ANNALES DE L'I. H. P., SECTION B

LAURENT YOUNES

Estimation and annealing for Gibbsian fields

Annales de l'I. H. P., section B, tome 24, n° 2 (1988), p. 269-294
<http://www.numdam.org/item?id=AIHPB_1988__24_2_269_0>

Ann. Inst. Henri Poincaré,

Vol. 24, n° 2, 1988, p. 269-294. *Probabilités et Statistiques*

Estimation and annealing for Gibbsian fields

by

Laurent YOUNES

École Normale Supérieure, 45, rue d'Ulm, 75005 Paris;
Laboratoire de Statistique, Université Paris Sud,
Département de Mathématiques, Bât. n° 425,
91405 Orsay France

ABSTRACT. — This paper introduces a stochastic gradient algorithm based on the Gibbs sampler in order to compute the maximum likelihood estimator for Gibbsian fields with finite number of states. It is then shown that it is possible to couple this algorithm with an annealing algorithm, so that estimation and restoration can be done at the same time in image processing.

RÉSUMÉ. — Ce papier introduit un algorithme de gradient stochastique basé sur l'échantillonneur de Gibbs, convergeant vers l'estimateur de maximum de vraisemblance pour un modèle de Gibbs à nombre d'états fini. Il est ensuite montré qu'il est possible sous certaines conditions de coupler cet algorithme avec un algorithme de recuit, et de mener ainsi de front des procédures d'apprentissage et de restauration en imagerie.

1. INTRODUCTION

In image processing, one deals with a picture composed of elements called pixels. It is a useful approach to consider it as a realization of a random process. The picture is then modelized by a stochastic field

$X=(X_s,\ s\in D)$, where D is a set of sites (which are generally pixels), and X_s takes its values in a set E_s which is assumed to be finite all through this paper (if s is a pixel E_s may be the set of grey levels or colours). Geman and Geman [3] have extended the set of sites to non observable elements such as edge location, label carriers, etc.

If $\Omega=\prod_s E_s$, we will call an element of Ω a configuration, or a state and the law of the field is a probability π on Ω. If for all x, $\pi(x)>0$ then the field is a Gibbsian field *i. e.* of the kind:

$$\pi(x)=\exp(-H(x)-\operatorname{Log} Z);$$

H is an energy function and Z is a constant. The energy function is often given by qualitative properties, or by local specifications on the probability π (Hammersley-Clifford theorem); in most cases one construct the gibbsian field by giving oneself the energy function at first. This is why one usually takes this kind of notation for the law of a the field (in particular, introducing the constant Z).

The Gibbsian model is then: $(\pi_\theta,\ \theta\in\theta)$, $\theta\subset\mathbb{R}^\nu$ with

$$\pi_\theta(x)=\exp[-H(\theta,\ x)-\operatorname{Log}(Z(\theta))].$$

These models have been used in other domains than image processing: in spatial statistic ([1]) or in statistical physics (Ising models).

The complexity of the normalizing constant, $Z(\theta)$ makes the estimation of θ difficult. Besag [1] and Guyon ([4], [5]) have studied some methods of pseudolikelihood or coding that avoid dealing with $Z(\theta)$; these methods, that have the drawback of computing only a pseudo likelihood maximum, have the advantage to be almost entirely analytically computable. We propose here a stochastic gradient algorithm, which uses the Gibbs sampler, in order to estimate the maximum likelihood estimator, when the model is exponential (*i. e.* $H(\theta,\ X)=\langle\alpha(X),\ \theta\rangle$, $\theta=\mathbb{R}^\nu$). Exponential models are almost exclusively used in practice; they cover Besag's auto-models [1] and all the models proposed by Geman in image restoration. The presentation of this algorithm and conditions for convergence are exposed in parts 2, 3 and 4.

In image restoration Geman and Geman have introduced annealing algorithms to obtain maximum *a posteriori* estimator, in a Bayesian context. In this case the aim is—roughly speaking—to maximize $\pi_\theta(X)$ no more in θ, but in X among all available configurations which are in finite number by assumption. For this one must start with a configuration X^0,

and form from it a sequence X^n of configurations in this manner: one gives oneself a sequence $u(n)$ of sites and a sequence (T_n) of "temperatures" $T_n > 0$; at step n X^n is known and one gets X^{n+1} by renewing only the value at site $u(n)$ so that $X_s^n = X_s^{n+1}$ for all $s \neq u(n)$; $X_{u(n)}^{n+1}$ is chosen in a stochastic way: the probability of $X_{u(n)}^{n+1} = x$ is the probability of $Y_{u(n)} = x$ conditional to $Y_s = X_s^n$ for $s \neq u(n)$, where Y follows the Gibbsian law of energy: $H(\theta, .)/T_n$. If T_n decreases slowly to 0 and if $u(n)$ visits each site infinitely often, with a finite period, Geman and Geman have shown that the sequence X^n converges in distribution to the uniform law on the set of configurations on which π_θ is maximum.

In practice it will *a priori* be necessary to work in two steps. One must first estimate the parameter θ and then make an annealing with the estimated θ. If, for the estimation of θ one uses an iterative algorithm (such as the one proposed in this paper) one can think of making the annealing at the same time in the following way: the estimation algorithm provides a sequence θ_n of parameters which converges to the estimated parameter θ_*; so at step n of the estimation algorithm θ_n is known and one can use at step n of a parallel annealing algorithm the energy $H(., \theta_n)/T_n$ instead on $H(., \theta_*)/T_n$ which is still unknown. We show in part 5 that under some conditions on θ_n and $H(\theta, .)$, the annealing algorithm still converges in law to the uniform law on the set where $\pi_{\theta_*}(.)$ is maximum.

2. PRESENTATION OF THE ESTIMATION ALGORITHM

2.1. Presentation of the model

We recall that D is the set of sites and we note N its cardinal. The field on D is $X = (X_s;\ s \in D)$ and we assume that for each s, X_s takes its values in a finite set E of cardinal L (for notation simplicity we assume that this set is the same for each site. The set of all configurations is then:

$$\Omega = E^D = \{x = (x_s;\ s \in D);\ x_s \in E\}.$$

We note $\Lambda = \text{card}(\Omega) = L^N$.

The law of X is given by:

$$\pi_\theta(x)=\exp[-\langle\theta,\alpha(x)\rangle-\mathrm{Log}(Z(\theta))],\qquad \theta\in\mathbb{R}^\nu \tag{1}$$
$$Z(\theta)=\sum_{y\in\Omega}\exp(-\langle\theta,\alpha(y)\rangle)$$

$\alpha:\Omega\to\mathbb{R}^\nu$ is a known sufficient statistic $\langle .,.\rangle$ is the scalar product on $\mathbb{R}^\nu$.

If $x=(x_s;\ s\in D)\in\Omega$ and if $s\in D$ we will note

$$\bar{x}_s=(x_u;\ u\in D-\{s\})\qquad\text{and}\qquad \Omega(x,s)=\{y\in\Omega/\bar{y}_s=\bar{x}_s\}.$$

The probability for $X_s=x_s$ conditional to $X_u=x_u$ for $u\neq s$ is then:

$$\pi^s_\theta(x_s\mid\bar{x}_s)=\exp[-\langle\theta,\alpha(x)\rangle-\mathrm{Log}(Z_s(\theta,\bar{x}_s))]. \tag{2}$$

Where $Z_s(\theta,\bar{x}_s)=\sum_{y\in\Omega(x,\,s)}\exp(-\langle\theta,\alpha(y)\rangle)$.

2.2. Equation to solve

Given a realisation x_0 of the field, we want to find the maximum likelihood estimator of θ. This leads to maximize $\pi_\theta(x_0)$ in θ. the function $(\theta\to\mathrm{Log}(\pi_\theta(x_0))$ is concave; its differential is:

$$h(\theta)=E_\theta(\alpha(.))-\alpha_0\qquad[\text{where }\alpha_0=\alpha(X_0)]$$

and its second derivative is: $-\mathrm{Var}_\theta(\alpha(.))$.

We need then to solve:

$$h(\theta)=0. \tag{3}$$

In the following we will always assume that:

- For all θ, $\mathrm{Var}_\theta(\alpha(.))$ is positive definite.
- There exists θ_* with $h(\theta_*)=0$.

2.3. Gibbs sampler

The Gibbs sampler is an algorithm that simulates Gibbsian fields. It is an annealing algorithm at constant temperature. Its law converges to the law π_θ. For it we need to visit each site infinitely often with a finite period.

This leads to define a sequence of sites $(u_k;\ k \in \mathbb{N})$ such that:

$$\exists R \in \mathbb{N} / \forall k \in \mathbb{N}: D \subset \{u_l;\ k+1 \leqq l \leqq k+R\}. \tag{4}$$

We then define a family of transition probabilities on Ω by: for $x, y \in \Omega$,

$$P_\theta^{n, n+1}(x, y) = \chi(x_{u(n)} = y_{u(n)})\, \pi_\theta^{u(n)}(y_{u(n)} \mid \bar{y}_{u(n)}). \tag{5}$$

This expression generally depends only of the values of y_s in a small neibourghood of site $u(k)$.

$\chi(A)$ is the characteristic function of the set A. We will also note $\chi(P)$ if P is any property and it will be equal to 1 if P is true and 0 if not. For any sequence (a_k) we will note a_k as well as $a(k)$.

We now put:

$$\begin{aligned} P_\theta^{n, n+p}(x, y) &= \sum_{z \in \Omega} P_\theta^{n, n+p-1}(x, z)\, P_\theta^{n+p-1, n+p}(z, y) \\ &= \sum_{z \in \Omega} P_\theta^{n, n+1}(x, z)\, P_\theta^{n+1, n+p}(z, y) \end{aligned} \tag{6}$$

and:

$$\begin{gathered} (P_\theta^{n, n+p} \varphi)(y) = \sum_{x \in \Omega} P_\theta^{n, n+p}(y, x)\, \varphi(x), \\ \text{for any } \varphi: \Omega \to \mathbb{R}^d. \end{gathered} \tag{7}$$

We have (*cf.* [3]): $\lim_{p \to \infty} P_\theta^{n, n+p}(x, y) = \pi_\theta(y)$ for all $x \in \Omega$ and all $n \in \mathbb{N}$.

2.4. Estimation algorithm

The stochastic gradient algorithm is defined by:

$$\begin{gathered} \theta_0,\ X_0 \text{ given} \\ \theta_{n+1} = \theta_n + [\alpha(X_{n+1}) - \alpha_0]/[(n+1)\, U] \\ P(X_{n+1} = x \mid X_n = y) = P_\theta^{n, n+1}(y, x) \end{gathered} \tag{8}$$

$U > 0$ is a constant ensuring a. s. convergence. (X_n) is then an inhomogeneous Markov chain.

We can rewrite the algorithm in the standard form:

$$\theta_{n+1} = \theta_n + h(\theta_n)/[(n+1)\, U] - g(\theta_n, X_{n+1})/[(n+1)\, U] \tag{9}$$

with:

$$g(\theta, x) = E_\theta(\alpha(.)) - \alpha(x) \tag{10}$$

$$h(\theta) = E_\theta(\alpha(.)) - \alpha_0. \tag{11}$$

2.5. First estimates

In the following we will use the constants:

$$\varphi_0 = \max\{\| \alpha(x) - \alpha_0 \|;\ x \in \Omega\} \tag{12}$$

$$\mu = \max\{\| \alpha(x) - \alpha(y) \|;\ x \in \Omega,\ s \in D,\ y \in \Omega(x, s)\} \tag{13}$$

μ estimates the variation of α when one changes the value at one site only. It is generally far smaller than φ_0.

We can already note that:

$$\begin{aligned} \| h(\theta) \| &\leqq \varphi_0 \\ \| g(\theta, x) \| &\leqq 2\varphi_0 \\ \| \theta_{n+1} - \theta_n \| &\leqq \varphi_0/[(n+1)U] \end{aligned} \tag{14}$$

And we can deduce the deterministic bound:

$$\| \theta_n \| \leqq \| \theta_0 \| + \varphi_0 (1 + \operatorname{Log} n)/U \tag{15}$$

In the paper we use the matrix norm associated to the euclidian norm. If A is an (n, d) matrice $A = (a_{ij})$ and $|a_{ij}| \leqq K$, we have $\| A \| \leqq (nd)^{1/2} K$. If u is n, 1 and v is 1, n, we have:

$$\| u\, {}^t v \| \leqq \| u \| \, \| v \|;$$

${}^tA$ is the transposed matrice of A.

Finally if f is any function from Ω to $\mathbb{R}^n$ we will call $\| f \| = \max(\| f(x) \|,\ x \in \Omega)$. There will be no risk of confusion between the different norms we use here.

3. PRELIMINAR RESULTS

3.1. Weak ergodicity

3.1.1. *Inhomogeneous Markov chains*

We shall first state a weak ergodicity lemma. Similar results have already been shown by Geman [3] or by Mitra *et al.* [8]. This lemma will be used several times in the following. We prove it here again for the sake of clarity and because we stated it in a larger context than in [3] or [8]. For it we will need some additional notations that will be used again in part 5. We also recall some property of inhomogeneous markov chains that can be found in (Isaacson-Madsen; [6]).

• For a sequence $(\mathbb{P}_n;\ n \geqq 1)$ of stochastic matrices and a row vector $f^0=(f^0(x);\ x\in\Omega)$ with $f^0(x)\geqq 0$ and $\Sigma f^0(x)=1$ (f^0 is an initial law) we note:

$$f^k=f^0\,\mathbb{P}_1\ldots\mathbb{P}_k \qquad \text{and} \qquad f^{m,k}=f^0\,\mathbb{P}_{m+1}\ldots\mathbb{P}_k.$$

If g^0 is an other initial law we note g^k and $g^{m,k}$.

The Markov chain associated to the $\mathbb{P}_n$ is called weakly ergodic if:
for all m, f^0, g^0 :

$$\| f^{m,k}-g^{m,k}\| \to 0 \text{ as } k\to\infty.$$

It will be strongly ergodic if there exists a row vector q with:

$$\forall m,\ \forall f^0:\ \| f^{m,k}-q\| \to 0 \qquad \text{if} \quad k\to\infty.$$

• The ergodic coefficient of a stochastic matrice $\mathbb{P}$ is by definition

$$\Delta(\mathbb{P})=1/2\sup_{x,y}\sum_z |p(x,z)-p(y,z)|=1-\inf_{x,y}\sum_z \min(p(x,z),p(y,z)) \quad (16)$$

where $\mathbb{P}=(p(x,y))$, $\sum_y p(x,y)=1$, $p(x,y)\geqq 0$.

We have:

$$\left.\begin{aligned}\Delta(\mathbb{P}) &\leqq 1\\ \Delta(\mathbb{P}\mathbb{Q}) &\leqq \Delta(\mathbb{P})\,\Delta(\mathbb{Q})\end{aligned}\right\} \quad (17)$$

If we note $\mathbb{P}^{m,k}=\mathbb{P}_{m+1}\ldots\mathbb{P}_k$, the chain is weakly ergodic iff. $\forall m:\ \Delta(\mathbb{P}^{m,k})\to 0$ if $k\to\infty$.

• If there exists a sequence of invariant vectors ψ_n such that $\psi_n \mathbb{P}_n = \mathbb{P}_n$, $\psi_n(x) \geqq 0$, $\sum_x \psi_n(x) = 1$, the chain is strongly ergodic if:

(i) it is weakly ergodic.

(ii) $\sum_n \| \psi_{n+1} - \psi_n \| < \infty$.

• In our context $\mathbb{P}_n = (P_{\theta(n)}^{n, n+1}(x, y);\ x, y \in \Omega)$ where $\theta_n = \theta(n)$ is a sequence of parameters that will be of two types.

(i) $\theta_n = \theta$ for all n.

(ii) $\theta_n = \eta_n / T_n$ where T_n decreases, with $T_n \geqq \mu NA / \text{Log}\, n$ and $\| \eta_n \| \leqq A$ for large enough n. The case (i) is included in case (ii) but we will need more precise results for it.

We can now state the lemma:

3.1.2. LEMMA. — *If* θ_n *is of type* (i) *or* (ii) *the chain is weakly ergodic. In case* (i) *we have for* $m - k \geqq R$:

$$\Delta(\mathbb{P}^{m, k}) \leqq [r(\theta)]^{(m-k)/R - 1}, \tag{18}$$

where $r(\theta) = 1 - \exp(-\mu N \| \theta \|)$ $(r(\theta) < 1)$.

3.1.3. *Proof of lemma* 3.2: The proof is based on two points;

(.) we have:

$$\min \{\pi_\theta^s(x_s \mid \bar{x}_s);\ s \in D\, x \in \Omega\} \geqq \delta_\theta \tag{19}$$

where $\delta_\theta = \exp(-\mu \| \theta \|)/L$; this is obvious.

(. .) let's put:

in case (i)

$$d_n = \delta_\theta$$

in case (ii)

$$d_n = \exp(-\text{Log}\, n/N)/L \qquad \text{if} \quad n \geqq n_0$$

$$d_n = 0 \qquad \text{if} \quad n < n_0$$

For both cases d_n is decreasing and we have $d_n \leqq \delta_{\theta(n)}$ for all n.

We have:

$$\min \{P^{n, n+R}(x, y);\ x, y \in \Omega\} \geqq (d_{n+R})^N \tag{20}$$

indeed, let's note: $m_s = \max\{k,\ k \leqq n + R - 1 \text{ and } u_n = s\}$. We order D: $D = \{s_1, \ldots, s_N\}$ with $n \leqq m_{s(1)} \leqq \ldots \leqq m_{s(N)} = n + R - 1$.

We then have:

$$\min_{x,y} P^{n,n+R}(x,y) = \min_{x,y} \Big(\sum_{z,z'} P^{n,m_s(1)}(x,z) \times P^{m_s(1),m_s(1)+1}(z,z') P^{m_s(1)+1,n+R}(z',y) \Big) \geqq d_{n+R} \inf_{y,z'} (P^{m_s(1)+1,n+R}(y,z'))$$

because $P^{m_s(1),m_s(1)+1}(z,z') \geqq d_{m_s(1)} \geqq d_{n+R}$.

We get (20) by doing this again for $s_2, \ldots, s_N$.

We now have:

$$\Delta(P^{n,n+R}) = 1 - \inf_{x,y} \sum_{z} \min(P^{n,n+R}(x,z) P^{n,n+R}(y,z)) \geqq 1 - \sum_{z \in \Omega} (d_{n+R})^N = 1 - \Lambda d_{n+R}^N \qquad [\Lambda = \mathrm{card}(\Omega)].$$

In case (i):

$$(d_{n+R})^N = \delta_\theta^N = e^{(-\mu N \| \theta \|)}/L^N$$

and hence:

$$\Delta(P^{n,n+R}) \leqq 1 - e^{-\mu N \| \theta \|} < 1$$

so that we can conclude with the help of properties (17); (since $P^{n,n+kR} = P^{n,n+R} \ldots P^{n+(k-1)R,n+kR}$).

In case (ii) we obtain:

for $n \geqq n_0$,

$$\Delta(P^{n,n+R}) \leqq (1 - 1/(n+R)).$$

Properties (17) lead us again to the desired result.

3.2. Poisson equation

3.2.1. *Introduction*

In [7] Métivier and Priouret have presented a method for showing convergence for markovian stochastic (gradient) algorithms. In particular they introduced a solution to a Poisson equation related to the markov chain. This is the method we will use here, with the slight difference that we deal with inhomogeneous Markov chains. It is easy to see that Métivier

and Priouret's results generalize to the inhomogeneous case. But these results do not enable us to obtain the almost sure convergence of (θ_n); we shall need more precise estimates for it. However, we first show the existence of that solution and give some properties. We will here only apply case (i) of lemma 3.1.2.

3.2.2. LEMMA. — *Let* $\zeta(\theta, x)$: $\mathbb{R}^v \times \Omega \to \mathbb{R}^d$ *verify:*

$$E_\theta[\zeta(\theta, .)] = 0 \qquad \textit{for all} \quad \theta$$

Then for each θ *there exist functions* ρ_θ^k, $k \in \mathbb{N}$ *from* Ω *to* $\mathbb{R}^d$ *such that:*

$$\rho_\theta^k(y) - (P_\theta^{k, k+1} \rho_\theta^{k+1})(y) = \zeta(\theta, y). \tag{19}$$

If $\partial\zeta/\partial\theta$ *exists then* $\partial\rho/\partial\theta$ *exists.*

3.2.3. *Proof.* — Let's first remark that:

$$\pi_\theta(.) = \sum_{y \in \Omega} \pi_\theta(y) P_\theta^{n, n+p}(y, .) \tag{20}$$

so that $\pi_\theta(.)$ is an invariant vector for the chain P_θ. This is easy to check.

We define

$$\rho_\theta^n(x) = \sum_{p \geq 0} P_\theta^{n, n+p}(\zeta(\theta, .))(x)$$

and

$$\sigma_{k, l}(x) = \sum_{k \leq p \leq l} P_\theta^{n, n+p}(\zeta(\theta, .))(x).$$

By (7) and (20) we have:

$$\sigma_{k, l}(x') = \sum_{k \leq p \leq l} \sum_{x \in \Omega} \zeta(\theta, x) \sum_{x'' \in \Omega} (P_\theta^{n, n+p}(x', x) - P_\theta^{n, n+p}(x'', x)) \pi_\theta(x'').$$

So according to lemma 3.1.2 for $k \geq R$:

$$\| \sigma_{k, l} \| \leq 2 M_\theta(\zeta) \sum_{k \leq p \leq l} \Delta(P_\theta^{n, n+p}) \leq 2 M_\theta(\zeta) r(\theta)^{(k/R)-1} / (1 - r(\theta)^{1/R}) \tag{21}$$

where $M_\theta(\zeta) = \max \{ \| \zeta(\theta, x), x \in \Omega \}$.

This shows that $\sigma_{k, l}$ converges to 0 if $k, l \to \infty$ and then the series in the definition of ρ_θ^n converges.

Let's now assume that $\partial\zeta/\partial\theta$ exists; we want to show that ρ is differentiable in θ. We have

$$\rho_\theta^n(x') = \sum_{p \geq 0} \sum_{x, x''} \zeta(\theta, x)(P^{n, n+p}(x', x) - P^{n, n+p}(x'', x))\pi_\theta(x'') \quad (22)$$

We will differentiate each term in the series in (22), and show that the obtained series converges uniformely on each compact set of $\mathbb{R}^v$.

We have:

$$\begin{aligned}\partial[\zeta(\theta, x)(P^{n, n+p}(x', x) - P^{n, n+p}(x'', x))\pi_\theta(x'')]/\partial\theta \\ = \pi_\theta(x'')(P_\theta^{n, n+p}(x', x) - P_\theta^{n, n+p}(x'', x))\partial(\zeta(\theta, x)/\partial\theta \\ + (P_\theta^{n, n+p}(x', x) - P_\theta^{n, n+p}(x'', x))\zeta(\theta, x)\,{}^t[\partial\pi_\theta(x'')/\partial\theta] \\ + \pi_\theta(x')\zeta(\theta, x)\,{}^t[\partial P_\theta^{n, n+p}(x', x)/\partial\theta - \partial P_\theta^{n, n+p}(x'', x)/\partial\theta]. \quad (23)\end{aligned}$$

This leads to the study of three series, one for each term of (23). By using the lemma 3.1.2 we can easily see that the first two converge. The third one leads to a little more calculation. We shall first study $\partial P_\theta^{n, n+p}(y, x)/\partial\theta$.

We have:

$$\begin{aligned}\partial P_\theta^{n, n+1}(x, y)/\partial\theta = \chi(\bar{x}_{u(n)} = \bar{y}_{u(n)})\,\partial\pi_\theta^{u(n)}(y_{u(n)} \mid \bar{y}_{u(n)})/\partial\theta \\ = \gamma_{u(n)}(y)\,P_\theta^{n, n+1}(x, y) \quad (24)\end{aligned}$$

where

$$\gamma_k(y) = \partial \mathrm{Log}(\pi_\theta^k(y_k \mid \bar{y}_k))/\partial\theta = E_\theta^k(\alpha(.) \mid \bar{y}_k) - \alpha(y).$$

Then:

$$P_\theta^{n, n+p}(x, y) = \sum_{x_1 \ldots x_{P-1}} P_\theta^{n, n+1}(x, x_1) \ldots P_\theta^{n+p-1, n+p}(x_{p-1}, y)$$

so that:

$$\begin{aligned}\partial P_\theta^{n, n+p}(x, y)/\partial\theta = \sum_{x_1 \ldots x_{P-1}} \sum_{1 \leq k \leq p} [P_\theta^{n, n+1}(x, x_1) \times \ldots \\ \times P_\theta^{n+p-1, n+p}(x_{p-1}, y)\,\gamma_{u(n+k-1)}(x_k)]\end{aligned}$$

(with $x_p = y$).

We arrange this sum in the following way: $\sum_k \sum_{x_k} \sum_{x_1 \dots x_{k-1}} \sum_{x_{k+1} \dots x_{p-1}}$ to obtain:

$$\partial P_\theta^{n, n+p}(x, y)/\partial\theta = \sum_{1 \leqq k \leqq p} \sum_{x_k} P_\theta^{n, n+k}(x, x_k) \times P_\theta^{n+k, n+p}(x_k, y)\, \gamma_{u(n+k-1)}(x_k) \quad (25)$$

(for $k=p$ we have the convention $P_\theta^{n+p, n+p}(x_p, y) = \chi(x_p = y)$).

We need then to show that the series with general term:

$$Z_p = \sum_{x, x''} \pi_\theta(x')\, \zeta(\theta, x)\, {}^t[\partial P_\theta^{n, n+p}(x', x)/\partial\theta - \partial P_\theta^{n, n+p}(x'', x)/\partial\theta]$$

converges. Using (25) we can write:

$$Z_p = \sum_{1 \leqq k \leqq p} \sum_{y,\, x''} \pi_\theta(x'')(P_\theta^{n, n+k}(x', y) - P_\theta^{n, n+k}(x'', y)) \left(\sum_x \zeta(\theta, x)\, P_\theta^{n+k, n+p}(y, x)\right) {}^t\gamma_{u(n+k-1)}(y).$$

Using our assumption on ζ we can replace $\sum \zeta(\theta, x)\, P_\theta^{n+k, n+p}(y, x)$ by:

$$\sum_x \zeta(\theta, x)(P_\theta^{n+k, n+p}(y, x) - \pi_\theta(x))$$

which can be estimated by: $2\, M_\theta(\zeta)\, r(\theta)^{[(p-k)/R]-1}$ for $p-k \geqq R$ according to (20) and to lemma 3.1.2. Using again lemma 3.1.2, we see that we can estimate Z_p by:

$$K M_\theta(\zeta)\, p r(\theta)^{(p/R)-2} \qquad \text{for} \quad p \geqq 2R$$

where K is independent of θ and p. This proves the convergence of Z_p and then the differentiability of ρ in θ.

3.2.4. *Application to g*

We have $g(\theta, x) = E_\theta(\alpha(.)) - \alpha(x)$ so that $E_\theta(g(\theta, .)) = 0$. We also see that g and its differential are bounded independently of θ. We have then proved the existence of functions v_0^k that verify:

$$v_\theta^k(y) - (P^{k, k+1} v_\theta^{k+1})(y) = g(\theta, y) \quad \text{for all } y, \theta, k. \quad (26)$$

If we look at the preceding proof with a little care, we can find estimates for the v and their derivatives; we have:

$$\| v_\theta^k \| \leqq K_1 (1 + [1 - r(\theta)^{1/R}]^{-1}) = C_1(\theta) \tag{27}$$

[use (22) and the above remarks]. Here K_1 is a constant independant of k and θ.

$$\| \partial v_\theta^k / \partial \theta \| \leqq K_2 (1 + [1 - r(\theta)^{1/R}]^{-2}) = C_2(\theta) \tag{28}$$

[use (24), and the estimate of Z_P].

Let's go back to the algorithm (8); because of (15), (27), (28) and the expression of $r(\theta)$ given in lemma 3.1.2 we can find *a priori* estimates for $C_1(\theta_n)$ and $C_2(\theta_n)$:

we have

$$r(\theta_n) = 1 - \exp(-\mu N \| \theta_n \|) \leqq 1 - (\text{Cte}) \exp(-\mu N \varphi_0 \operatorname{Log} n/U)$$
$$= 1 - (\text{Cte})\, n^{-\mu N \varphi_0 / U} = M_n$$

then:

$$C_1(\theta_n) \leqq K_1 (1 + [1 - M_n^{1/R}]^{-1} = \beta_n \tag{29}$$

and we can see that there is a constant C such that β_n is equivalent to $C n^{\mu N \varphi_0 / U}$ if $n \to \infty$.

In the same way, we get

$$C_2(\theta_n) \leqq \beta'_n \qquad \text{with} \quad \beta'_n \tag{30}$$

equivalent to $C' n^{2 \mu N \varphi_0 / U}$, where C' is another constant.

We now state the convergence theorem.

4. ALMOST SURE CONVERGENCE OF THE ESTIMATION ALGORITHM

4.1. THEOREM. — Let's consider the algorithm given in (8):

$$\theta_{n+1} = \theta_n + [\alpha(X_{n+1}) - \alpha_0]/[(n+1)\, U]$$

where X_{n+1} is obtained from X_n according to the method of the Gibbs sampler (*see* 2.3):

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

Let θ_* be the solution of $E_\theta[\alpha(.)]=\alpha_0$.

We have the following convergence result:

If $U > 2\mu N\varphi_0$ then $\theta_n \to \theta_*$ a. s. and there exists an $\varepsilon_0 > 0$ such that a. s.: $\|\theta_n-\theta_*\|=o(n^{-\varepsilon})$ for $\varepsilon < \varepsilon_0$.

4.2. *Proof.* — The proof is in two steps; the first one deals with the stochastic part of the algorithm; it uses some methods exposed in [7] (proposition 3.1) but requires some care because we do not have *a priori* constant bounds for the v_θ. We obtain here a stronger result than the one we would have had by applying directly the theorems in [7], which estimates the probability of non convergence of (θ_n).

The second part of the proof uses deterministic estimates to reach the conclusion. There are some technical calculations that are needed to obtain the second fact of the theorem.

4.2.1. *First step*

Let $a=\mu N\varphi_0/U$ and $\varepsilon_1=1/2-a$; by assumption we have $\varepsilon_1 > 0$. We will first show that for almost all trajectory, (X_n), and for all $\varepsilon < \varepsilon_1$:

$$\sum_{0 \leqq k \leqq n} < g(\theta_k, X_{k+1}), \theta_k-\theta_* > /(k+1)^{1-\varepsilon} \tag{31}$$

converges if $n \to \infty$

One can see that, if (31) is true for one value of ε, then it is true for any smaller value. So, by considering ε of the kind ε_1-1/n, one only needs to prove that (31) is true for almost all trajectory, with fixed ε.

Let's put

$$S_{n,m}=\sum_{n \leqq k \leqq m} < g(\theta_k, X_{k+1}), \theta_k-\theta_* > /[(k+1)^{1-\varepsilon}]$$

Let's note $w_\theta^n=\langle v_\theta^n, \theta-\theta_* \rangle$ and $\lambda_k=(k+1)^{-(1-\varepsilon)}$; by (26) we have:

$$S_{n,m}=\sum_{n \leqq k \leqq m} \lambda_k[w_{\theta_{(k)}}^{k+1}(X_{k+1})-(P_{\theta_{(k)}}^{k+1,k+2} w_{\theta_{(k)}}^{k+2})(X_{k+1})].$$

We cut $S_{n,m}$ in S_1+S_2: $S_{n,m}=S_1+S_2$ where:

$$S_1=\sum_{n\leqq k\leqq m}\lambda_k[w^{k+1}_{\theta(k)}(X_{k+1})-P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)]$$

and

$$S_2=\sum_{n\leqq k\leqq m}\lambda_k[P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)-P^{k+1,k+2}_{\theta(k)}w^{k+2}_{\theta(k)}(X_{k+1})].$$

Let's consider now the Markov process X_n on $\Omega^{\mathbb{N}}$. We call $\mathcal{F}_n$ the σ-algebra generated by $X_1, \ldots, X_n$ and we note $\mathbb{E}$ the expectation according to the law of this process. θ_k is $\mathcal{F}_k$ mesurable and we have:

$$\mathbb{E}[w^{k+1}_{\theta(k)}(X_{k+1})\mid\mathcal{F}_k]=P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k).$$

So,

$$\mathcal{Q}_n=\sum_{0\leqq k\leqq n}\lambda_k[w^{k+1}_{\theta(k)}(X_{k+1})-P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)]$$

is an $\mathcal{F}$-martingale where $\mathcal{F}=(\mathcal{F}_n, n\in\mathbb{N})$, moreover we have:

$$\mathbb{E}(\|\mathcal{Q}_n\|^2)=\sum_{0\leqq k\leqq n}\lambda_k^2\,\mathbb{E}[\|w^{k+1}_{\theta(k)}(X_{k+1})-P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)\|^2]$$

$$\leqq 4\sum_{0\leqq k\leqq n}\beta_k^2\|\theta_k-\theta_*\|^2\lambda_k^2 \quad [\text{by (29)}]$$

β_k is equivalent to Ck^a when $k\to\infty$; moreover, $\|\theta_k-\theta_*\|\leqq C''\ \mathrm{Log}\,k$ where C'' is a constant. We then have:

$$\beta_k^2\|\theta_k-\theta_*\|^2\lambda_k^2\leqq\beta_k^2 C''^2\,\mathrm{Log}^2 k\,\lambda_k^2\ (\mathrm{cte}\ \mathrm{Log}^2 k\,.\,k^{2-2(\varepsilon+a)})$$

and this is the general term of a convergent series, because we assumed that $\varepsilon<1/2-a$. Then $\mathbb{E}(\|\mathcal{Q}_n\|^2)$ is bounded, $\mathcal{Q}_n$ converges a. s. and $S_1\to 0$ a. s. if $n, m\to\infty$.

We can write S_2:

$$S_2=\sum_{n\leqq k\leqq m}\lambda_k P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)-\sum_{n+1\leqq k\leqq m+1}\lambda_{k-1}P^{k,k+1}_{\theta(k-1)}w^{k+1}_{\theta(k-1)}(X_k)$$

then:

$$S_2=\lambda_n P^{n,n+1}_{\theta(n)}w^{n+1}_{\theta(n)}(X_n)-\lambda_m P^{m+1,m+2}_{\theta(m)}w^{m+2}_{\theta(m)}(X_{m+1})$$
$$+\sum[\lambda_k P^{k,k+1}_{\theta(k)}w^{k+1}_{\theta(k)}(X_k)-\lambda_{k-1}P^{k,k+1}_{\theta(k-1)}w^{k+1}_{\theta(k-1)}(X_k)] \quad (32)$$

But $\| w^{k,k+1}_{\theta(k)} \| \leqq \beta_k \| \theta_k - \theta_* \| \leqq \text{Cte}\, k^a \operatorname{Log} k$. We can deduce from this that the first two terms of (32) vanish if $m, n \to \infty$. We decompose the sum in (32) into several sums whose general terms are:

$$\xi^1_k = [P^{k,k+1}_{\theta(k)} w^{k+1}_{\theta(k)}(X_k)](\lambda_k - \lambda_{k-1})$$

$$\xi^2_k = [P^{k,k+1}_{\theta(k)} w^{k+1}_{\theta(k)}(X_k) - P^{k,k+1}_{\theta(k-1)} w^{k+1}_{\theta(k)}(X_k)]\lambda_{k-1}$$

$$\xi^3_k = \langle P^{k,k+1}_{\theta(k-1)} v^{k+1}_{\theta(k)}(X_k) - P^{k,k+1}_{\theta(k-1)} v^{k+1}_{\theta(k-1)}(X_k),\ \theta_k - \theta_* \rangle \lambda_{k-1}$$

$$\xi^4_k = \langle P^{k,k+1}_{\theta(k-1)} v^{k+1}_{\theta(k-1)}(X_k),\ \theta_k - \theta_{k-1} \rangle \lambda_{k-1}.$$

We have:

- $\| \xi^1_k \| \leqq \text{Cte}\, \beta_k \operatorname{Log} k\ (\lambda_k - \lambda_{k-1})$ and

$$\lambda_{k-1} - \lambda_k = \{1 - [k/(1 - 1/(k+1))]^{1-\varepsilon}\}/k^{1-\varepsilon}(1-\varepsilon)/k^{2-\varepsilon}$$

so that $\beta_k \operatorname{Log} k\,(\lambda_k - \lambda_{k-1})$ Cte $\operatorname{Log} k/k^{2-\varepsilon-a}$ $2-\varepsilon-a > 1$ and hence $\sum \xi^1_k$ converges

On the other hand:

- for any ψ:

$$\| P^{k,k+1}_{\theta(k)} \psi - P^{k,k+1}_{\theta(k-1)} \psi \| \leqq \sum \| \psi(x) \| \| P^{k,k+1}_{\theta(k)}(.,x) - P^{k,k+1}_{\theta(k-1)}(.,x) \|$$

$$\leqq \| \psi \| \mu \| \theta_k - \theta_{k-1} \|$$

(*See* (24), we have $\| \gamma_k \| \leqq \mu$).

So: $\| \xi^2_k \| \leqq \text{Cte}\, \beta_k \| \theta_k - \theta_* \| (\varphi/U)/k^{2-\varepsilon} \leqq \text{Cte} \operatorname{Log} k/k^{2-\varepsilon-a}$ and $\sum \xi^2_k$ converges.

- Because of the bounds on the derivatives of v_θ we have

$$\| \xi^3_k \| \leqq \beta'_k \| \theta_k - \theta_{k-1} \| \| \theta_k - \theta_* \| / k^{1-\varepsilon} \leqq \text{Cte}\, \beta'_k \operatorname{Log} k/k^{2-\varepsilon}$$

from $\beta'_k \leqq \text{Cte}\, k^{2a}$ we get: $\| \xi^3_k \| \leqq \text{Cte}\ \operatorname{Log} k/k^{2-\varepsilon-2a} \leqq 2-\varepsilon-2a > 1$ and $\sum \xi^3_k$ converges.

- Finally $\| \xi^4_k \| \leqq \text{Cte}\, \beta_k/k^{2-\varepsilon}$ and $\sum \xi^4_k$ converges.

We have then shown that $S_2 \to 0$ if $n, m \to \infty$ and which concludes the first step of the proof.

4.2.2. *End of the proof*

Recall that $h(\theta)=E_\theta(\alpha)-\alpha_0$ is the differential of a concave function; we then have:

$$\langle h(\theta),\ \theta-\theta_*\rangle < 0 \quad \text{for all } \theta.$$

Moreover, if θ is in a convex compact set, Q, such that $\theta_*\in Q$ we have for $\theta\in Q$:

$$\langle h(\theta),\ \theta-\theta_*\rangle \leqq -b(Q)\,\|\theta-\theta_*\|^2 \qquad (33)$$

where $b(Q)=\inf \|\operatorname{Var}_\theta(\alpha)\|$ for $\theta\in Q$.

We note $b_*=\|\operatorname{Var}_{\theta_*}(\alpha)\|$. We will show that we can take $\varepsilon_0=\min(\varepsilon_1/2,\ b_*/U)$.

The estimates we are going to make in the following depends on the trajectories of the process (X_n). So, we now assume that we are given a trajectory, and thus that the X_n are fixed—such that (31) is true for all $\varepsilon<\varepsilon_1$.

Let n_0 be a positive integer and $b\geqq 0$ and let's assume that for all $n\geqq n_0$ θ_n is in a convex set Q such that $\theta_*\in Q$ and for all $\theta\in Q$, $\|\operatorname{var}_\theta(\alpha)\|\geqq b$ (we can always take $b=0$).

We have

$$\theta_{n+1}-\theta_*=\theta_n-\theta_*+h(\theta_n)/[(n+1)U]-g(\theta_n,\ X_{n+1})/[(n+1)U].$$

and then:

$$\|\theta_{n+1}-\theta_*\|^2 \leqq [1-2b/((n+1)U)]\,\|\theta_n-\theta_*\|^2$$
$$-2\langle g(\theta_n,\ X_{n+1}),\ \theta_n-\theta_*\rangle/[(n+1)U]+\varphi_0^2U^2/(n+1)^2.$$

[*Cf.* (32) and (14).]

Let ε be a positive number with $\varepsilon<\varepsilon_1/2$.

Let's put: $\omega_n=n^\varepsilon\,\|\theta_n-\theta_*\|$ and

$$V_{k,n}=[(n+1)/(k+1)]^{2\varepsilon}\prod_{l=k}^{n}[1-2b/(U(l+1))].$$

We have for $n \geqq n_0$:

$$\omega_{n+1}^2 \leqq V_{n_0-1,n}\omega_{n_0}^2 - 2/U \sum_{k=n_0}^{n} V_{k,n} \langle g(\theta_k, X_{k+1}), \theta_k - \theta_* \rangle / (k+1)^{1-2\varepsilon} + \varphi_0^2/U^2 \sum_{k=n_0}^{n} V_{k,n}/(k+1)^{2-2\varepsilon} \quad (34)$$

One can easily prove the following facts:

If $b > \varepsilon U$:

- $V_{k,n} \to 0$ if $n \to \infty$ (35)
- there exists an $n_1 > 0 / V_{k-1,n} \leqq V_{k,n}$ for all $k \geqq n_1$ and $n \geqq k$.

Let η be a positive number.

The first step of the proof shows that:

$\exists p_0 / \forall p \geqq p_0$:

$$\left\| 2/U \sum_{k=p_0}^{p} \langle g(\theta_k, X_{k+1}), \theta_k - \theta_* \rangle / (k+1)^{1-2\varepsilon} \right\| < \eta$$

and

$$\left\| \sum_{k=p_0}^{p} (\varphi_0/U)^2/(k+1)^{2-2\varepsilon} \right\| < \eta$$

(we have $2\varepsilon < \varepsilon_1$ and then $2-2\varepsilon > 1$).

If we apply (34) with $\varepsilon = b = 0$ we see that θ_n is bounded, so for any choice of n_0, we can take Q compact and assume $b > 0$.

We now assume that $\varepsilon < b/U$ and take $p_0 \geqq \max(n_0, n_1)$.

According to (35) for large enough n, the quantity:

$$\left\| V_{n_0-1,n}\omega_{n_0}^2 - 2/U \sum_{k=n_0}^{p_0-1} V_{k,n} \langle g(\theta_k, X_{k+1}), \theta_k - \theta_* \rangle / (k+1)^{1-2\varepsilon} + \varphi_0^2/U^2 \sum_{k=n_0}^{p_0-1} V_{k,n}/(k+1)^{2-2\varepsilon} \right\|$$

is smaller than η.

Let

$$B_n = 2/U \sum_{k=p_0}^{n} V_{k,n} \langle g(\theta_k, X_{k+1}), \theta_k - \theta_* \rangle / (k+1)^{1-2\varepsilon}$$

and

$$U_n = 2/U \sum_{k=p_0}^{n} \langle g(\theta_k, X_{k+1}), \theta_k - \theta_* \rangle / (k+1)^{1-2\varepsilon}$$

we have $B_n = \sum_{k=p_0}^{n} V_{k,n}(U_k - U_{k-1})$ and

$$\|B_n\| \leqq \|V_{n,n} U_n\| + \sum_{k=p_0}^{n-1} \|U_k\| \, |V_{k,n} - V_{k+1,n}|$$

$$\leqq \eta + \eta \sum_{k=p_0}^{n-1} (V_{k+1,n} - V_{k,n})$$

$$\leqq \eta + \eta (V_{n-1,n} - V_{p_0,n}).$$

We have $|V_{p_0,n}| < |V_{n-1,n}| < |V_{n,n}| = 1$ and then $\|B_n\| \leqq 3\eta$.

If we put:

$$B'_n = \varphi_0^2 / U^2 \sum^{n} V_{k,n} / (k+1)^{2-2\varepsilon}$$

and

$$k = p_0$$

$$U'_n = \varphi_0^2 / U^2 \sum_{k=p_0}^{n} 1/(k+1)^{2-2\varepsilon}$$

we can make the same kind of estimates and show that $\|B'_n\| \leqq \text{Cte}\ \eta$.

So, we have shown that if n_0 is given, and Q is a convex compact set with $\theta_* \in Q$ and $\theta_n \in Q$ for all $n \geqq n_0$, then: for $\varepsilon < \min(\varepsilon_1/2, b(Q)/U)$, $\omega_n \to 0$ a. s.

We now know that $\theta_n \to \theta_*$ a. s. So by choosing n_0 large enough, we can take Q as near to θ_* as we want and then choose $b(Q)$ to be any number $> b_*$.

So we can take any $\varepsilon < \varepsilon_0 = \min(\varepsilon_1/2, b_*/U)$.

The proof of theorem 4.1 is complete.

5. ANNEALING

5.1. Goals

As announced in the introduction, we now examine the possibility of coupling estimation and annealing. The exponential case ($H(\theta, x) = \langle \alpha(x), \theta \rangle$) is the one of main interest and it is the one we have studied since the beginning of this paper; so, in order to keep the same notations, we will state the result in this case and give a detailed proof of it. We will then show how it can be generalized to other energy functions and give the main ideas of the proof. So, until part 5.5 we still keep in the exponential case.

A classic annealing algorithm boils down to defining a markovian process Y_n with values in Ω, by:

$$P[Y_{n+1} = y \mid Y_n = x] = P^{n, n+1}_{\theta_*/T_n}(x, y). \tag{36}$$

(as explained in part 1 we take at each step the conditional probability associated to the Gibbsian field of energy $H(\theta_*, .)/T_n = \langle \alpha(.), \theta_*/T_n \rangle$).

Let's note M_θ for the set where $\pi_\theta(.)$ is maximum in the present case:

$$M_\theta = \{x / \forall y \in \Omega, \langle \alpha(x) - \alpha(y), \theta \rangle \leqq 0\}.$$

We will put M_* for M_{θ_*}. We know [3] that (36) defines a strongly ergodic markov chain as soon as T_n decreases slowly enough to 0; the limit law is the uniform law on M_*.

We wonder here wether we can replace in (36) θ_* by θ_n where θ_n converges to θ_*. We would like it to be true in particular for the sequence obtained with the estimation algorithm described in this paper. The following theorem gives this result, under some additional hypothesis. To have lighter expressions we will call η_n the sequence that converges to θ_* and $\theta_n = \eta_n/T_n$.

5.2 THEOREM. — *Let $\eta_n \in \mathbb{R}^\nu$ be a sequence that converges to θ_*. We put $\theta_n = \eta_n/T_n$ ($T_n \geqq 0$) and we define a Markov process Y_n by:*

$$P[Y_{n+1} = y \mid Y_n = x] = P^{n, n+1}_{\theta(n)}(x, y). \tag{37}$$

• We assume that there exists constants $C > 0$, $\varepsilon > 0$, $A > \|\theta_*\|$ such that:

$$\begin{aligned} \|\eta_{n+1} - \eta_n\| &\leqq C/(n+1) \\ \|\eta_n - \theta_*\| &\leqq C n^{-\varepsilon} \\ T_n &= A\,\mu\,N/\mathrm{Log}\,n \end{aligned} \tag{H1}$$

• We also assume:

$$\mathrm{card}\,(\alpha(M_*)) = 1. \tag{H2}$$

Then the chain defined by (37) is strongly ergodic and its limit law is the uniform law on M_*, noted π_*.

5.3. Remarks on the hypothesis

The conditions H1 on η_n are verified by the sequence of the estimation algorithm we propose here. They are also true for gradient type algorithm for which the step is of the same order as $1/n$.

The condition on T_n is the one that is usually given in annealing algorithm. This condition needs only to be true for large enough n, and the conclusion of the theorem are still valid if we assume: $T_n \geqq A\,\mu\,N/\mathrm{Log}\,n$ provided that $1/T_n$ has small oscillations i. e.:

$$(1/T_{n+1} - 1/T_n) = O\,(1/n).$$

The condition H2 is a stability hypothesis on M_θ for θ near θ_*. One can easily see that it is equivalent to $M_\theta = M_*$ in the neighbourghood of θ_*. The set of θ for which H2 is false is negligible in $\mathbb{R}^\nu$ because it is included in the union of the hyperplans orthogonal to the $\alpha(x) - \alpha(y)$ for $x, y \in \Omega$ and $x \neq y$. In addition it is always true for $\theta \in \mathbb{R}\,(\nu = 1)$; this enables us to see that the result for the algorithm (36) is a consequence of theorem 5.2 (put $\eta_n = 1$ for all n).

5.4. Proof of theorem 5.2

We can first remark that θ_n satisfies to the conditions given in lemma 3.1.2, case (ii) and hence the chain defined by (37) is strongly ergodic. As the $\pi_{\theta(n)}$ are invariant vectors for this chain we only need to

show (*cf.* [6]):

$$\forall x \in \Omega : \sum_n \left| \pi_{\theta_{(n)}}(x) - \pi_{\theta_{(n+1)}}(x) \right| < \infty \tag{38}$$

[it is easy to see from H1 that $\lim_n \pi_{\theta_{(n)}}(x) = \pi_*(x)$].

Because of H2 we have $M_{\eta_n} = M_*$ for large enough n and then $M_{\theta_{(n)}} = M_*$ for large enough n (we have $M_{l\theta} = M_\theta$ for any θ and any $l > 0$). To show (38) we will use the differential of $\pi_\theta(x)$ in θ which is:

$$d\pi_\theta(x)/d\theta = \{E_\theta[\alpha(.)] - \alpha(x)\}\, \pi_\theta(x)$$

and then:

$$\left| \pi_{\theta_{(n)}}(x) - \pi_{\theta_{(n+1)}}(x) \right| \leqq \left\| E_{\tau_n}[\alpha(.)] - \alpha(x) \right\| \pi_{\tau_n}(x) \left\| \theta_{n+1} - \theta_n \right\| \tag{39}$$

where $\tau_n \in [\theta_n, \theta_{n+1}]$.

Conditions H1 show us that $\left\| \theta_{n+1} - \theta_n \right\| = O(\operatorname{Log} n/n)$ and this implies that

$$\left\| \tau_n \right\| \geqq K \log n \quad \text{where K is a positive constant.} \tag{40}$$

This implies that there exists a positive constant $c > 0$ with $\pi_{\tau_n}(x) \leqq n^{-c}$ for all $x \notin M_*$. Indeed we have:

$$\pi_\theta(x) = \{ \sum_{y \in \Omega} \exp[\langle \alpha(x) - \alpha(y), \theta \rangle] \}^{-1} \leqq \{ \sum_{y \in M_*} \exp[\langle \alpha(x) - \alpha(y), \theta \rangle] \}^{-1}.$$

Let Q be a compact neighborhood of θ_* such that $M_\theta = M_*$ for $\theta \in Q$. Let $\rho = \inf \{ \langle \alpha(x) - \alpha(y), \theta / \left\| \theta \right\| \rangle,\ x \notin M_*,\ y \in M_*,\ \theta \in Q \}$ (we can notice that H2 implies $\theta_* \neq 0$).

For $\theta \in Q$ and $l > 0$ we have then $\pi_{l\theta}(x) \leqq \exp(-l\rho)$. If n is large enough to have $\eta_n \in Q$ we use (40) to get:

$$\pi_{\tau_n}(x) \leqq n^{-\rho K}; \text{ take } c = \rho K.$$

We can know use (39) and H1 to see that (38) is true for $x \notin M_*$

Consider now $x \in M_*$ and let θ be such that $M_\theta = M_*$. We have:

$$E_\theta(\alpha) - \alpha(x) = \sum_{y \in \Omega} [\alpha(y) - \alpha(x)]\, \pi_\theta(y) = \sum_{y \notin M_*} [\alpha(y) - \alpha(x)]\, \pi_\theta(y)$$

[because of condition H2, for all $y \in M_*$ we have $\alpha(x) = \alpha(y)$]. This implies that:

$$\| E_{\tau_n}(\alpha) - \alpha(x) \| \leqq \varphi_0 N n^{-c}$$

and we obtain (38) for $x \in M_*$.

This ends the proof of theorem 5.2.

5.5. Extension to nonexponential cases

One can see that theorem 5.2 can be extended to more general energy functions. Here we don't have anymore: $H(\theta, .)/T_n = H(\theta/T_n, .)$ so notations will be more complicated.

We define the probabilities and conditional probabilities $\pi_{\theta, t}$ as in (1) and (2), replacing $\langle \alpha(\theta), . \rangle$ by $t H(\theta, .)$ and then the analog of (5) gives us transition probabilities on Ω, noted $P^{n, n+1}_{\theta, t}(x, y)$. So, the Markov chain associated to the modified annealing will have transition probabilities:

$$P^{n, n+1}_{\eta_n, t_n}(x, y) \text{ with } t_n = 1/T_n.$$

We finally note $\mu(\theta) = \max \{ | H(\theta, x) - H(\theta, y) | \; x, y \in \Omega \}$.

For this chain one has the same results than theorem 5.2 with the hypothesis:

$H(\theta, .)$ is twice continuously differentiable in θ.

H1′: the sequence θ_n is the same as in 5.2

$$T_n = AN/\mathrm{Log}\, n \text{ with } A > \mu(\theta_*).$$

H2′: for all $x, y \in M_*$: $dH(\theta, x)/d\theta = dH(\theta, y)/d\theta$.

In the exponential case H2′ reduces to H2. It says that if θ is near θ_*, the differences between the $H(\theta, x)$ when x varies in M_* is an $O(\| \theta - \theta_* \|^2)$.

To show this we must extend lemma 3.1.2, part (ii) to the non exponential case. The hypothesis on T_n becomes: $T_n \geqq AN/\mathrm{Log}\, n$ with $A > \mu(\eta_n)$ for $n \geqq n_0$. The proof is the same as in the exponential case; just replace δ_θ by $\exp(-\mu(\theta))/L$.

One proves strong ergodicity as in 5.4, the basis of the proof is still a differentiation of $\pi_{\theta, t}(x)$ which will be done here in $\Theta = (\theta, t)$. We obtain,

setting $\pi_n(x) = \pi_{\theta(n), t(n)}(x)$:

$$|\pi_n(x) - \pi_{n+1}(x)| \leqq \text{Cte} \max \{ \| E_{\Theta_n}(dH(\tau_n, .)/d\theta) - dH(\tau_n, x)/d\theta \| u_n \| \theta_{n+1} - \theta_n \|, \quad (41)$$
$$|E_{\theta_n}[H(\tau_n, .)] - H(\tau_n, x)| | t_{n+1} - t_n | \} \pi_{\theta_n}(x)$$

where $\Theta_n = (\tau_n, u_n)$ is in the segment of $\mathbb{R}^{\nu+1}$: $[(\theta_n, t_n), (\theta_{n+1}, t_{n+1})]$.

We then have to notice that if $x \notin M_*$ then $x \notin M_\theta$ for θ near θ_* and that we still have $\pi_{\theta(n), t(n)}(x) \leqq n^{-c}$ for these x and large enough n (c is a positive constant). For $x \in M_*$ we use Taylor formula and H2 to find:

If $x, y \in M_*$:

$$|H(\theta, x) - H(\theta, y)| = O(\| \theta - \theta_* \|^2)$$

and

$$\| dH(\theta, x)/d\theta - dH(\theta, y)/d\theta \| = O(\| \theta - \theta_* \|)$$

so we still can estimate the $E(H) - H$ and $E(dH) - dH$ in (41) by an $n^{-c'}$.

6. SOME PRACTICAL REMARKS ON THE ESTIMATION ALGORITHM

The estimation algorithm has been tested on simulations. We used Ising models; these models are defined on the set $D = \{1, \ldots, I\} \times \{1, \ldots, I\}$ and takes binary values; for such models the conditional law at site (i, j) knowing the other sites depends only of the values at the four nearest neighbourgs of (i, j) and is:

$$\pi_\theta(x_{ij} \mid \bar{x}_{ij}) = \frac{\exp[-x_{ij}(\theta_1(x_{ij-1} + x_{ij+1}) + (\theta_2(x_{i-1\,j} + x_{i+1\,j}))]}{1 + \exp[-\theta_1(x_{ij-1} + x_{ij+1}) + (\theta_2(x_{i-1\,j} + x_{i+1\,j})]}.$$

For this two-parameter model, the sufficient statistic, α, is $\alpha = (\alpha_1, \alpha_2)$ where:

$$\alpha_1 = \sum_{i,j} x_{ij} x_{ij-1} \qquad \text{and} \qquad \alpha_2 = \sum_{i,j} x_{ij} x_{i-1\,j}.$$

In the computation of the conditional probability, there are problems that occur at the edges of the domain D. To avoid them one can consider that one deals with the restriction of a field defined on all $\mathbb{Z}^2$, and fix, for

the computation $x_{ij}=0$ outside D; one then uses in fact the conditional law on D, knowin $x_{ij}=0$ outside D, and this is a good approximation of the absolute law on D.

The size of the "picture" we used here is 16×16 (256 sites). For such a model 1 000 sweeps of D by the algorithm take about 50 sec. on an IBM 4143.

When using the estimation algorithm, one must start with an initial value, θ_0, for the parameter. One needs then a preliminar estimation procedure; a rather efficient method is to just let the algorithm wind up a little time with a constant step. Other preliminar estimation methods are currently being studied.

A second problem that occures when one uses the estimation procedure is the choice of the step of the stochastic gradient algorithm. It is impossible in practice to use a step of the kind $1/n\,\mathrm{U}$ with the theoretical value of U given in theorem 4.1, which is far too large to expect the convergence to be achieved in a reasonable time, the best results have been obtained by using a step $a_n=1/(n\,\mathrm{U}+n_0)$, where we used $\mathrm{U}=1$ or 10 and $n_0=1{,}000$. One can remark that the probability of non convergence for the algorithm is a $O(1/n_0)$. [This result is a straitforward application of [7], theorem 1.3 (ii)].

Finally, one must define a stopping procedure for the algorithm. One can iteratively compute the mean of the $\alpha(\mathrm{X}_n)$, which tends to $\mathrm{E}_{\theta_*}[\alpha]=\alpha_0$ and stop when the estimated mean is close enough to α_0. Another possibility (which is the one we used) is to choose a criterium of the kind:

$$\| \theta_{n+\mathrm{P}}-\theta_n \| \leqq \delta p/(n_0+n\,\mathrm{U}),$$

where p is fixed and δ is a constant that can be choosen as a level of accuracy for $\mathrm{E}_\theta(\alpha)-\alpha_0$.

When one looks at the behaviour of this algorithm, one can note the following facts: the parameter θ_n comes near its limit after a small amount of sweeps (about 100); then, in order to get closer to the limit one must wait a longer time, as the stochastic behaviour becomes more important than the deterministic behaviour. One must also notice that, in statistical applications, the approximation θ_n of the maximum likelihood estimator, θ_*, needs not be better than the precision of θ_*, as an estimator of the true parameter of the model. More results about these questions will be published elsewhere [8].

ACKNOWLEDGEMENTS

I thank Pr. R. Azencott for his suggestions during the elaboration of this paper, which is part of a Université Paris Sud thesis prepared under his guidance.

REFERENCES

[1] J. BESAG, Spatial Interaction and Statistical Analysis of Lattice Systems, *J. Royal Stat. Soc.*, B, vol. 36, 1974, p. 192-236.

[2] J. BESAG, *Statistical Analysis of Dirty Picture*, preprint.

[3] D. GEMAN and S. GEMAN, Stochastic Relaxation and Bayesian Restauration of Images, *I.E.E.E. Proc. Patt. Anal. Mach. Int.*, 6, 1985.

[4] X. GUYON, *Pseudomaximum de vraisemblance et champs markoviens*, Preprint, 1985.

[5] X. GUYON, *Estimation d'un champ de Gibbs*, preprint, 1986.

[6] D. L. ISAACSON and R. W. MADSEN, *Markov Chains, Theory and Applications*, Wiley, 1976.

[7] M. MÉTIVIER and P. PRIOURET, *Théorèmes de convergence presque sure pour une classe d'algorithmes stochastique à pas décroissants*, École Polytechnique Rapport Interne, n° 116, 1984.

[8] L. YOUNES, *Thesis at University*, Paris-XI, 1988.

(*Manuscrit reçu le* 4 *décembre* 1986.)