. To any such graph g, one can associate the undirected tree obtained by forgetting the orientation of the edges.

If $g \in G(\{m\})$, we have:

$$-\log(\pi(g)) - (N_s-1)\log(N_b) = \sum_{x\to y\in g} \beta(J(y)-J(x))^+$$

$$= \sum_{x\to y\in g} \beta(J(y)VJ(x)-J(x)) \qquad V \text{ stands for sup}$$

$$= \sum_{x\to y\in g} \beta(J(y)VJ(x)) - \sum_{x\in E} \beta J(x) + \beta J(m)$$

$$= \beta K(t) + \beta J(m) - \sum_{x\in E} \beta J(x)$$

where t is the tree associated to g and K(t) is the length of the tree (the length of an edge e=(x,y) of t being K(e) = J(y)VJ(x) for x and y neighbors).

If $g=(g',g'') \in G(\{m',m''\})$, we have in the same way:

$$-\log(\pi(g)) - (N_s-2)\log(N_b) = K(t') + K(t'') + \beta J(m') + \beta J(m'') - \sum_{x\in E} \beta J(x)$$

where t' and t'' are the trees associated to g' and g''.

The two last equalities have reduced the problem to a problem of minimum spanning trees. We have obviously:

Lemma2: *g_0 is associated to a minimum spanning tree on E, where K is the length function.*

The following result will be needed:

Theorem4:

(a) t_0 is a minimum spanning tree iff any spanning tree t_1 obtained by removing one edge out of t_0 and adding another one somewhere else satisfies $K(t_1) \geq K(t_0)$.

(b) the edge $e=(x,y)$ is in some minimum spanning tree iff for any path p leading from x to y there exists an edge $e' \in p$, $e' \neq e$, such that $K(e') \geq K(e)$.

(c) the path p is in some minimum spanning tree iff for any p' having the same extreme vertices the following is satisfied:

$$\max_{e' \in p'} K(e') \geq \max_{e \in p} K(e).$$

Proof: This theorem is contained in remarks1 and 4 of [3] in the case where $K(e) \neq K(e')$, $e \neq e'$ (t_0 is unique, the three inequalities are strict and (b) and (c) are characterization of the edges and paths of t_0). For the general case, consider t_0 (resp. e, p) satisfying one of the conditions of (a) (resp. (b), (c)); modify K into $K' = K + \varepsilon K_0$, where K_0 is non-positive on the edges t_0 (resp. e, p) and non-negative out of t_0 (resp. e, p) so that $K'(e) \neq K'(e')$, $e \neq e'$, and utilize the theorem with K' and let ε tend to zero to prove (a) (resp. (b), (c)).

We will use (a) to prove the following

lemma3: The tree (t_1,t_2) (associated to (g_1,g_2)) may be obtained from t_0 (associated to g_0) by removing an edge out of it.

Proof: Consider the sets A_1 and A_2 of vertices of t_1 and t_2 and the points x_1 and x_2 which minimize $J(x)VJ(y)$ over all the couples of neighbor points $(x,y)\in A_1\times A_2$. Denote by e the edge (x_1,x_2) and by t the tree obtained as the union of t_1, t_2, and $\{e\}$ (note that $K(e)=J(x)VJ(y)$). Clearly, t is a spanning tree. t may be represented:

$$t_1 \overset{e}{—} t_2$$

We will prove by contradiction that it satisfies (a). If it does not, there exist two edges e_1 and e_2 such that $t'=\{e_2\}\cup t\backslash\{e_1\}$ is still a spanning tree and $K(t')<K(t)$. Three cases are possible: $e_1=e$, $e_1\in t_1$, $e_1\in t_2$. We only have to consider the two first ones (if we are not in the first case, we rename t_1 as the tree which has e_1 in it, t_2 being the other one).

<u>case</u>1, $e_1=e$:

The relation $K(t')<K(t)$ implies $K(e_2)<K(e)$ which is in contradiction with the choice of e.

<u>case</u>2, $e_1\neq e$, $e_1\in t_1$:

In that case, A_1 is the union of two sets B_1 and B_2 connected by e_1, B_2 being the one which is connected to A_2 by e. We have the following picture for t:

$$B_1 \overset{e_1}{—} B_2 \overset{e}{—} A_2.$$

The optimality of t_1 (it is necessarily a minimum spanning tree of A_1) implies that e_2 does not connects B_1 to B_2 (because in that case we t' would satisfy $K(t')\geq K(t)$). Consequently, e_2 connects B_1 to A_2 and t' is organnized as follows:

$$B_2 \overset{e}{—} A_2 \overset{e_2}{—} B_1.$$

For any set A we will denote by m(A) the minimum value of the function J over A. We consider $t_1'\cup t_2'=t'\backslash\{e_2\}$ and the graph $(g_1',g_2')\in G(\{m(B_1),m(A_2\cup B_2)\})$ associated to (t_1',t_2'). The relation

$$\pi((g_1',g_2')) \leq \pi((g_1,g_2))$$

becomes

$K(t_1') + K(t_2') + \beta J(m(A_2 \cup B_2)) + \beta J(m(B_1)) \geq K(t_1) + K(t_2) + \beta J(m_1) + \beta J(m_2)$

$K(t') - K(e_2) + \beta J(m(A_2 \cup B_2)) + \beta J(m(B_1)) \geq K(t) - K(e) + \beta J(m_1) + \beta J(m_2)$

which implies (using $K(e_2) \geq K(e)$ and $K(t') < K(t)$):

(12) $\beta J(m(A_2 \cup B_2)) + \beta J(m(B_1)) > \beta J(m_1) + \beta J(m_2)$.

On the other hand, considering $t_1'' \cup t_2'' = t' \setminus \{e\}$ and the graph $(g_1'', g_2'') \in G(\{m(B_2), m(A_2 \cup B_1)\})$ associated to (t_1'', t_2''). The relation

$\pi((g_1'', g_2'')) < \pi((g_1, g_2))$

becomes

$K(t_1'') + K(t_2'') + \beta J(m(A_2 \cup B_1)) + \beta J(m(B_2)) > K(t_1) + K(t_2) + \beta J(m_1) + \beta J(m_2)$

$K(t') - K(e) + \beta J(m(A_2 \cup B_1)) + \beta J(m(B_2)) > K(t) - K(e) + \beta J(m_1) + \beta J(m_2)$

which implies:

(13) $\beta J(m(A_2 \cup B_1)) + \beta J(m(B_2)) > \beta J(m_1) + \beta J(m_2)$.

Defining $a = m(A_2)$, $b_1 = m(B_1)$, $b_2 = m(B_2)$, relations (12) and (13) may be rewritten as:

$a \Lambda b_2 + b_1 > b_1 \Lambda b_2 + a$

$a \Lambda b_1 + b_2 > b_1 \Lambda b_2 + a$

If $b_1 \leq b_2$, the first equation gives a contradiction (because the inequality is strict), if not, consider the second one.

This ends the proof of lemma3.

<u>Proof of theorem4</u>:

The last lemma implies that (g_1, g_2) may be obtained from g_0 by removing an arrow $x \rightarrow y$ and reversing the path going from $m'(x)$ to x, where $m'(x)$ is the point which minimizes J over all the points leading to x. The arrow choosen will be one which maximizes

$J(x)VJ(y)\text{-}m'(x)$.

Note that property (c) implies that for any x, the sequence of arrows of g_0 leading from x to m_0 will be one which minimizes the supremum of J along the path. Consequently, maximizing the expression above will give

$$\delta = \max_{m} \min_{p\in P(m)} \max_{x\in p} J(x)\text{-}J(m)$$

where P(m) is the set of paths leading from m to m_0, and

$$\frac{\pi(g_0)}{\pi(g_1,g_2)} = \frac{1}{N_b} \exp(-\beta\delta)$$

Finally, using (11), we get

$$(14) \quad \frac{\sigma_1}{\sigma_2} \sim \frac{N_1}{N_2 N_b} \exp(-\beta\delta).$$

Note that, because $\frac{\sigma_2}{\sigma_1}$ is the sum of the inverted roots of Q, we have the bounds

$$\frac{\sigma_1}{\sigma_2} \le 1\text{-}\lambda(\beta) \le N_s \frac{\sigma_1}{\sigma_2}$$

which, with (14), gives the first assertion of theorem4.

if we now replace P by P^2, the same reasonning may be done. The new function K is:

(15) $K'(e,\beta) = (-\log(p^{(2)}(x,y)) - 2\log(N_b) + \beta J(x))/\beta$

for $e=(x,y)$.

Note that because equation(5) remains true for P^2, switching x and y in (15) does not change the result.

If x and y are not neighbors, $K'(e,\beta)=K'(e)$ does not depend on β and

$$(16) \quad K'(e) = (J(y)\text{-}J(z))^+ + (J(z)\text{-}J(x))^+ + J(x) \ge K((x,z))VK((y,z))$$

where z is a neighbor of x and y which minimmizes the second expression.

If x and y are neighbors and x or y $K'(e,\beta)$ tends to a limit $K'(e)$; if x or y is not a local maximum, then

$$K'(e) = K(e)$$

(because, if for instance $J(x) \geq J(y)$, $p^{(2)}(x,y)$ tend to a positive limit). If x and y are local maxima, $p(x,x)$ and $p(y,y)$ are zero,$J(x)=J(y)$, and

$$K'(e) \geq K(e).$$

A minimum spanning tree t_0' for K' is still characterized by property (b) and if t_0 does not have any edge made of two local maxima, t_0 will be a minimum spanning tree for K'. If not, we only have to make a slight modification of t_0: consider three points x,y,z, successively connected in t_0, such that y and z are local maxima, the modification consists in removing the edge (y,z) and connecting x to z and it is easily verified, using (16), that

$$K'((x,z)) = J(x) \vee J(z) = K((x,z))$$
$$K'(x,y) = K((x,y)).$$

This can be done even if more than two local minima are successively connected (except in the case where E consists in two points x and y with $J(x)=J(y)$, but in this case $N_b=1$). Once we have a minimum spanning tree t_0' the proof is easily carried out and we get the same value for δ.

IV) Proof of theorem1.

A) Continuous case:

For any differentiable schedule β_t, we will consider

$$\mu_t = \alpha_t - \pi(\beta_t) = \alpha_t - \pi_t.$$

We have

$$(17) \quad d\mu_t = \mu_t\,(P(\beta_t)-I)\,dt - d\pi_t = \mu_t\,(P_t - I)\,dt - d\pi_t.$$

Let now

$$\nu_t = \mu_t\, D_t^{-1}$$

where D has been defined in part two (and depends on t because of the schedule).

Then

$$d\nu_t = \mu_t\,(P_t-I)D_t^{-1}\,dt - (d\pi_t)D_t^{-1} + \mu_t\,dD_t^{-1}$$
$$= \nu_t\,D_t(P_t-I)D_t^{-1}\,dt - (d\pi_t)D_t^{-1} + \nu_t\,D_t\,dD_t^{-1}$$
$$= \nu_t\,(S_t-I)\,dt - (d\pi_t)D_t^{-1} + \nu_t\,D_t^{-1}dD_t$$

(18) $d\nu_t\nu_t^T = 2\nu_t\,(S_t-I)\,\nu_t^T\,dt - 2(d\pi_t)D_t^{-1}\nu_t^T - 2\nu_t\,D_t^{-1}(dD_t)\,\nu_t^T.$

An elementary calculation shows that ($d_t(x)$ is the x^{th} diagonal entry of D):

$$d\pi_t(x) = \pi_t(x)\,(-J(x) + \sum_y J(y)\pi_t(y))\,d\beta_t.$$
$$dd_t(x) = d_t(x)^{-1}/2\;d\pi_t(x)$$

and we get the bounds:

$$\|(\frac{d\pi_t}{d\beta_t})\,D_t^{-1}\|^2 = \sum_y J(y)^2\pi_t(y) - (\sum_y J(y)\pi_t(y))^2 = Var_t(J) \le V$$

$$|2\,\frac{dd_t(x)}{d\beta_t}\,d_t(x)^{-1}| = |\,-J(x) + \sum_y J(y)\pi_t(y)\,| \le \Delta$$

$d\nu_t\nu_t^T = 2\nu_t\,(S_t-I)\,\nu_t^T\,dt - 2(d\pi_t)D_t^{-1}\nu_t^T - 2\nu_t\,D_t^{-1}(dD_t)\,\nu_t$

where V is the maximum variance of J over all the laws π_t and Δ is the difference between the two extreme values of the function J. Equation (18) becomes (we assume that $\beta(t)$ is a decreasing function):

$$d\nu_t\nu_t^T \le 2(\lambda\,(\beta_t)-1)\nu_t\nu_t^T\,dt + 2V^{1/2}\,(\nu_t\nu_t^T)^{1/2}\,\beta_t'\,dt + \Delta\,\nu_t\nu_t^T\,\beta_t'\,dt$$

Finally, setting $n_t = (\nu_t\nu_t^T)^{1/2}$, we get

$$2n_tdn_t \le 2(\lambda\,(\beta_t)-1)\,n_t^2\,dt + 2V\,n_t\,\beta_t'\,dt + \Delta\,n_t^2\,\beta_t'\,dt$$

(19) $dn_t \le (\lambda(\beta_t)-1)\,n_t\,dt + V\,\beta_t'\,dt + (\Delta/2)\,n_t\,\beta_t'\,dt.$

Theorem2 implies that for any $\eta>0$, there exists B such that if $\beta>B$

$$\lambda(\beta)-1 < -\exp(-\beta(\delta-\eta))$$

Taking β_t=hlog(t) in (19) gives, for t large enough

(20) $dn_t \leq -t^{-(\delta-\eta)/h}\, n_t\, dt + V \frac{h}{t} dt + (\Delta/2)\; n_t \frac{h}{t} dt.$

If h has been choosen smaller than δ, η can be choosen so that $\delta-\eta$ is larger than h and n_t will converge exponentially fast to zero. This gives:

$$\sum (\alpha_t(x)-\pi_t(x))^2/\pi_t(x) \leq c_1\ \exp(-c_2 t)$$

for some constants c_1 and c_2, and then

$$(\alpha_t(x)-\pi_t(x))^2 \leq c_1\ \exp(-c_2 t).$$

B) Discrete case:

We consider now $\mu_n = \alpha_n - \pi(\beta_n)$ which satisfies now the equation

$$\mu_{n+1} = \mu_n P(\beta_n) - \pi(\beta_{n+1}) + \pi(\beta_n).$$

$\nu_n = \mu_n D_n^{-1}$ satisfies the equation

(21) $\nu_{n+1} = \nu_n S_n + \nu_n S_n(D_n D_{n+1}^{-1} - I) - (\Delta_n\pi)D_{n+1}^{-1}$

where

$$\Delta_n\pi = \pi(\beta_{n+1}) - \pi(\beta_n).$$

As before, we have

(22) $\|(\Delta_n\pi)D_{n+1}^{-1}\|^2 \leq V\, (\Delta_n\beta)^2$

where

$$\Delta_n\beta = \beta_{n+1} - \beta_n.$$

And

(23) $\|D_n D_{n+1}^{-1} - I\| \leq c\Delta_n\beta$ for some c.

Using (21), (22) and (23), we get

$$\|v_{n+1}\| \leq \|v_n S_n\| + \|v_n S_n(D_n D^{-1}_{n+1} - I)\| + \|(\Delta_n \pi) D^{-1}_{n+1}\|$$

(24) $\|v_{n+1}\| \leq \|v_n\| \, |\lambda_2(\beta)| \, (1 + c\Delta_n\beta) + V^{1/2} \Delta_n \beta$

where $\lambda_2(\beta)$ is the eigenvalue of $P(\beta)$ with largest absolute value and different from 1. Theorem2 gives an assymptotic for $\lambda_2(\beta)$ (because the eigenvalues of P are real) and one can now easily carry out the proof as before.

References

[1] B.Hajek, "Cooling Schedules for Optimal Annealing", Mathematics of Operation Research, vol.13, n.2, may 1988.

[2] B.Delyon, "Expansions for determinants and for characteristic polynomials of stochastic matrices". LIDS-P-1802.

[3] P.Rosentiehl, "L'arbre minimum d'un graphe", in "Theory of Graphs, International symposium, Rome 1966",(P.Rosentiehl ed.), Dunod, Paris, 1966.