Prob. Th. Rel. Fields 82, 625–645

Probability Theory and Related Fields

Parametric Inference for Imperfectly Observed Gibbsian Fields

Laurent Younes

Université Paris Sud, Laboratoire de Statistique Appliquée, Bat 425, F-91405 Orsay Cedex, France

Summary. This paper presents a maximum likelihood estimation method for imperfectly observed Gibbsian fields on a finite lattice. This method is an adaptation of the algorithm given in Younes [28]. Presentation of the new algorithm is followed by a theorem about the limit of the second derivative of the likelihood when the lattice increases, which is related to convergence of the method. Some practical remarks about the implementation of the procedure are eventually given.

1. Introduction

We study here parametric inference problems for Gibbs fields. For this purpose, numerous methods have been developed. The most popular one is pseudo-likelihood estimation, which has been proposed by Besag (Besag [2]), as an alternative to maximum likelihood, which was then thought impossible to compute, at least in general set-up (maximum likelihood could be evaluated, or approximated, for particular fields, for example in the Gaussian homogeneous case, or in the Ising case; see Guyon [10], and Künsch [13] for the Gaussian case, and Pickard [20–22] for the Ising model). For binary markov fields, an estimation framework has been studied by Possolo (Possolo [24]). In preceding papers (Younes [27], Younes [28]), I proposed a stochastic gradient algorithm which converges to the maximum likelihood estimator for general Gibbs models with finite state space.

All these methods are very strongly dependent on the structure of the neighbourhood system of the field. They all assume that the conditional law at one site knowing all the other ones only depends on a few sites that are called its neighbours. This assumption is in general very natural, and is satisfied for all models that are used in practice. But in the case when the modeled field is imperfectly observed (for example perturbed by a noise), the noisy field, which is the only information that is available for inference, does not have this Markovian property, even when the neighbourhood system of the original field is very simple. A standard estimation set-up is thus impossible to use, and one must search for different

procedures. For example, Chalmond (Chalmond [5]) proposed such a procedure for parameter estimation of noisy fields in view of image restoration. His method, which has the advantage of being easy to implement, depends in a large part on the particular form of the model he chose. We here propose a method that can (in theory) be used in very general situations. It can be seen as a generalisation of the algorithm in Younes [28].

In this paper, we shall present this algorithm, (Sect. 3), and give some results about its convergence, results that are mainly inspired by Métivier and Priouret's work on stochastic gradient algorithms (Métivier-Priouret [18], Benveniste-Métivier-Priouret [1]). We shall point out practical and theoretical problems that this procedure involves. Section 4 will be devoted to the statement and the proof of a theorem that is related to the convergence of the algorithm and to the asymptotic normality of the maximum likelihood estimator. In Sect. 5, we shall give some aspects of practical problems that one may encounter.

First, we shall present, in the next section, the modelization we use, and recall some results about random fields that will be useful in the paper.

2. Random Fields

2.1. Generalities

We shall consider here imperfectly observed random fields. We shall then define an original field, X, which will be modeled in a standard way, and an observed field Y, which will be a perturbation of X.

The field X will be indexed by $\mathbf{Z}^2$. The restriction to 2-dimensional fields is not crucial for what follows; it is mainly done in order to simplify notations, and because 2-D problems are, at the moment, the most frequent cases in spatial statistics.

Let F be a given finite set. The field X, that can be written as

$$X=(X_s, s\in\mathbf{Z}^2),$$

will be assumed to take its values in $F^{\mathbf{Z}^2}$.

For any subset D of $\mathbf{Z}^2$, and any $x\in F^{\mathbf{Z}^2}$ we shall note $x_D=(x_s, s\in D)$.

We give ourselves a family of potentials: $(\lambda_C(\theta, x))$ where C runs over finite subsets of $\mathbf{Z}^2$, and $\lambda_C(\theta, x)$ is defined on $F^{\mathbf{Z}^2}$ but only depends on coordinates of x that are elements of C. θ is the parameter we want to estimate: the law of X will be associated to a potential $(\lambda_C(\theta_*, .))$, for an unknown θ_*.

Y will be a function of X, which will be assumed, and this is the principal limitation of our method, to be calculated from X component by component. More precisely, we shall consider a finite set G and a function b from F onto G, and assume that for all s, $y_s=b(x_s)$. Calling $\mathbf{b}$ the application defined on $F^{\mathbf{Z}^2}$, which has all its components equal to b, we have: $Y=\mathbf{b}(X)$.

In addition, if D is any subset of $\mathbf{Z}^2$, we shall call b_D the application defined on F^D with all components equal to b.

The case of main interest included in this situation is the case of noisy data. To see this, consider an original field, X^1, and a noise $N=(N_s)$. We call $X=(X^1, N)$, and we model the joint law of X^1 and N, making no assumption such as independence,

for example. The observed data is thus a function of X. In the case of additive noise, it is X^1+N. Of course many variations on the way the noisy field is constructed may be taken into account, with the limitation that it must be computed independently for each component.

Another context which is included in this set-up is the case when the original field is partially unobservable; for example, one often considers, in image processing, fields of the kind $X=(X_P, X_E)$ where X_P is the intensity (grey level) at each pixel, and X_E is the occurence of an edge at the site. The function b corresponding to the true observation is then $b(x_p, x_e)=x_p$.

2.2. *Assumptions*

Throughout this paper, we shall assume the following facts on the potentials:

Assumption 1. i) There exists a $\gamma \geqq 0$ such that, for all x and all $\theta, \lambda_C(\theta, x)=0$ whenever diam $(C) \geqq \gamma$. diam (C) is the diameter of C for the metric on $\mathbf{Z}^2$ defined by:

$$d((i,j),(i',j'))=\max(|i-i'|, |j-j'|). \tag{1}$$

ii) We call, for $s \in \mathbf{Z}^2$ T_s the shift operator of step s, i.e.

$$(T_s x)_t = x_{s+t}.$$

We assume then that for all C and for all s:

$$\lambda_{C+s}(\theta, T_s x)=\lambda_C(x).$$

iii) $\lambda_C(\theta, x)$ is twice continuously differentiable in θ for all x.

We have thus assumed that the field X is of bounded range, uniformly in θ, and that the potential is spatially homogeneous.

This assumption ensures the existence of a law on $\mathbf{Z}^2$ associated to the potential for any θ. This means that for all θ there exists a law for X such that, for any $D \subset \mathbf{Z}^2$, finite, for any $x \in F^D$, and $x' \in F^{D^c}$, where D^c is $\mathbf{Z}^2 \backslash D$, the probability of $X_D=x$ knowing $X_{D^c}=x'$ is given by:

$$\pi_\theta(x|x')=\exp(-\Lambda_D(\theta, x \cdot x'))/Z(\theta, x') \tag{2}$$

where $x \cdot x'$ is the concatenation of x and x', and:

$$\Lambda_D(\theta, x \cdot x')=\sum_{C, C \cap D \neq \emptyset} \lambda_C(\theta, x \cdot x')$$

Z is a normalizing constant.

As we assume that the potential is homogeneous, we can (and shall) assume *that the law of X is homogeneous* (see Preston [24]).

Let's point out that this model does not ensure uniqueness of a law on $\mathbf{Z}^2$ associated to a parameter θ. This uniqueness is closely related to mixing properties of the field X. Although no uniqueness nor mixing assumption is needed for the definition of the algorithm, we shall use such properties in the asymptotics we shall make in Sect. 4. This is why we recall now some sufficient conditions related to them.

2.3. Uniqueness and Mixing

We recall here Dobrushin's uniqueness conditions and more precisely conditions given by Simon, ensuring Dobrushin's ones. These conditions apply to finite state spaces. They say that, if we are given a potential (λ_C), a sufficient condition for uniqueness of a field associated to it is:

Condition 1. There exists an $\alpha \in [0, 1[$ such that:

$$\forall t \sum_{C, t \in C} (|C|-1)\|\lambda_C\|_\infty < \alpha \tag{3}$$

(Dobrushin [5], Künsch [14], Simon [26]).

For a set $A \subset \mathbf{Z}^2$ we shall call $\mathscr{F}_A$ the σ-algebra generated by the projection of $F^{\mathbf{Z}^2}$ on F^A. We define, for a law P on $F^{\mathbf{Z}^2}$:

$$\phi(A, B) = \sup |P(V \cap W) - P(V)P(W)|$$

where the supremum is taken over $V \in \mathscr{F}_A$ and $W \in \mathscr{F}_B$. One can easily check the well known estimate: if A, B are subsets of $\mathbf{Z}^2$, and f, g, are two bounded and positive functions, that are respectively $\mathscr{F}_A$ and $\mathscr{F}_B$ mesurables, we have:

$$|E(fg) - E(f)E(g)| \leqq \|f\|_\infty \|g\|_\infty \phi(A, B) \tag{4}$$

Under condition 1, we have the following theorem:

Theorem 1. *If we assume condition 1, and if the potential λ is of bounded range, we have (P being the law associated to λ):*

$$\phi(A, B) \leqq K \cdot \bar{\alpha}^{d(A,B)}$$

Where $d(A, B)$ is the distance between sets A and B, according to the distance in $\mathbf{Z}^2$ defined in (1); *K is a constant that depends on A and B, and $\bar{\alpha}$ can be taken as being equal to $\alpha^{1/\gamma}$.*

We shall say that the law is exponentially mixing, with rate $\bar{\alpha}$ (see Dobrushin [5], Künsch [14], Guyon [10].

Let's remark that the results we have given for a field on $\mathbf{Z}^2$ are also true for a field defined on a subset S of $\mathbf{Z}^2$, so that we can replace all the "$\in \mathbf{Z}^2$" by "$\in S$". We also can notice that the field doesn't need to be homogeneous, and we don't even need the space F to be the same for all s.

The following result about comparison of expectations will be useful.

Let S be a finite subset of $\mathbf{Z}^2$, and two potentials, (λ_C^0) and (λ_C^1), indexed by the subsets of S.

On F^S we define the law R_i by:

$$R_i(x) = \exp\left[-\sum_C \lambda_C^i(x)\right]/Z_i,\ i = 0, 1\,. \tag{5}$$

We call E_0 and E_1 the corresponding expectations. We want to estimate the difference between these expectations when both laws are mixing. This is the subject of the following proposition:

Proposition 1. *We assume*:

i) (λ_C^0) *and* (λ_C^1) *of bounded range* (*i.e. there exists a* $\gamma \geqq 0$ *such that* $\lambda_C^0 = \lambda_C^1 \equiv 0$ *if* $\operatorname{diam}(C) > \gamma$).

ii) *for* $\alpha \in [0, 1[$

$$\forall t \in S, \sum_{C, t \in C} (|C| - 1) \max(\|\lambda_C^0\|_\infty, \|\lambda_C^1\|_\infty) < \alpha \tag{6}$$

Let $A \subset S$ *and* f $\mathcal{F}_A$*-measurable. We have*:

$$|E_0(f) - E_1(f)| \leqq K \|f\|_\infty \sum_C \|\lambda_C^0 - \lambda_C^1\|_\infty \alpha^{d(A,C)/\gamma} \tag{7}$$

K is a constant that only depends on A, γ *and* α.

This is standard, references for this are, for example Künsch [14] or Föllmer [6]. In our case, it can also be proved very easily by defining

$$\lambda_C^\tau = \lambda_C^0 + \tau(\lambda_C^1 - \lambda_C^0),$$

and expectations E_τ, as in (5), and making simple estimates on the derivative of E_τ.

As a consequence of this result, or as a consequence of Künsch [14], Corollary 2.4, one can compare conditional expectations to absolute expectation for mixing fields. Let's give ourselves a field defined on $S \subset \mathbf{Z}^2$, not necessarily finite and assume that this field is associated to a potential (λ_C). If $D \subset S$ is finite, we want to compare for f, depending only on a small number of coordinates, the expectation of f conditionally to D^c, which is computable, to the expectation under the marginal law of the field (the absolute expectation) which is most of the time not computable. We have:

Corollary 1. *We assume a field on S satisfying conditions 1* (*we call* (λ) *the associated potential*). *Let* $D \subset S$, *and* $A \subset D$ *be finite sets. We denote by E the absolute expectation for the field and by* E^{D^c} *the conditional expectation knowing* D^c. *We assume that D and A are rectangles* (*i.e. of the form* $[a_1, a_2] \times [b_1, b_2]$).

Let then f be $\mathcal{F}_A$ *measurable. We have the following estimate*:

$$\|E^{D^c}(f) - E(f)\|_\infty \leqq K \|f\|_\infty \alpha^{d(A, D^c)/(2\gamma)} \tag{8}$$

K is a constant which depends on α, *A, and of the range of the field.*

We omit the proof of this standard result.

2.4. Law of X Knowing Y

As we shall see in Sect. 3, the conditional law of X knowing Y is of main interest for our problem. We give now some results about it.

θ is fixed in this section and we omit it in the notations.

Fix an element y in $G^{\mathbf{Z}^2}$. We shall study the law of X conditionally to $Y = y$. Its support is the set $\mathbf{F}$ of all x such that $\mathbf{b}(x) = y$. As we have defined $\mathbf{b}$ component by component, this set is the product:

$$\mathbf{F} = \prod_s \bar{F}_s$$

where $\bar{F}_s$ is by definition: $b^{-1}(y_s)$.

Let's put for $D \subset \mathbf{Z}^2$:

$$\bar{F}_D = \prod_{s \in D} \bar{F}_s$$

and let $\bar{\mu}_D$ be the counting measure on $\bar{F}_D$; $\bar{\mu}_D$ is also product of the $\bar{\mu}_s$, counting measures on $\bar{F}_s$, for $s \in D$.

Let x' be in $\bar{F}_{D^c}$. Let $\pi^y(x|x')$ be the probability of $x \in \bar{F}_D$, conditional to $X_{D^c} = x'$ and $Y = y$. This probability is equal to:

$$P(X_D = x | X_{D^c} = x',\ Y = y) = P(X_D = x | X_{D^c} = x',\ Y_D = y_D)\,.$$

And this is therefore equal to:

$$\frac{\pi(x|x')}{\sum_{u,\, b_D(u) = y_D} \pi(u|x')}\,. \tag{9}$$

This shows the following result:

Result 1. If y is fixed, the law of X knowing $Y = y$ is in general non homogeneous. The state space at site s is: $\bar{F}_s = b^{-1}(y_s)$, and the field is associated to the same potential as the field X.

This implies the corollary:

Corollary 2. *If (λ_C) satisfy conditions* 1, *the law of X conditionally to $Y = y$ is, for all y, exponentially mixing, with the same rate as the law of X.*

This is trivial under Dobrushin-Simon conditions. But this result appears to be far harder to obtain, – and maybe false – under weaker hypotheses.

We now come back to the estimation process.

3. Estimation of θ

3.1. Likelihood

Let's recall that we have a model for the field X, and that we want to make statistical inference from the observation of Y, on a finite domain $D \subset \mathbf{Z}^2$, i.e., we observe $Y_D = b_D(X_D)$.

First of all, we must note that it is impossible to obtain the marginal law of X on D. We must then choose an approximation and a possible choice is the conditional probability in (2), in which we fix x' arbitrarily. We can make numerous different choices of the likelihood, making the energy Λ vary, especially for edge sites. Anyway, we shall here impose to the "likelihood" to be given by (2), and then to be of the kind:

$$\pi_\theta(x) = \exp(-\Lambda(\theta, x))/Z(\theta)\,.$$

We assume for the moment that D and x' are fixed and don't mention them in the notations.

We can now compute the equations that have to be solved in order to obtain the maximum likelihood estimator. The law of Y is given by

$$\psi_\theta(y) = \sum_{x,\, b_D(x) = y} \pi_\theta(x)\,.$$

If we differentiate the logarithm of ψ, we get:

$$\frac{d}{d\theta}\log\psi_\theta(y)=\frac{\sum_{x,b_D(x)=y}\frac{d}{d\theta}[\log\pi_\theta(x)]\pi_\theta(x)}{\sum_{x,b_D(x)=y}\pi_\theta(x)}.$$

This is equal to the conditional expectation of $\frac{d}{d\theta}(\log\pi_\theta)$ knowing $Y=y$, and we shall denote it by:

$$\frac{d}{d\theta}\log\psi_\theta(y)=E_\theta\left[\frac{d}{d\theta}\log\pi_\theta|Y=y\right]. \tag{10}$$

If we differentiate again, we obtain:

$$\frac{d}{d\theta^2}\log\psi_\theta(y)=E_\theta\left[\frac{d^2}{d\theta^2}\log\pi_\theta|Y=y\right]+\mathrm{var}_\theta\left[\frac{d}{d\theta}\log\pi_\theta|Y=y\right]. \tag{11}$$

In fact, it was not necessary to assume that the state space was finite in order to obtain these equations; this only formally simplifies the computations. With essentially the same kind of calculus, and simple considerations on conditional expectations, one can check that Eq. (10) and (11) are true for any law π_θ.

If we express the derivatives of $\log\pi_\theta$ by means of Λ, we get the following result:

Result 2. The derivatives of the log-likelihood of Y are:

$$\frac{d}{d\theta}\log\psi_\theta(y)=E_\theta[\Lambda']-E_\theta[\Lambda'|Y=y]. \tag{12}$$

And

$$\frac{d^2}{d\theta^2}\log\psi_\theta(y)=E_\theta[\Lambda'']-E_\theta[\Lambda''|Y=y]-(\mathrm{var}_\theta[\Lambda']-\mathrm{var}_\theta[\Lambda'|Y=y]). \tag{13}$$

In order to find the maximum likelihood, one must thus solve:

$$h(\theta)=0 \tag{14}$$

Where h is defined by:

$$h(\theta)=E_\theta[\Lambda']-E_\theta[\Lambda'|Y=y]. \tag{15}$$

3.2. Estimation method

3.2.1. Introduction

If we assume that the field is perfectly observed, i.e. $Y=X$, the expression of $h(\theta)$ boils down to:

$$h(\theta)=E_\theta(\Lambda')-\Lambda'(y)=E_\theta(\Lambda'-\Lambda'(y))$$

(y is the observed value of $Y=X$).

This equation is thus of the kind $E_\theta(f)=0$, and fits into the standard context of stochastic gradient algorithm: one tries to use a recurrent equation of the kind:

$$\theta_{n+1}=\theta_n+\sigma_n f(\theta_n, X^{n+1})$$

where σ_n is an appropriate step and X^{n+1} tends to simulate the law of parameter θ_n. This is the method we proposed in (Younes [28]).

In the case of imperfectly observed data, the basic idea lies in the remark that (15) is also of the kind $\tilde{E}_\theta(f)=0$, where $\tilde{E}$ is an expectation for a properly chosen law. Indeed, let's consider a field on $(F^D)^2$, denoted by (X_1, X_2). Let X_1 have the law π_θ and X_2 the law π_θ^y, which is by definition the conditional law of X knowing $Y=y$. Let's call E the expectation for the joint law and

$$f(\theta, x_1, x_2)=\Lambda'(\theta, x_1)-\Lambda'(\theta, x_2)$$

We have $\tilde{E}_\theta(f)=h(\theta)$.

In theory, the joint law of (X_1, X_2) may be any law with marginals π_θ and π_θ^y; in practice it seems very difficult to find Gibbsian fields with given marginals, excepted the obvious one which corresponds to independent X_1 and X_2. However, the joint law of X_1 and X_2 will be of the kind:

$$\tilde{\pi}_\theta(x_1, x_2)=\exp(\tilde{\Lambda}(\theta, x_1, x_2))/Z(\theta, y)$$

where $Z(\theta, y)$ is the sum of the $\tilde{\Lambda}(\theta, x_1, x_2))$ for all x_1 and x_2 such that $b_D(x_2)=y$. When X_1 and X_2 are independent, we have: $\tilde{\Lambda}(x_1, x_2)=\Lambda(x_1)+\Lambda(x_2)$.

To solve $h(\theta)=0$, we can thus follow the same process as in Younes [28]. We briefly recall it now.

3.2.2. Presentation of the Algorithm

We want to solve the equation $E_\theta(f)=0$ for a given function $f(\theta, x)$ and a Gibbsian law $\pi_\theta(x)$, of a field X on F^D for a finite $D\subset \mathbf{Z}^2$. It is important for what follows that for each s, the conditional laws of $X_s=x_s$ knowing $X_t=x_t$, for $t\neq s$ are easily computable. This is true for the law $\tilde{\pi}$ we introduced before (at least when X_1 and X_2 are independent), but this is not true for the marginal law of Y, ψ_θ, in Sect. 3.1.

As already said, we shall use an algorithm of the kind

$$\theta_{n+1}=\theta_n+\sigma_n f(\theta_n, X^{n+1}) \tag{16}$$

where σ_n is a step (gain) that we shall choose, for simplicity, to be of the kind $[(n+1)U]^{-1}$, although, in practice, the choice of matricial gain can accelerate convergence (see Benveniste-Métivier-Priouret [1]). X^n must be a process that tends to simulate the law π_{θ_n}. Unfortunately, there exists no direct method of simulation of a Gibbsian field on D. The only tool we can use is the Gibbs sampler, which is an iterative simulation process of Gibbsian fields. Before explaining the way we form X^{n+1}, we must recall the definition of the Gibbs sampler, that can be found for example in Geman [7].

We want to simulate a law π on F^D. The exact value of π is impossible to compute, but one can easily use the conditional probabilities:

$$\pi^s(x_s|x_t, t\neq s),$$

which are the probabilities of having x_s in s, knowing the other sites. This is what is done in the Gibbs sampler: its philosophy is to start with any configuration x_0, to sweep the domain D site by site, and at each time a site is visited, to renew the current configuration by changing only this site, according to the corresponding conditional probability.

More precisely, we must define a sequence (s_n) of sites that sweeps D; we shall impose to the s_n the following periodicity condition: there exists an integer R such that, for all n

$$D \subset \{s_{n+1}, \ldots, s_{n+R}\}.$$

We start with a configuration x_0 and define from it a sequence of configuration, X^n, such that: $X_s^{n+1} = X_s^n$ for all $s \neq s_n$ and we obtain the new value, $X_{s_n}^{n+1} = X_s^n$, at random, according to the law:

$$\pi^{s_n}(.|X_t^n, t \neq s_n).$$

One can easily check that (X^n) is an inhomogeneous Markov process with transition probabilities:

$$P(X^{n+1} = dx | X^n = x') = P^{n,n+1}(x', dx) = \bigotimes_{s \neq s_n} \delta_{x_s'}^{(dx_s)} \otimes \pi^{s_n}(dx_{s_n} | \bar{x}'_{s_n}) \tag{17}$$

$\delta_{x_s'}$ is the Dirac measure at point x_s'.

This Markov chain converges in law to the distribution π on F^D. We can now return to the estimation algorithm (16), and give the definition of X^{n+1}.

For each θ, we can define a Gibbs sampler and then a family of transition kernels, $P_\theta^{n,n+1}$ given by (17). Now X^{n+1} is defined as taken from X^n according to the kernel $P_{\theta_n}^{n,n+1}$.

The exact definition of the algorithm is:

$$\begin{cases} X_0, \theta_0 \quad \text{given} \\ \theta_{n+1} = \theta_n + \dfrac{1}{(n+1)U} f(\theta_n, X^{n+1}) \\ P(X^{n+1} = x | X^n = x') = P_{\theta_n}^{n,n+1}(x', x). \end{cases} \tag{18}$$

The important point in this algorithm, is that we do not expect the Gibbs sampler to converge for each θ_n, or equivalently, we do not mind if X^n really follows the law π_{θ_n} for all n (this will be generally untrue, unless θ_n has converged). Thus, we do not have to wait as in (Lippman [17]), at each step, for the convergence to be achieved. In (18), we renew the value of θ at each step of the simulation algorithm; it is possible to wait a few sweeps between each renewal, in order to reduce the calculus that is needed, but even in this case, X^n doesn't follow the law π_{θ_n}. Of course, when the parameter has converged, and thus when its variations are small enough, the law of X^n begins to stabilize. One can even show that empirical estimations of means of functions computed on the basis of the (X^n) converge to the expectation of the function under the limit parameter.

The proof of convergence of this algorithm, in the case of perfectly observed data and exponential model lies in (Younes [28]) and (Younes [29]). In this case, we proved almost sure convergence provided the constant U is large enough. We cannot follow this proof in the present case, because the dynamic of the algorithm is

far from being as simple. We shall in the next section, give an idea of what is needed for the study of the behaviour of an algorithm such as (18). We shall apply results of Métivier-Priouret [18] to obtain what we shall call quasi-convergence of (18), which is a weaker result than what has been made in (Younes [28]).

First, let's look at what (18) gives in the precise context of Sect. 3.2.1.

We have a field (X_1, X_2). We choose X_1 and X_2 independent, and we recall that X_1 follows the law π_θ and X_2 the law π_θ^y. The simulation of (X_1, X_2) can then be done by simulating each component separately. This leads us to define two Gibbs samplers, and thus two families of conditional probabilities, noted $P_\theta^{n,n+1}$ and $Q_\theta^{n,n+1}$ respectively corresponding to π_θ and π_θ^y. As we know exactly the potentials for both these laws, these kernels are easily computable. Now the algorithm is obtained by simply transcripting what has been said before into the present context, and this gives:

$$\begin{cases} X_1^0, X_2^0, \theta_0 \quad \text{given} \\ \theta_{n+1} = \theta_n + \dfrac{1}{(n+1)U}(\Lambda'(\theta_n, X_1^{n+1}) - \Lambda'(\theta_n, X_2^{n+1})) \\ P(X_1^{n+1} = x_1 | X_1^n = x_1') = P_{\theta_n}^{n,n+1}(x_1', x_1) \\ P(X_2^{n+1} = x_2 | X_2^n = x_2') = Q_{\theta_n}^{n,n+1}(x_2', x_2). \end{cases} \tag{19}$$

In the particular case of exponential models, which are almost exclusively used in practice, we have $\Lambda(\theta, x) = \langle \theta, H(x) \rangle$ and the increment $\theta_{n+1} - \theta_n$ is independent of θ_n. In the case of noisy data, X_2 follows the conditional law of the original field knowing the noisy one. It will thus become more deterministic as the noise becomes weaker, and for negligible noise, we get back into the algorithm of (Younes [28]).

We now recall some results about the behaviour of algorithms such as (18).

3.3. *Quasi-convergence of the Algorithm*

3.3.1. *Introduction*

The results we cite now are taken from (Métivier-Priouret [18]) and (Benveniste-Métivier-Priouret [1]). Like most of the results related to stochastic algorithms, they *do not provide* almost sure convergence of the parameter θ_n in (18). In fact, they give estimates of the probability of non-convergence of the algorithm, and one of the conclusion we can draw from them is that, if the parameter comes back infinitely often in a given compact set, then it converges.

Métivier-Priouret [18] have studied Markovian stochastic algorithms, i.e. of the kind:

$$\theta_{n+1} = \theta_n + \frac{1}{(n+1)} f(\theta_n, X^{n+1}) \tag{20}$$

$$P(X^{n+1} \in A | X^n = x') = P_{\theta_n}(x', A).$$

Such an algorithm is thus controled by homogeneous transition kernels. (The transition only depends on n by the parameter θ_n). In this context they gave sufficient conditions which ensure quasi-convergence. These conditions are

numerous, and are made to fit in very general situations. Some of them boil down to trivialities in our case, and we shall not recall them. In fact, only two groups of hypothesis in Métivier-Priouret [18] are really significant here (they are called H5 and H6 in [18]). One group is related to stochastic behaviour of the process X^n and the other one to dynamical behaviour of the mean differential equation associated to (18).

This equation is:

$$\frac{d}{dt}(\theta)=h(\theta). \tag{21}$$

Where $h(\theta)$ is given by (15), i.e. $h(\theta)=E_\theta(f(\theta,.))$.

If we put:

$$g(\theta,x)=E_\theta(f(\theta,.))-f(\theta,x). \tag{22}$$

Equation (18) becomes:

$$\theta_{n+1}-\theta_n=\frac{1}{n+1}(h(\theta_n)-g(\theta_n,X_{n+1})).$$

This equation can be interpreted as a discretization of (21) perturbed by the "noise" g. This remark forms the basis of the theory of stochastic algorithms. There exists numerous results which compare the sequence (θ_n) to the trajectories of (21). When this differential equation possesses an asymptotically stable point one can expect θ_n to converge to it.

It is clearer now why one needs stochastic and dynamical conditions: on one hand, we must control the behaviour of the perturbation g, and on the other hand one has to control the stability of the mean differential equation. We begin by the stochastic ones.

3.3.2. Conditions H5 in Métivier-Priouret [18]

These conditions essentially assume the existence of solutions of Poisson equations associated to g and to the transition kernels. In the homogeneous case this means that there exist functions $(\varrho(\theta,x))$ that satisfy:

$$\varrho(\theta,x)-(P_\theta\varrho)(\theta,x)=g(\theta,x),$$

and possess some regularity properties in θ, ensured for example by differentiability.

3.3.3. Conditions H6 in Métivier-Priouret [18]

As already noted, the asymptotic stability of the mean differential equation is a natural condition of convergence of the algorithm. Let $\mathscr{D}$ be an open subset of $\mathbf{R}^d$, the space where the parameter θ may vary.

Asymptotic stability with attraction domain $\mathscr{D}$ is implied by the existence of a Liapunov function, $\mathscr{L}$, twice continuously derivable, such that, denoting by $\hat{\theta}$ the solution of (14):

1. $\mathscr{L}(\hat{\theta})=0$ and $\mathscr{L}(\theta)>0$ for $\theta\in\mathscr{D}$ and $\theta\neq\hat{\theta}$.
2. $\langle\mathscr{L}'(\theta)|h(\theta)\rangle<0$ if $\theta\in\mathscr{D}$ and $\theta\neq\hat{\theta}$.
3. $\mathscr{L}(\theta)\to+\infty$ if $\theta\to\partial\mathscr{D}$ or if $|\theta|\to+\infty$, $\theta\in\mathscr{D}$.

3.3.4. Main result in Métivier-Priouret [18]

The main result of Métivier-Priouret [18], is the following:

Theorem 2. *Let $\mathscr{D}$ be a subset of $\mathbf{R}^d$ for which conditions* H6 *are true. Under conditions* H5, *we have*:

For all compact set $\mathscr{Q}$ included in $\mathscr{D}$, the probability of non-convergence of the sequence θ_n defined in (18) *to the solution of* (14), *conditional to $\theta_{n_0}\in\mathscr{Q}$, is lower than $C(\mathscr{Q}).1/n_0$, where $C(\mathscr{Q})$ is a constant that depends on $\mathscr{Q}$.*

This implies that if θ_n returns infinitely often in a given compact subset $\mathscr{Q}$ of $\mathscr{D}$, then it converges. This is what we can call quasi-convergence of the algorithm.

3.3.5. Application to the Estimation Algorithm

Before applying the preceding result, we must make the following remark. In (20), transition kernels associated to a fixed θ are homogeneous; (18) does not possess this property and thus seems more general. In fact, one can include the non-homogeneous situation into the homogeneous one by the following construction: if X_n is a non-homogeneous Markov chain, with transition kernels $P^{n,n+1}$, and initial law μ, the process (n, X_n) is homogeneous, with initial law $\delta_0\otimes\mu$ and transition:

$$P((p,x'),(q,x))=\delta_{p+1}(q)P^{p,p+1}(x',x).$$

Using this remark, conditions H5 become:
there exists a family of functions $(\varrho_n(\theta,x))$, that satisfies:

$$\varrho_n(\theta,x)-(P_\theta^{n,n+1}\varrho_{n+1})(\theta,x)=g(\theta,x),$$

ϱ_n must be regular in θ. These functions have been studied in (Younes [28]), where existence and regularity have been checked. As (18) is still controled by a Gibbs sampler, conditions H5 are also satisfied in the present context.

When the field X is perfectly observed, and when the model is exponential, it is easily checked that $\dfrac{dh}{d\theta}$ is a negative matrix. In fact, h is the differential of a concave function (the likelihood), and conditions H6 are trivially true; quasi-convergence is thus true in this context; as already said, one can even show (under some additional conditions on the gain) that the algorithm converges almost surely (Younes [28]). Unfortunately, in the present case, the behaviour of the equation is far more unstable and the only thing one can say about it is that, as the maximum of the likelihood ψ_θ is a solution of (14), one can expect the differential of h to be negative in a neighbourhood $\mathscr{D}$ of this solution, and result 2 is true with this $\mathscr{D}$. This says that provided that the algorithm stays not too far from the maximum likelihood estimator, it will converge to it: there exists a compact set $\mathscr{Q}$ containing θ and the

algorithm can diverge only if it goes of $\mathcal{Q}$ and never comes back. The meaning of "not too far" (or, equivalently, the size of $\mathcal{Q}$) will depend on the amount of information about the original field X that is contained in the observed field Y. The following theorem shows this precisely, in the case of mixing fields. It provides the expression of the second derivative of the likelihood, for large observation domain D. It shows that, in a neighbourhood of the true parameter θ_*, this likelihood is concave; as it can be shown that the maximum likelihood estimator is convergent (cf. Younes [29]), this gives another means to check the local stability of the mean differential equation; it will also point out where the risks of unstability, and "explosion" of the algorithm are.

4. Second Derivative of the Likelihood

4.1. Notations

We keep the notations of Sect. 2 and give some new ones.

Let's order $\mathbf{Z}^2$ with respect to the lexicographic order. In order to write the energy Λ in (2) as an homogeneous sum of terms, we define:

$$\beta(\theta, x) = \sum_{C/\max(C)=0} \lambda_C(\theta, x). \tag{23}$$

Assumptions 1 imply that:

$$\beta(\theta, x) = \sum_{C/\max(C)=s} \lambda_C(\theta, T_s x),$$

for all $s \in \mathbf{Z}^2$.

As we assumed bounded range, $\beta(\theta, x)$ only depends on coordinates of x whose indices are included in a finite subset Γ of $\mathbf{Z}^2$, containing 0; the diameter of Γ is at most 2γ.

According to (2), the approximate density is equal to

$$\pi_\theta(x) = \pi_\theta(x|x') = \exp(-\Lambda^{x'}(\theta, x))/Z(\theta)$$

Where we noted:

$$\Lambda_D^{x'}(\theta, x) = \sum_{C, C \cap D \neq \emptyset} \lambda_C(\theta, x \cdot x') \tag{24}$$

Most of the time, we shall not express D nor x' in the notations, writing $\pi(x)$ and $\Lambda(x)$. We also define:

$$\Lambda_D(\theta, x) = \sum_{s, \Gamma + s \in D} \beta(\theta, T_s x). \tag{25}$$

One has:

$$\Lambda_D(\theta, x) = \sum_{C/\Gamma + \max(C) \subset D} \lambda_C(\theta, x) \tag{26}$$

In fact, the expressions Λ and Λ only differ on the edge of D.

We shall denote by $\mathbf{E}_\theta$ the expectation for the absolute law of the field, under the parameter θ. When the energy on D is $\Lambda^{x'}$, we shall denote by $E_\theta^{x'}$ or by E_θ the corresponding expectation. We shall use the same kind of notations for expressing variances (or covariances) according to these laws (Var, $\mathrm{var}^{x'}$,...).

In the following, we shall consider limits of expressions when the domain D tends to $\mathbf{Z}^2$. This must be understood as limits of the expression for any sequence of squares centered at 0 that increases to $\mathbf{Z}^2$.

4.2. Theorem

Theorem 3. *The field X is associated to potentials, $\lambda_C(\theta_*, .)$ satisfying to conditions* 1. *We assume in addition that Dobrushin's conditions are true for θ_*; for notation convenience, we will express it by:*

there exist an $\alpha \in [0, 1[$ such that

$$\sum_{C, 0 \in C} (|C|-1) \left\| \lambda_C(\theta_*, .) \right\| \infty \leqq \alpha^{\gamma} \tag{27}$$

We assume that x is a realization of X under P_{θ_*}, the only law on $F^{\mathbf{Z}^2}$ associated to θ_*.

We consider a function b from F onto G and the associated function $\mathbf{b}$ on the space of configurations over $\mathbf{Z}^2$. We call $Y = \mathbf{b}(X)$, and $y = \mathbf{b}(x)$ the observed realization of Y. We note $\bar{F}_s = b^{-1}(y_s)$

For all D, finite subset of $\mathbf{Z}^2$, let x' be any edge condition defined on $\bar{F}_{D^c}$. For a given θ, we use the approximate likelihood $\pi_\theta(. | x')$.

We denote by $\psi_\theta(. | x')$ the associated likelihood for Y on D, and put:

$$L_D(\theta) = \frac{1}{|D|} \frac{d^2}{d\theta^2} \log \psi_\theta(y_D | x')$$

Then, for all θ for which Dobrushin's condition is true: $L_D(\theta)$ converges for P_{θ_*} almost all x to the sum of two matrices: $-I(\theta, \theta_*)$ and $S(\theta, \theta_*)$ with:

$$I(\theta, \theta_*) = \sum_{s \in \mathbf{Z}^2} \mathrm{Cov}_\theta (\beta'(\theta, .), \beta'(\theta, .) \circ T_s) - \sum_{s \in \mathbf{Z}^2} \mathbf{E}_{\theta_*} \mathrm{Cov}_\theta (\beta'(\theta, .), \beta'(\theta, .) \circ T_s | Y) \tag{28}$$

And:

$$S(\theta, \theta_*) = \mathbf{E}_\theta(\beta''(\theta, .)) - \mathbf{E}_{\theta_*}(\mathbf{E}_\theta(\beta''(\theta, .) | Y)) \tag{29}$$

If $\theta = \theta_*$, S vanishes and $I(\theta, \theta_*)$ is equal to I_* with:

$$I_* = \sum_{s \in \mathbf{Z}^2} \mathrm{Cov}_{\theta_*}(\mathbf{E}_{\theta_*}(\beta' | Y), \mathbf{E}_{\theta_*}(\beta' \circ T_s | Y)) \tag{30}$$

4.3. Proof

4.3.1. First Step

Remark first that (27) implies exponential mixing with rate α for the global law as well as for the approximate ones. We fix a θ. For simplicity of notation, we assume that the associated laws are mixing with the same rate as the law of parameter θ_*.

We have

$$L_D(\theta) = \frac{1}{|D|} [E_\theta[\Lambda'']] - E_\theta[\Lambda'' | Y = y] - (\mathrm{var}_\theta[\Lambda'] - \mathrm{var}_\theta[\Lambda' | Y = y])].$$

Expectations and variances are here computed with respect to the approximating law. We shall intend to replace these expectations by the absolute ones and Λ by the homogeneous sum $\boldsymbol{\Lambda}$.

We split $L_D(\theta)$in two:

$$L_D(\theta)=L_D^1(\theta)-L_D^2(\theta),$$

with

$$L_D^1=\frac{1}{|D|}\,[E_\theta[\Lambda'']-\mathrm{var}_\theta[\Lambda']]$$

and

$$L_D^2(\theta)=\frac{1}{|D|}\,[E_\theta[\Lambda''|Y=y]-\mathrm{var}_\theta[\Lambda'|Y=y]].$$

We study L_D^1, and show that it converges to:

$$\mathbf{E}_\theta(\beta'')-\sum_{s\in\mathbf{Z}^2}\mathrm{Cov}_\theta(\beta',\beta''\circ T_s)$$

In fact this reduces to show that L_D^1 has the same limit as:

$$\frac{1}{|D|}\,[\mathbf{E}_\theta(\boldsymbol{\Lambda}''(\theta,.))-\mathrm{Var}_\theta[\boldsymbol{\Lambda}']].$$

Indeed, by homogeneity:

$$\frac{1}{|D|}\,\mathbf{E}_\theta(\boldsymbol{\Lambda}''(\theta,.))=\frac{1}{|D|}\sum_s \mathbf{E}_\theta\beta''\circ T_s$$

tends to $\mathbf{E}_\theta(\beta''(\theta,.))$ (the number of terms in the sum is equivalent to $|D|$).

And, $\mathrm{Var}_\theta(\boldsymbol{\Lambda}')$ can be written as:

$$\begin{aligned}\mathrm{Var}_\theta(\boldsymbol{\Lambda}')&=\sum_{s,t}\mathrm{Cov}_\theta(\beta'\circ T_s,\beta''\circ T_t)\\&=\sum_{s,t}\mathrm{Cov}_\theta(\beta',\beta'\circ T_{t-s})\end{aligned}\tag{31}$$

The sums are over s and t such that $\Gamma+s\subset D$ and $\Gamma+t\in D$. We note, for $k\in\mathbf{Z}^2$:

$$r_k(D)=\mathrm{card}\,\{s,\Gamma+s\subset D \text{ and } \Gamma+s+k\subset D\}\tag{32}$$

we can write:

$$\frac{1}{|D|}\,\mathrm{Var}_\theta(\boldsymbol{\Lambda}')=\sum_{k\in\mathbf{Z}^2}\frac{r_k(D)}{|D|}\,\mathrm{Cov}_\theta(\beta',\beta'\circ T_k)$$

Since Γ is finite, $r_k(D)$ is equivalent to $|D|$ when D tends to $\mathbf{Z}^2$. Because of mixing properties, the covariance between β' and $\beta'\circ T_k$ tends exponentially to 0, if k tends to $+\infty$; thus, $\frac{1}{|D|}\,\mathrm{Var}_\theta(\Lambda')$ tends to

$$\sum_{s\in\mathbf{Z}^2}\mathrm{Cov}_\theta(\beta',\beta'\circ T_s)$$

by the dominated convergence theorem.

We must now show that we had the right to replace E by $\mathbf{E}$, and Λ by $\boldsymbol{\Lambda}$. It is clear that the difference: $\frac{1}{|D|}(E^y(\Lambda'')-E^y(\boldsymbol{\Lambda}''))$ tends to 0, because both energies differ only at the edge of D, which is of order $\sqrt{|D|}$ for square domains. We must thus show that $E(\Lambda'')-\mathbf{E}(\boldsymbol{\Lambda}'')$ is negligible before $|D|$.

According to Corollary 1, we have:

$$\|E^y_\theta(f)-\mathbf{E}_\theta(f)\| \leqq \text{cst}\,\|f\|_\infty \alpha^{d(A,D^c)/2}$$

if f is $\mathscr{F}_A$-measurable. If we remember that β'' depends only on coordinates indexed by Γ, we can estimate $\|E(\Lambda'')-\mathbf{E}(\boldsymbol{\Lambda}'')\|$ by:

$$\text{cst}\sum_s \alpha^{d(\Gamma+s,D^c)/2} \tag{33}$$

the constant depends on Γ. It is easy to check that the expression in (33) is an $O(\sqrt{|D|})$.

We now estimate the differences of the variances. We first consider:

$$\frac{1}{|D|}(\text{var}(\boldsymbol{\Lambda}')-\text{Var}(\boldsymbol{\Lambda}'))$$

This is equal to:

$$\frac{1}{|D|}\sum_{s,k}(\text{cov}(\beta'\circ T_s,\beta'\circ T_{s+k})-\text{Cov}(\beta'\circ T_s,\beta'\circ T_{s+k}))$$

the sum is over s and k such that $\Gamma+s\subset D$, et $\Gamma+s+k\subset D$.

We note, for $\Gamma+k\subset D$

$$a_k(D)=\frac{1}{|D|}\sum_{s,\Gamma+s+k\subset D}(\text{cov}(\beta'\circ T_s,\beta'\circ T_{s+k})-\text{Cov}(\beta'\circ T_{s+k}))$$

and $a_k(D)=0$ if $\Gamma+k\not\subset D$. We must study the limit of $\sum_k a_k(D)$. Mixing properties (for both absolute and approximate laws) imply:

$$\|a_k(D)\| \leqq \text{cst}\,\alpha^{|k|}$$

The constant depends on the diameter of Γ.

To apply the dominated convergence theorem, we only need to show that $a_k(D)$ tends to 0 for a fixed k, when D tends to $\mathbf{Z}^2$. But, putting $f_t=\beta'\circ T_t$,

$$\begin{aligned}\text{cov}(f_s,f_{s+k})-\text{Cov}(f_s,f_{s+k}) &= E(f_s\cdot f^t_{s+k})-\mathbf{E}(f_s\cdot f^t_{s+k})\\ &\quad -E(f_s)(E(f_{s+k})-\mathbf{E}(f_{s+k}))^t\\ &\quad -(E(f_s)-\mathbf{E}(f_s))\,\mathbf{E}(f_{s+k})^t\end{aligned} \tag{34}$$

k being fixed, we can show convergence of $a_k(D)$ to 0 as before, using Corollary 1.

In order to finish this part of the proof, we need to compare $\frac{1}{|D|}\text{var}_\theta(\Lambda')$ and $\frac{1}{|D|}\text{var}_\theta(\boldsymbol{\Lambda}'')$. We can do this in exactly the same way as before, writing the difference as a sum over k of terms $b_k(D)$, of the same type as the $a_k(D)$, such that each of them exponentially tends to 0 with k (uniformly in D), and tends to 0 when D increases.

4.3.2. Second Step

With essentially the same technics, we study convergence of L_D^2 to:

$$\mathbf{E}_{\theta*}(\beta'') - \sum_{s\in\mathbf{Z}^2} \mathbf{E}_{\theta*}[\mathrm{Cov}_\theta(\beta', \beta'\circ T_s|Y)]$$

This is mainly based on Corollary 2 and on the following remark:

If ϕ is a function defined on $F^{\mathbf{Z}^2}$, P_θ integrable, we have, for $s\in\mathbf{Z}^2$:

$$E(\phi\circ T_s(X)|Y) = E(\phi(X)|Y)\circ T_s \tag{35}$$

As before, we shall show that we can replace Λ by Λ, and the approximate expectation (E) by the absolute one ($\mathbf{E}$) in L_D^2. If we assume this, we have to study on one hand the limit of:

$$\frac{1}{|D|}\sum_{s,\Gamma+s\subset D} \mathbf{E}_\theta(\beta''\circ T_s|Y)$$

which is equal to:

$$\frac{1}{|D|}\sum_{s,\Gamma+s\subset D} \mathbf{E}_\theta(\beta''|Y)\circ T_s.$$

Since X is homogeneous and mixing, so is Y and this sum converges to:

$$\mathbf{E}_{\theta*}(\mathbf{E}_\theta(\beta''|Y)).$$

On the other hand, we must study the limit of

$$\frac{1}{|D|}\sum_{s,t} \mathrm{Cov}_\theta(\beta'\circ T_s, \beta'\circ T_t|Y=y).$$

The sum extends over s and t such that $\Gamma+s$ and $\Gamma+t$ are included in D. If we remark that:

$$\mathrm{Cov}_\theta(\beta'\circ T_s, \beta'\circ T_t|Y=y) = \mathrm{Cov}_\theta(\beta', \beta'\circ T_{t-s}|Y=y)\circ T_s$$

we can order the sum in the following manner:

$$\sum_k \frac{1}{|D|}\sum_s \mathrm{Cov}_\theta(\beta', \beta'\circ T_k|Y=y)\circ T_s. \tag{36}$$

The sums are made on sets of cardinals equivalent to $|D|$. According to the ergodic theorem, for each k, the sum over s, normalized by $|D|$ converges to

$$\mathbf{E}_{\theta*}[\mathrm{Cov}_\theta(\beta', \beta'\circ T_k|Y)]$$

As Corollary 2 implies that the covariance between $\beta'\circ T_s$ and $\beta'\circ T_{s+k}$ converges exponentially to 0 if k tends to infinity, we obtain that expression (36) converges, if $|D|$ tends to $\mathbf{Z}^2$ to

$$\sum_{k\in\mathbf{Z}^2} \mathbf{E}_{\theta*}[\mathrm{Cov}_\theta(\beta', \beta'\circ T_k|Y)].$$

To finish the proof, we must show that we did not change the limits by taking absolute expectations and homogeneous sums. But result 1 implies that, for fixed y, the law of X conditional to $Y=y$ satisfies corollary 1. This allows us to make exactly

the same estimates for L_D^2 as we made for L_D^1, so that we do not need to write them again.

We have then shown that $L_D(\theta)$ converges to:

$$\mathbf{E}_\theta(\beta'') - \sum_{k \in \mathbf{Z}^2} \mathrm{Cov}_\theta(\beta', \beta' \circ T_k) - \mathbf{E}_{\theta_*}(\mathbf{E}_\theta(\beta''|Y)) + \sum_{k \in \mathbf{Z}^2} \mathbf{E}_{\theta_*}[\mathrm{Cov}_\theta(\beta', \beta' \circ T_k|Y)]$$

as announced in Theorem 3.

The expression of I_* for $\theta = \theta_*$ is obtained by noting that:

$$\mathrm{Cov}_\theta(\beta', \beta' \circ T_k) - \mathbf{E}_\theta[\mathrm{Cov}_\theta(\beta', \beta' \circ T_k|Y)] = \mathrm{Cov}_\theta(\mathbf{E}_\theta(\beta'|Y), \mathbf{E}_\theta(\beta' \circ T_k|Y)).$$

4.4. Extensions and Remarks

If one uses an other energy than the one we chose in order to define the approximate likelihood, the conclusions of the preceding theorem are still true, under the following hypothesis: For each D, we can choose a family $(\bar{\lambda}_C)$ of potentials, to which we associate the energy $\bar{\Lambda}_D$ on D. It is easy to change the proof to check that the theorem is still true if we assume:

$$\sum_{C, t \in C} (|C| - 1) \max(\|\lambda_C(\theta, .)\|_\infty, \|\bar{\lambda}_C(\theta, .)\|_\infty) \leqq \alpha_\theta = \alpha < 1$$

for all t, (to be able to apply proposition 1), and:

1. $$\frac{1}{|D|} \sum_C \|\lambda_C - \bar{\lambda}_C\|_\infty \to 0,$$
2. $$\frac{1}{|D|} \sum_C \|\lambda'_C - \bar{\lambda}'_C\|_\infty \to 0,$$
3. $$\frac{1}{|D|} \sum_C \|\lambda''_C - \bar{\lambda}''_C\|_\infty \to 0.$$

Another important extension is the possibility to solve: $h(\theta) = \bar{E}_\theta(f) = 0$, when f is not exactly the derivative of the likelihood. The differential of h is now:

$$\bar{E}_\theta(f') - \overline{\mathrm{cov}}(f, \bar{\Lambda}'(\theta, x))$$

($\overline{\mathrm{cov}}$ means the covariance for the approximate law).

In order that this expression (normalised by $|D|$) has the same limit as

$$\frac{1}{|D|} \bar{E}_\theta(\bar{\Lambda}'') - \bar{\Lambda}''(\theta, x_0) - \overline{\mathrm{var}}(\bar{\Lambda}'(\theta, x)),$$

it suffices to assume that f is a sum of potentials:

$$f = \sum_C \tilde{\lambda}'_C,$$

and that the $(\tilde{\lambda}_C)$ satisfy the same kind of conditions as the $\bar{\lambda}$ before (the mixing condition excepted).

These variations in the hypothesis have a great importance if one tries to use an estimation procedure that avoids bias problems. It is not in the intent of the present paper to develop these questions. We refer in particular to (Künsch [15]), (Guyon

[10]), and (Younes [29]) to obtain some answers. Finally, let's remark that Theorem 3 can be used to obtain asymptotic normality results for maximum likelihood estimators. If the matrix I_* is invertible, I_*^{-1} will be the asymptotic variance of the maximum likelihood estimator. This is proved in (Younes [29]), jointly with consistency – which is true under much weaker assumptions – and will be the subject of a future paper. See also Gidas [9] for consistency and asymptotic normality in the perfectly observed case.

We can make a few remarks about Theorem 3. First, we stated it for a quite general family (λ) of potentials. In fact, the case of main interest in practice is when these potentials are associated to exponential models for X, i.e. $\lambda(\theta, x) = \langle \theta, l(\theta) \rangle$. In this case, the term $S(\theta, \theta_*)$ vanishes, and $\beta'(\theta, .)$ does not dependend on θ anymore in the expression of I. However, dealing with the general case was not far more expensive.

As S is in the applications most of the time equal to 0, we shall make some remarks on $I(\theta, \theta_*)$. It can be easily checked, by taking trivial examples, or by making simulations, that, if θ is far from θ_*, this matrix is no more positive. As already pointed out, this is a real danger for the convergence of the algorithm, and can have as a consequence the explosion of the algorithm, i.e. the parameter θ_n in (19) can tend to infinity. One must then try to find some methods to prevent this kind of accident.

The risk of unstability will of course highly depend on the sharpness of the crest of the likelihood near its maximum. This is given by the matrix I_* in (30). This matrix measures the amount of information about X that is included in Y. If Y is independent of X, this matrix vanishes, and it is maximum as a positive matrix when $Y = X$. *In order that the algorithm* (19) *has a chance to converge, one must assume that this matrix is strictly positive definite.* Even more, it must be "positive enough" to provide a sufficiently sharp maximum, which means that it might be very difficult (and quite impossible by our method) to make precise statistical inference from too strongly perturbed data. Finally, let's remark that nothing ensures uniqueness of a local maximum of the likelihood; this means that there might be situations for which the algorithm converges, but does not reach the maximum likelihood estimator, especially if the starting point is too far from it. This situation is of course caused by the loss of the concavity of the log-likelihood.

However, these limitations leave a rather large domain of application for parametric inference, and one obtains encouraging results with significant perturbation.

We now make some practical remarks on the algorithm.

5. Practical Remarks

The preceding procedure has been tested on simulated data, namely on noisy Gaussian fields (sometimes coupled with another binary field, that represents edge location), Ising models (including external field parameter), perturbed by various noises (addition of a white Gaussian noise, replacement of some values by 0 with a certain probability...). We also used it for parametric estimation on real data, in view of image restoration, using a model of the type of Chalmond's one

(Chalmond [5]). All numerical results will be given in Younes [29]. Simulations enabled us to make estimations of asymptotic mean square errors of estimators; comparison to standard estimators such as pseudo-maximum, has been used to evaluate the significancy of perturbations.

All these experiments brought out the following general facts about the procedure. We separate the algorithm in three phases: its beginning, i.e. the choice of θ_0 in (19), the time it runs, and the choice of a stopping rule for it.

1. The choice of a good starting point is of main importance. It does not only reduce the number of steps that will be needed until convergence, but a choice of a starting point that is too far from the maximum likelihood can cause explosion of the algorithm.

Unfortunately, this choice appears to be very difficult in our context. For totally observed data, some fast and good preliminar estimators can easily be found, using for example pseudo likelihood. In the presence of noise, these estimators are no more available; when the noise is not too strong, pseudo-likelihood estimators, computed as if the data were not perturbed can give an idea of the parameter. But this estimator becomes very bad as the noise becomes singificant.

In our applications, we chose to let (19) wind up a certain time with constant step, starting with any θ_0; if the step is not too large, this provides a good – but long – preliminar estimation procedure.

2. In (19) the gain is a scalar. In fact, it has been seen on simulations that a matricial gain can make convergence faster, and reduce risks of explosion. The choice of the gain is done according to standard gradient algorithms. It is an approximation of the "optimal" one:

$$\left[\text{cst}\left\|\frac{d\log\psi_\theta}{d\theta}\right\|\cdot I+\frac{d^2\log\psi_\theta}{d\theta^2}\right]^{-1}. \tag{37}$$

These derivatives are given by result 2. They can't be exactly computed, but can be estimated by simulations. In our applications, we used empiric estimations computed on a certain number of preceding iterations of (19), based on $\Lambda'(X_i^n)$, $i=\mathbf{1,2}$. As all approximates in the algorithm, these are not expected to provide the exact values of the expectations, unless the algorithm has begun to converge. But they give a sufficiently good idea of them to ameliorate the performances of the algorithm, provided that the constant in (37) is large enough.

3. In order to decide whether the algorithm has converged or not, we test if the difference:

$$(\Lambda'(\theta_n, X_1^{n+1})-\Lambda'(\theta_n, X_2^{n+1}))$$

in (19) has zero mean or not. Indeed, once the algorithm has converged to θ_*, X_1^n and X_2^n respectively follow the laws π_{θ_*} and $\pi_{\theta_*}^y$. One can show that, for large enough domains, under each law, Λ' follows a Gaussian law. We use then a χ^2 test to implement the stopping rule.

Acknowledgements. I wish to thank Professor R. Azencott for his helpful advice during the preparation of this paper. I also thank Professor H. Künsch for valuable suggestions on the treated subject.

References

1. Benviste, A., Métivier, M., Priouret, B.: Algorithmes adaptatifs et approximations stochastiques. Techniques stochastiques. Paris: Masson 1987
2. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. B**36**, 192–236 (1974)
3. Besag, J.: On the statistical analysis of dirty pictures. J. Roy. Stat. Soc. B**48**, 259–303 (1986)
4. Chalmond, B.: Image restoration using an estimated Markov model. Preprint Université Paris Sud (1988)
5. Dobrushin, R.L.: The description of a random field by means of conditional probabilities and conditions of its regularity. Theory Probab. Appl. **8**, 197–224 (1968)
6. Föllmer, H.: A covariance estimate for Gibbs measures. J. Funct. Anal. **46**, 387–395 (1982)
7. Geman, D., Geman, S.: Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. IEEE TPAMI, vol. PAMI 6, pp. 721–741 (1984)
8. Geman, S., Graffigne, C.: Markov random field image models and their applications to computer vision. Proceedings of the international congress of mathematicians. Gleason, A.M. (ed.). Providence: AMS 1987
9. Gidas, B.: Paramter estimation for Gibbs distributions. In: Fully observed data. Preprint, Brown University (1988)
10. Guyon, X.: Parameter estimation for a stationary process on a d-dimensional lattice. Biometrika **69**, 95–105 (1982)
11. Guyon, X.: Pseudo maximum de vraisemblance et champs Markoviens. In: Droesbeke, F. (ed.) Spatial processes and spatial time series analysis. Proc. 6th. Franco-Belgian Meeting of Statisticians. Publication des Pascaltés Universitaires de Saint Louis-Bruxelles 15–62 (1987)
12. Guyon, X.: Estimation d'un Champ de Gibbs. Preprint, Université Paris Sud (1986)
13. Künsch, H.: Thermodynamics and the statistical analysis of Gaussian random fields. Z. Wahrscheinlichkeitstheor. Verw. Geb. **58**, 407–421 (1981)
14. Künsch, H.: Decay of correlations under Dobrushin's uniqueness condition and its applications. Commun. Math. Phys. **84**, 207–222 (1982)
15. Künsch, H.: Asymptotically unbiased inference for Ising models. J. Appl. Probab. **19**, 345–357 (1982)
16. Künsch, H.: Discussion on Besag's paper [3]
17. Lippmann, A.: A maximum entropy method for expert systems. Brown University Thesis (1986)
18. Métivier, M., Priouret, P.: Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant. Probab. Th. Rel. Fields **74**, 403–428 (1987)
19. Xanh, N.X., Zessin, H.: Ergodic theorems for spatial processes. Z. Wahrscheinlichkeitstheor. Verw. Geb. **48**, 133–158 (1979)
20. Pickard, D.: Asymptotic inference for an Ising lattice. J. Appl. Probab. **13**, 486–497 (1976)
21. Pickard, D.: Asymptotic inference for an Ising lattice, II. Adv. Appl. Probab. **9**, 479–501 (1977)
22. Pickard, D.: Asymptotic inference for an Ising lattice, III. J. Appl. Probab. **16**, 12–24 (1979)
23. Pickard, D.: Inference for general Ising models. J. Appl. Probab. **19**A, 345–357 (1982)
24. Possolo, A.: Estimation of binary Markov random fields. University of Washington, Technical Report (1986)
25. Preston, C.: Random fields. (Lect. Notes Math., vol. 534) Berlin Heidelberg New York: Springer 1976
26. Simon, B.: A remark on Dobrushin uniqueness theorem. Commun. Math. Phys. 183–185 (1979)
27. Younes, L.: Couplage de l'estimation et du recuit pour des champs de Gibbs. C. R. Acad. Sci. Paris, Ser. I, **303**, 659–662 (1986)
28. Younes, L.: Estimation and annealing for Gibbsian fields. Ann. Inst. Henri Poincaré **24**, 269–294 (1988)
29. Younes, L.: Estimation pour des Champs de Gibbs et application au traitement d'images. Université Paris-Sud, Thesis (1988)

Received April 1, 1988; in revised form January 15, 1989