

Parameter Estimation For Imperfectly Observed Gibbs Fields and Some Comments on Chalmond's EM Gibbsian Algorithm.

Laurent Younes

28 January 1991

Abstract

We make a short review of existing algorithms for parameter estimation from imperfectly observed Gibbs field. We then focus on one of these methods: the EM Gibbsian algorithm, making new comments and simulations.

⁰Université Paris Sud. Laboratoire de statistique appliquée. Bat 425. 91405 ORSAY Cedex (France) and DMI, Ecole Normale Supérieure.

1 Introduction.

In this paper, we are concerned with general parameter estimation methods for partially observed Markov random fields (MRF) (by general methods, we mean methods that are valid for a reasonably wide class of models). This issue is of main importance for numerous problems that involve MRF (we particularly have in mind the Bayesian framework for image analysis developed in [15], which represents one of the most powerful and one of the most popular methods for low level processing of pictures).

Unfortunately, this problem is quite awkward, and there does not exist, as far as we are informed, totally satisfactory methods designed for it. The point is that, due to the computational restrictions that are inherent to MRF modeling, standard algorithms of parameter estimation for hidden models (such as EM) are almost useless in this context. Beside various clever solutions to this problem for very particular models, (cf [13], [5], and other references given in [17]), we are aware of only three general algorithms for it : stochastic gradient (SG) algorithm for solution of maximum likelihood equations ([47], [48]) , modification of variational estimator (VE) for partially observed data ([1]), and the so-called EM Gibbsian (EMG) algorithm ([6]). In the following, we shall briefly present the first two of them, and then, focus on the third making some remarks concerning its properties that did not appear in the original paper of Chalmond. This will be done in sections 3 and 4. Before that, in section 2, we fix some notations and present the model that we use.

2 Model.

We consider MRF on the regular lattice \mathbf{Z}^d : let F be a set, the single site state space, and denote by Ω the set of all configurations $x = (x_s, s \in \mathbf{Z}^d)$. If $D \subset \mathbf{Z}^d$, we denote Ω_D the set of all $x_D = (x_s, s \in D)$. We assume that F is provided with a measure μ_1 and denote by μ (resp. μ_D) the corresponding product measure on Ω (resp. Ω_D).

Let $\Theta \subset \mathbf{R}^\nu$ be the set of parameters, and consider a family of functions: $(h_C, C \subset \mathbf{Z}^d, C \text{ finite})$, where h_C is defined on Ω and takes values in \mathbf{R}^ν , and $h_C(x)$ only depends of coordinates $x_s, s \in C$. In all the sequel, we shall assume that h_C vanishes for all C of diameter larger than a given non-negative number γ (bounded range). This assumption simplifies notations and proofs and is absolutely necessary for any of the preceding methods to be feasible.

For $\theta \in \Theta$ we define the potential:

$$\lambda_\theta = (\langle \theta, h_c \rangle, C \subset \mathbf{Z}^d).$$

($\langle \cdot, \cdot \rangle$ being the usual euclidian scalar product).

Let $\mathcal{G}(\theta)$ be the set of Gibbs fields associated to this potential; this means that, if D is a finite subset of \mathbf{Z}^d , and $x' \in \Omega_{D^c}$ is given ($D^c = \mathbf{Z}^d \setminus D$), a regular version of the conditional law for $\pi \in \mathcal{G}(\theta)$ on Ω_D given x' outside D has a density with respect to μ_D which is given by:

$$\pi(x|x') = \exp(-\langle \theta, H(x|x') \rangle) / Z_\theta(x') \quad (1)$$

where $H(x|x') = \sum_{C \subset \mathbf{Z}^d} h_c(x.x')$, $x.x'$ being the element of Ω for which the coordinate at site s is x_s if $s \in D$ and x'_s if not (and it is assumed that the normalizing constant Z_θ is well defined, ie. that the exponential is μ_D integrable).

Although it has no importance in practice, we shall need, for rigorous results, to assume that h_C is space homogeneous: let's denote T_s the translation operator in the direction s , ie. $(T_s x)_t = x_{s+t}$, we assume that $h_{C+s} = h_C \circ T_s$.

In the case of perfect observation, the inference problem is the following: estimate θ_* from the restriction to Ω_D of some realization $x \in \Omega$ of one law $\pi_{\theta_*} \in \mathcal{G}(\theta_*)$. We assume that there exists at least one law in $\mathcal{G}(\theta_*)$, (this is a consequence of our assumptions when F is finite), but note that there may exist several ones. Identification is possible provided $\mathcal{G}(\theta)$ and $\mathcal{G}(\theta')$ are disjoint for distinct $\theta, \theta' \in \Theta$ (note however that it is in general impossible to estimate π_{θ_*} itself on the basis of a single realization of the field). In this context, several estimation methods exist, the most popular being the maximum pseudo likelihood (MPL) estimator (see section 3.3.2), other are maximum likelihood using stochastic gradient ([46]) which involves more computation while remaining feasible and has the advantage to be efficient, variational estimators ([1]), logit estimator for binary fields ([39]), various estimators based on approximations of the likelihood in the Gaussian case ([26], [23]), etc... All these estimators are proven to be consistent (this means that if D tends to \mathbf{Z}^d in a reasonable way, the corresponding estimators converge to the true θ_*). See for example [18], [21], [22], [9] for various results concerning minimum of contrast estimation with this context.

In the case of partial observations, the situation is the following. Let F^h and F° be some sets, and μ_1^h, μ_1° be measures on these sets. Denote Ω^h (resp. Ω°) the set of configurations on \mathbf{Z}^d with coordinates in F^h (resp. F°), and let $F = F^h \times F^\circ$, $\mu_1 = \mu_1^h \otimes \mu_1^\circ$ be the one site state spaces and measures for the joint process. We are therefore given a Gibbs field on the corresponding set $\Omega = \Omega^\circ \times \Omega^h$. Inference has to be made on the basis of the observation of the restriction to Ω_D° of the second coordinate, $y \in \Omega^\circ$ of some realization $(x^h, y) \in \Omega$ of a law $\pi_{\theta_*} \in \mathcal{G}(\theta_*)$ (we shall use letter y as well as x° to denote the observed data).

Let's remark that for the VE, F^h should be \mathbf{R}^k , whereas for the EMG estimator, F^h has to be a finite set with a small number of elements. Concerning the SG method, all kind of state spaces may be used, provided some good Monte Carlo simulation algorithm is available, but all rigorous results related to it were proven only for finite (or compact) F^h . Consistency for maximum of likelihood estimator in this context is proven in [8] and [48].

In both cases of complete and incomplete observations, it is of course possible to assume that, in the preceding model, part of the coordinates of the true parameter are known. For example, for noisy data, the noise mechanism can have been estimated before. Most of the time, we shall not consider this, because this would introduce extra notations. But, since it is straightforward to modify all the equations for this case, we shall make references to this situation in the remarks.

If π is some field over Ω , we shall denote by π° (or ψ) and π^h its marginals on Ω° and Ω^h .

In almost all cases, the marginal on Ω_D of some $\pi_\theta \in \mathcal{G}(\theta)$ is impossible to compute. The first reason is that, if we are only given θ , a precise description of $\mathcal{G}(\theta)$ is seldom available, so that π_θ itself cannot be described. The second reason is that, even if we know exactly what π_θ we are dealing with (for example in the absence of phase transition), there is no close form for the expression of probability on Ω_D in term of the potential (which is the only thing provided with the model). Therefore, in practice one is forced to use some approximate expression of the likelihood, which is most of the time of the kind (1) with some fixed boundary condition x' (variants are free

boundary condition, periodic boundary condition, differing only in terms involving coordinates near D^c). Consistency of maximum of likelihood estimation, for any of the preceding variants is not affected. It is unfortunately not the case for asymptotic normality and efficiency (see [22]), unless some more stringent conditions are imposed, such as Dobrushin's ones for unicity of the field, and suitable modifications are made in the energy $H(x|x')$ to eliminate edge effects (see [49]).

In the following, we assume that we have chosen some $x' \in \Omega_D$ for each D to approximate π_θ by $\pi_\theta(\cdot|x')$, and we shall not, unless necessary, let this appear in the notations, keeping abusively the expression π_θ , or π_θ^D . For each D , we are therefore using as raw materials for estimation a probability measure on Ω_D of the kind:

$$\pi_\theta^D(x^h, x^o) = \exp(-\langle \theta, H(x^h, x^o) \rangle) / Z_\theta.$$

We shall denote by E_θ expectations with respect to this law, by ψ_θ its marginal law over the observed space Ω_D^o , and for a given y , by π_θ^y the conditional law on Ω_D^h given that the field on Ω_D^o is y . We shall also use (X^h, Y) to designate a field on Ω .

Implementation of the algorithms requires some methods for simulating Markov random fields. They generally consist in a sequence of local updatings of the configurations ("local" means that only one or a small number of sites are changed), using transition kernels that are in detailed balance with the law to simulate. There also exist non local methods, such as multi-grid Monte Carlo algorithms, Swendsen-Wang algorithms, that are more efficient for some classes of models. See [44] for a precise description of these algorithms, as well as for a bibliography. All these methods are dynamic, which means that they provide a Markovian sequence of configurations that converges in distribution.

3 General methods for inference on partial observations

3.1 Maximum of likelihood.

Let y be the observation. A straightforward computation shows that the maximum of likelihood estimator is a solution of the equation

$$E_\theta(H) - E_\theta(H|y) = 0. \tag{2}$$

(In the case of perfect observation, this equation simplifies in: $E_\theta(H) - H(y) = 0$).

Computation of expectations with respect to some Gibbs field can only be done by simulation and such an operation is time consuming. Solving this equation, even for perfect observations is therefore a non trivial problem. If one is sure to use the same model a large number of times, and if the dimension of θ is small (1 or 2), it is possible to draw once for all the function $f(\theta) = E_\theta(H)$ and use this for subsequent estimation (see [18] for such an attempt). However, this cannot be done in general, in particular in the case of partial observation. We thus are generally reduced to using a gradient like algorithm to solve (2). An exact gradient descent algorithm would be

$$\theta_{n+1} = \theta_n + a[E_{\theta_n}(H) - E_{\theta_n}(H|y)].$$

Evaluating with a reasonable precision these expectations at each step requires a lot of computation time, but such an approach was successfully used in [30] for a maximum of entropy estimation

that led to similar equations. It is however possible to significantly improve on this by applying ideas from stochastic approximation techniques. Basically, the method is to mix the gradient descent with Monte Carlo simulation, without necessarily waiting for stabilization of the estimated expectations. A short formulation of this kind of algorithms is the following. Let $P_\theta(x, x')$ (resp. $P_\theta^y(x, x')$) be a transition kernel corresponding to some elementary part of the chosen Monte Carlo simulation algorithm for the distribution π_θ (resp. π_θ^y), for example, one site updating or one sweep of a Gibbs sampler. Then define the sequence (θ_n, X_n^1, X_n^2) by fixing $\theta_0 \in \Theta$, $X_0^1, X_0^2 \in \Omega_D^h$, and by the following stochastic algorithm:

$$\begin{aligned} P(X_{n+1}^1 = x' | X_n^1 = x) &= P_{\theta_n}(x, x') \\ P(X_{n+1}^2 = x' | X_n^2 = x) &= P_{\theta_n}^y(x, x') \\ \theta_{n+1} &= \theta_n + \frac{a}{n+b} [H(X_{n+1}^1) - H(X_{n+1}^2)] \end{aligned} \tag{3}$$

In [48] and [49] details are given on practical remarks about the implementation of the algorithm, and improvements that can be done. Applying results and methods from the stochastic approximation literature, (see [2]), it is possible to prove some rigorous facts concerning convergence. In particular, when observation is complete (in which case the second process X_n^2 disappears), θ_n converges almost surely, provided the constant a is small enough. In the general case, one can prove a local result of quasi convergence, which is much weaker. See [46], [47] for details. Recently, some stochastic algorithms have been studied for global optimization of functions for a context that includes ours ([24], [27], [14]); they are in some way random perturbations of (3) reminiscent of annealing algorithms for discrete or continuous optimization ([15], [20]). They might give interesting results in this context although we are not aware of any related numerical experiments.

Finally, we point out that, since the conditional law given the observations is simulated during (3), it is possible, in the context of a Bayesian analysis of images, to perform some reconstruction on line, using the MPM criterium (cf. [30]). This possibility is also available for both algorithms that follow.

3.2 Variational Estimator.

This estimator, which is preferably designed for complete observations, is mainly valid for $F = \mathbf{R}^n$ and $\mu_1 = \text{lebesgue measure on } F$. We give of it a summarized version, referring to [1] for details and variants. In the case of perfect observations, the principles are as follows. The first step is to choose statistics

$$(W_s^\alpha, \alpha = 1, \dots, \nu, s \in D)$$

which are local (ie. $W_s^\alpha(x)$ depends on x_t for t in a restricted neighborhood of s), and such that the divergence theorem holds:

$$\int_{\Omega_D} \{\nabla \cdot [W_D^\alpha(x) \pi_\theta(x)]\} dx = 0$$

where W_D^α is the $|D|$ -dimensional vector with components equal to W_s^α .

Developing these identities and replacing expectations with respect to π_θ with empirical ones yields a set of linear equations in θ , and the variational estimator is defined as a solution of these

equations. One possibility for W_s^α is $\partial H_s^\alpha / \partial x_s$, where $H_s = \sum_{C|s \in C} h_C$ and H_s^α is the α -th component of H_s . The obtained equations, that are very simple to solve, are proven to provide consistent estimators.

In the case of partial observations, the W 's are chosen so that

$$\int_{\Omega_D} \{\nabla \cdot [W_D^\alpha(y) \psi_\theta(y)]\} dy = 0$$

(the state space for y still being \mathbf{R}^n).

This leads to the equations :

$$\sum_{i=1}^{\nu} \theta_\beta [W^\alpha(y) \cdot E_\theta(\nabla H^\beta | Y = y)] = \nabla \cdot W^\alpha(y). \quad (4)$$

It is not obvious to obtain some natural way of finding W in this context. There is no proof of consistency for estimators obtained from (4), but authors in [1] report satisfying numerical results, at least for small noise.

Note that for solving (4), it is necessary to use a (stochastic) Newton algorithm, since conditional expectations given Y generally cannot be put into close form. This algorithm, however, is less complex than the one used for maximum of likelihood estimation. This is balanced by the fact that (4) does not come from the gradient of some function to maximize, and this can yield more risk of instability. Note also that in both cases, it is likely that several solutions of the equations exist.

3.3 EM algorithm algorithm and EM Gibbsian algorithm.

3.3.1 EM algorithm.

The EM algorithm is a way for obtaining local maxima of the likelihood for partial observations. For several applications, it is a very efficient tool, widely used, especially for one dimensional hidden Markov chains.

As said before, the maximum likelihood estimator solves the equation:

$$E_\theta \left[\frac{d}{d\theta} l(\theta) | Y = y \right]$$

where $l(\theta)$ is the logarithm of the likelihood. The principle of the EM is as follows : start with some parameter θ_0 , and construct θ_{n+1} from θ_n as a maximizer of:

$$g_n(\theta) = E_{\theta_n} \left[\frac{d}{d\theta} l(\theta) | Y = y \right] \quad (5)$$

For exponential models, under some identifiability condition, g_n is strictly concave and this provides a well defined algorithm. One can prove that the sequence θ_n is such that $\psi_{\theta_n}(y)$ is increasing. Thus, if θ_0 is suitably chosen, the algorithm will converge to the maximum of likelihood, while wrong choices of θ_0 can lead to a non global maximum, or even divergence. In the absence of *a priori* knowledge on the true value of the parameter, it is generally recommended to try several θ_0 .

We now specialize to the case of partially observed MRF. The difficulty is of course to maximize g_n , which is equivalent to solve the equation:

$$E_\theta(H) = E_{\theta_n}(H|y). \quad (6)$$

Computation of $E_{\theta_n}(H|y)$ requires Monte Carlo simulation, and the solution of (6) can hardly be obtained without a stochastic gradient algorithm. We propose a procedure for that purpose, in the vein of what has been done in section 3.1.

Let θ_0 be a starting point, and define the doubly indexed sequences $(\theta_{n,k} \ n \geq 1, 0 \leq k \leq K(n))$, $(X_{n,k}^1)$, $(X_{n,k}^2)$ by :

- i) $\theta_{n,0} = \theta_n = \theta_{n-1, K(n-1)}$.
- ii) $(X_{n,k}^2, \ k = 1, \dots, K(n))$ is an ergodic sequence of configurations that converge in distribution to $\pi_{\theta_n}(\cdot|y)$.
- iii) $\theta_{n,k+1} = \theta_{n,k} + \gamma_k [H(X_{n,k+1}^1) - \hat{E}_{\theta_n}]$
 where $(X_{n,k+1}^1)$ is obtained from $(X_{n,k}^1)$ by making one step of some Monte Carlo simulation algorithm for $\pi_{\theta_{n,k}}(\cdot)$, γ_k is a decreasing sequence of numbers, and \hat{E}_{θ_n} is an empirical estimation of $E_{\theta_n}(H|y)$ on the basis of the sequence $(X_{n,k}^2, \ k = 1, \dots, K(n))$.

We believe that some rigorous study of this algorithm, using results concerning stochastic approximations (cf [2]) might be feasible, maybe requiring some technicalities. We shall however not get into that direction, mainly because the full application of the procedure would be quite unrealistic in practice, because of the large number of computations it involves ($K(n)$ should be large enough so that $\theta_{n, K(n)}$ should be close to the maximizer of g_n). On the other hand, this algorithm is surely more stable than the one in section 3.1, since it mimics the EM algorithm rather than a gradient algorithm. In the case of high noise, when the latter performs poorly, it might be fruitful to make a few iterations of the former as a preliminar analysis. This would have the effect to place the initial point of the SG algorithm deeper into some attraction basin of the log-likelihood, increasing its trend to convergence (the risk of terminating in a non global maximum is of course still the same).

In [6], Chalmond proposed a modified procedure for the case of finite F^h with a small number of elements, in which the maximization step is simplified. He proposed to replace the likelihood in the expression of g_n (cf (5)) by the pseudo likelihood, of which we now recall the definition.

3.3.2 Pseudo likelihood.

Pseudo likelihood has been introduced in the case of complete observations as an alternative to maximum of likelihood estimation. It requires much less computation at the risk of losing some efficiency. It has been inspired by the coding technique that Besag ([3]) developed for fields with bounded range. For such fields, it is possible to extract some maximal subset of \mathbf{Z}^d such that the variables at sites from this sublattice are mutually independent given variables from the rest of the sites. Let's denote this sublattice by L , the coding estimator associated to L is a maximizer of the logarithm of the conditional probability on $L \cap D$, given the data at $D \setminus (L \cap D)$:

$$p_L(\theta) = \sum_{s \in L \cap D} \log \pi_\theta(x_s | x_t \ t \neq s)$$

where $\pi_\theta(x_s|x_t, t \neq s)$ is the conditional law at site s given the rest of the sites.

Noting that efficiency can be expected to be improved if a sum of p_L 's for several L 's is maximized, it is natural to maximize the function (pseudo likelihood):

$$p(\theta) = p_D(\theta) = \sum_{s \in D} \log \pi_\theta(x_s|x_t, t \neq s)$$

Since conditional probabilities for bounded range models are easy to deal with, this function can be maximized with the help of some deterministic gradient descent algorithm. This method does not naturally generalize to the case of incomplete observations.

3.3.3 EM Gibbsian algorithm.

As said before, this algorithm essentially consists in replacing $l(\theta)$ by $p(\theta)$ in (5). We now describe it, with a parametric formulation a little bit different from the one given in [6].

The function g_n is now

$$g_n(\theta) = E_{\theta_n} \left[\frac{d}{d\theta} p(\theta) | Y = y \right] \quad (7)$$

Let's put:

$$H_s(x) = \sum_{c/ s \in C} h_c(x)$$

and

$$K_s(x) = \sum_{c/ s \in C} \frac{1}{|C|} h_c(x)$$

We have: $H(x) = \sum_{s \in D} K_s(x, x')$. Since the boundary condition x' will have negligible influence on what will follow, we shall make the abuse of notation: $H(x) = \sum_{s \in D} K_s(x)$.

We also have: $\frac{d}{d\theta} p(\theta) = \sum_{s \in D} \hat{H}_s(\theta) - H_s$, where

$$\hat{H}_s(\theta, x) = E(H_s | x_t, t \neq s)$$

(if $f(\theta, x)$ is a function on $\mathbf{R}^\nu \times \Omega$, we denote by $f(\theta)$ the function on Ω that associates $f(\theta, x)$ to x).

The maximization step is therefore done by solving:

$$E_{\theta_n} \left[\sum_{s \in D} \hat{H}_s(\theta) - H_s | Y = y \right] = 0$$

This is not completely trivial because \hat{H}_s can depend in a non trivial way on θ . This however can be simplified when $|F^h|$ is small as well as the range of the field. This implies that $\hat{H}_s(\theta, x)$ will only depend on a few coordinates x_t for t in a neighborhood $V(s)$, of s , and will therefore take a reasonable number of different values. It is thus possible, by considering the different configurations $(x_t^h, t \in V(s))$, to divide $\Omega^y = \Omega^h \times \{y\}$ into K subsets $\Omega_l^y(s)$, $l = 1, \dots, K$, independent of θ such that $\hat{H}_s(\theta, x) - H_s$ takes the constant value $f_{l,s}(\theta, y)$ on $\Omega_l^y(s)$. Since it depends on conditional probabilities for π , $f_{l,s}(\theta, y)$ can be explicitly expressed in terms of θ, y . Because of the bounded

range, the property for a configuration to be element of $\Omega_t^y(s)$ will depend on its coordinates in a small neighborhood of s . Moreover, because of stationarity, one can arrange these sets so that:

$$\Omega_t^y(s+t) = T_s \Omega_t^y(s)$$

and

$$f_{l,s+t}(\theta, y) = f_{l,t}(\theta, T_s y).$$

It is indeed always possible to take as many $\Omega_t^y(s)$ as configurations $(x_t^h, t \in V(s))$, but symmetry properties for particular models can allow to choose larger $\Omega_t^y(s)$, and thus smaller K .

With these notations, the equation to be solved becomes:

$$\sum_{s \in D} \sum_{l=1}^K f_{l,s}(\theta, y) \pi_{\theta_n}[\Omega_t^y(s)|y] = 0 \quad (8)$$

The $\pi_{\theta_n}[\Omega_t^y(s)|y]$ have to be computed by Monte Carlo simulation. The final equation can be solved by using standard deterministic procedures. We illustrate this with an example at the end of section 4.

In [6], the treatment is somewhat different. The conditional law of the observed field given the original one, ie. $P(Y = y | X^h = x^h)$ is assumed to be of the kind:

$$\prod_{s \in D} p(y_s | x_s). \quad (9)$$

The model is overparametrized by the family $p(y|x)$, for $y \in F^o$, $x \in F^h$ (both these sets being assumed finite; a different approach, consistent with ours, is used for Gaussian noise), and by the family of all local specifications of the law of the hidden field X^h . This yields explicit formulas for updating, once similar quantities to the $\pi_{\theta}[\Omega_t^y(s)|y]$ have been computed. At the end of the algorithm, the parameter θ_* is computed on the basis of local specifications by a least square procedure. We refer to [6] for details. Our approach, while requiring some extra computation, has the advantage of being more consistent with the parametric formulation; moreover, it permits more flexibility in the structure of the noise, since there is no need for assumptions like (9).

In the next section, we explore a little more some features of this algorithm.

4 Comments and simulation results on the EM Gibbsian algorithm.

The first remark that can be made, is that, if the algorithm converges, it provides a solution of the equation

$$\begin{aligned} E_{\theta} \left[\frac{d}{d\theta} p_{\theta} | y \right] &= 0 \\ \text{ie. } E_{\theta} \left[\sum_{s \in D} (\hat{H}_s(\theta) - H_s) \right] &= 0 \end{aligned} \quad (10)$$

We are therefore dealing with a moment estimator. Remark that there are similarities between the preceding equation, and equation (4) for the variational estimator. A stochastic gradient algorithm could also be used, as an alternative to the EMG for solving (10).

Several questions arise at this point. The first one is whether equation (10) has or not a unique solution. The answer is likely to be false, since this is not the case for the maximum of likelihood equations. The second question is whether the method provides a consistent estimator. In view of the first point, we do not expect to obtain more than a local result of consistency, of the kind: there exists a fixed neighborhood, \mathcal{V} , of θ_* such that, for large enough D , there exists only one solution of (10) in \mathcal{V} , denoted by $\hat{\theta}_D$, and $\hat{\theta}_D$ converges to θ_* when D tends to \mathbf{Z}^d . The last issue is of course: does the EM Gibbsian algorithm converge? there again, we should not expect more than local results, saying, for example that the algorithm converges provided its initial point lies in a small enough neighborhood of a solution of (10). Unfortunately, we were not able to give any positive answer to these questions, except under the quite restrictive hypothesis that are made in propositions 1 and 2. We shall then present a few numerical results.

Let D be fixed for a moment. Let's denote by $\phi(\theta_1, \theta_2)$ the function $E_{\theta_1}[p'(\theta_2)|y]$, where p' is the derivative of the logarithm of the pseudo likelihood. Step n of the EMG algorithm requires therefore the solution of

$$\phi(\theta_n, \theta) = 0.$$

In the following we assume that quantities that have to be estimated with simulation are exactly computed. We just study the deterministic part of the algorithm, which has its own complexity. To take into account perturbations due to imperfect estimation of these quantities would require to introduce concepts related to stochastic approximation. However, it is clear from this theory that, since estimation can never be perfect, the sequence θ_n computed in the algorithm will have some "large deviation" behaviour, preventing it from converging. In good cases, this theoretical limitation will never be seen in practice (because time required for "explosion" is very large), but it can happen (and it indeed happens, see the end of section 4) that this behaviour becomes typical. To remedy to this, one possibility is to replace θ_{n+1} by $\bar{\theta}_{n+1} = \theta_n + \frac{a}{n+b}(\theta_{n+1} - \theta_n)$ (reminiscent of stochastic approximation with decreasing steps), and, if necessary, to determine some suitable compact set in which the parameter must lie, and prevent θ_n to go out of it.¹ At this point, it might even be better to directly use a stochastic gradient algorithm of the kind:

$$\theta_{n+1} = \theta_n + \frac{a}{n+b}(\hat{\phi}(\theta_n, \theta_n))$$

where $\hat{\phi}(\theta_n, \theta_n)$ is an approximation of $\phi(\theta_n, \theta_n)$ computed by Monte Carlo simulation. Following ideas of section 3.1, this approximation can be rough provided a is small enough and simulation and updating of the parameter are mixed.

We first study the behavior of ϕ with regard to the second coordinate. Note that if θ_1 is fixed, $\phi(\theta_1, \theta)$ is the derivative of $E_{\theta_1}[p(\theta)|y]$, which is concave in θ . Indeed, we have, by a straightforward computation:

¹In fact, this procedure (in the case when it converges) does not exactly provide a solution of (10). If, at each step, a sequence $X = (X_1, \dots, X_N)$ of configurations is simulated to estimate all probabilities that are involved, the stable point of the algorithm is a solution of $\bar{E}_\theta(\hat{\eta}(\theta, X)) = \theta$, where \bar{E} means expectation for the law of the sequence X when simulation is made with parameter θ , and $\hat{\eta}(\theta, X)$ is the solution τ of the approximated version of equation $\phi(\tau, \theta) = 0$ based on the simulated data X . If N is reasonably large, this limit can be expected to be very close to a fixed point of η (see below for a definition of η).

$$\frac{\partial \phi}{\partial \theta_2} = -E_{\theta_1} \left\{ \sum_{s \in \mathcal{D}} \text{var}_{\theta_2} [H_s | X_t, t \neq s] | y \right\}$$

We shall assume that there exists no non zero solution $\tau \in \mathbf{R}^\nu$ of the system of equations:

$${}^t \tau \text{var}_{\theta_2} [H_s | x_t, t \neq s] \tau = 0, \text{ for } x \in \Omega.$$

Note that the number of equations is finite, because of the bounded range assumption, and that, by stationarity, this set of equations does not depend on s . This is equivalent to the assumption that there is no solution to the system ${}^t \tau H_s(x) = 0$, for $x \in \Omega$ (because all conditional probabilities are positive). Since all marginal probabilities on finite domains for the law given y are positive, and since $\text{var}_{\theta_2} [H_s | x_t, t \neq s]$ only depends on a finite number of coordinates x_t , this assumption implies that $E_{\theta_1} [p(\theta) | y]$ is a strictly concave in θ , and one obtains a well defined function, that we denote $\eta(\theta_1) = \eta_y(\theta_1)$ by:

$$\theta_2 = \eta(\theta_1) \iff \phi(\theta_1, \theta_2) = 0.$$

The EMG algorithm is thus an iterative procedure defined by : $\theta_{n+1} = \eta(\theta_n)$.

Let $\hat{\theta}$ be a solution of (10), or equivalently a fixed point of η , a sufficient condition for the existence of a neighborhood \mathcal{V} of $\hat{\theta}$ such that $\theta_0 \in \mathcal{D} \Rightarrow \theta_n \rightarrow \hat{\theta}$, is that the largest eigenvalue of $\eta'(\hat{\theta})$ is smaller than one. Here we have

$$\eta'(\hat{\theta}) = - \frac{\partial \phi}{\partial \theta_2}^{-1} \Big|_{(\hat{\theta}, \hat{\theta})} \cdot \frac{\partial \phi}{\partial \theta_1} \Big|_{(\hat{\theta}, \hat{\theta})}$$

where

$$\frac{\partial \phi}{\partial \theta_1} \Big|_{(\hat{\theta}, \hat{\theta})} = \text{cov}_{\hat{\theta}} \left\{ \sum_{s \in \mathcal{D}} [H_s - \hat{H}_s(\hat{\theta})], \sum_{t \in \mathcal{D}} K_t | y \right\}.$$

Two simple remarks can be made at this point.

- 1) In a situation of "maximal noise", when the conditional law given y does not depend on y (for example if $\langle \theta, H \rangle$ can be written $\langle \bar{\theta}, H_1(x^h) \rangle + \langle \alpha, H_2(x^h, y) \rangle$ and α is *known* to be 0), we have $\phi(\theta, \theta) = 0$ for all θ , and η' is the identity matrix, whose eigenvalues are 1. So, in this very bad situation, these values are at the boundary of their validity domain. Of course, there is no reason for this situation (which is the worst, for example in the case of maximum of likelihood estimation) to be the worst for the EMG case.
- 2) The opposite situation is when there is a one to one relation between Y and X . With our modeling, this typically corresponds to cases when some parameters tend to infinity. In that case, we obviously have $\eta' = 0$ (we compute pseudo maximum of likelihood for perfect observations), and there therefore exists some neighborhood of these (maybe infinite) parameters for which EMG algorithm is valid. In other terms, this algorithm can be used for large enough signal to noise ratio.

In this last remark, the validity domain can depend on D . To be able to make some uniform asymptotics, we shall need some additional assumptions.

Let y be given. For $D \subset \mathbf{Z}^d$, and some configuration $x' \in \Omega_{D^c}^h$, we define the law $\pi_{\theta, D}^y$ on Ω_D^h by

$$\pi_{\theta, D}^y(x|x') = \exp\{-\langle \theta, H_D[(x^h, y_D)|(x', y_{D^c})] \rangle\} / Z_{\theta, D}(y) \quad (11)$$

where H_D has been defined near equation (1), and y_V is the restriction of y to the subset V of \mathbf{Z}^d .

This is the ‘‘approximate conditional neighborhood’’ with edge condition x' . The family $\pi_{\theta, D}^y(\cdot|x')$, $D \subset \mathbf{Z}^d$, $x' \in \Omega_{D^c}^h$ forms a consistent family of conditional laws. We shall assume that, for all y , the family satisfies Dobrushin’s unicity condition (cf [10]) which is based on bounds on the distances between $\pi_{\theta, \{s\}}(\cdot|x')$ and $\pi_{\theta, \{s\}}(\cdot|x'')$, for x' and x'' that differ on only one coordinate, and this for θ in a neighborhood of θ_* .

For example, assume that the energy has the following form (we omit D and x' in the notations):

$$\langle \theta, H(x^h, y) \rangle = \langle \theta_1, H_1(x^h) \rangle + \langle \theta_2, H_2(y) \rangle + \theta_3 \sum_s h_3(x_s^h, y_s).$$

(this is the case when the original field is degraded independently at each site). For this energy, Dobrushin’s conditions are true provided θ_1 is small enough, or, in the case when F° is finite and when for all $y_s \in F^\circ$, $h_3(\cdot, y_s)$ has a unique minimum, when θ_3 is large enough. These conditions are very convenient to use, and yield explicit bounds (cf [43]). The weaker conditions given in [11] would be sufficient, but they appear to be very difficult to check (even with the help of a computer), because of the non-homogeneity of $\pi_{\theta, D}^y$.

We assumed these conditions for all y . In fact, it would be enough to assume them for ψ_{θ_*} , almost all y , where ψ_{θ_*} is the true (unknown) law of the observations.

With these assumptions, there exists a neighborhood \mathcal{V} of θ_* such that for all $\theta \in \mathcal{V}$ the following is true

- i) There exists a unique field, denoted π_θ^y associated to this family, that is, for all $D \subset \mathbf{Z}^d$, finite:

$$\pi_{\theta, D}^y(x^h) = \int_{\Omega_{D^c}^h} \pi_{\theta, D}^y(x|x') \pi_{\theta, D^c}^y(dx')$$

where $\pi_{\theta, D}^y$ is the marginal of π_θ^y on Ω_D^h and we made the abuse of notation of writing $\pi_{\theta, D}^y(x^h)$ in spite of $\pi_{\theta, D}^y(x_D^h)$ (and the same for D^c).

- ii) Let T_s be the shift operator on configurations (we use the same notation whenever this operator acts on Ω^h , Ω° or Ω). We have $\pi_\theta^{T_s y} = \pi_\theta^y \circ T_{-s}$.

To see that, put $q = \pi_\theta^{T_s y} \circ T_s$.

We fix a finite D and a configuration $x^h = (x_t^h, t \in D) \in \Omega_D^h$, and denote by $T_{-s} x^h$ the element of Ω_{-s+D}^h for which the coordinate at site $-s+t$ is x_t^h . We omit to write the index θ .

$$\begin{aligned} q_D(x) = \pi_{-s+D}^{T_s y}(T_{-s} x^h) &= \int_{\Omega_{-s+D^c}^h} \pi_{-s+D}^{T_s y}(T_{-s} x^h | x') \pi_{-s+D^c}^{T_s y}(dx') \\ &= \int_{\Omega_{D^c}^h} \pi_{-s+D}^{T_s y}(T_{-s} x^h | T_{-s} x') q_{D^c}(dx') \\ &= \int_{\Omega_{D^c}^h} \pi_D^y(x|x') q_{D^c}(dx') \end{aligned}$$

(Note that these equalities do not rely on the uniqueness assumption; in the general case they imply relations between the sets of laws associated to families $\pi_{\theta,D}^y(\cdot|x')$ and $\pi_{\theta,D}^{T_s y}(\cdot|x')$; in the last one, we used translation invariance of the potential).

- iii) Correlations for π_{θ}^y decay exponentially fast, in the sense that if f_1 and f_2 are respectively measurable functions of x_{D_1} and x_{D_2} , with $D_1 \subset \mathbf{Z}^d$ finite and $D_2 \subset \mathbf{Z}^d$, we have:

$$\text{cov}(f_1, f_2) \leq A \|f_1\|_{\infty} \|f_2\|_{\infty} |\partial D_1| \exp(-g \cdot \text{dist}(D_1, D_2))$$

where A and g are positive constants that only depend on the potential, and ∂D_1 is the boundary of D_1 in \mathbf{Z}^d . We assume that they do not depend on θ (by restricting if needed the set \mathcal{V}).

- iv) Let $\pi_{\theta} \in \mathcal{G}(\theta)$; then π_{θ}^y is a version of the conditional law for π_{θ} given y .

Assuming this, we have the asymptotic behavior of the derivatives of ϕ :

Proposition 1 *Assume that $\pi_{\theta_{\star}}$ is ergodic. Then for all $\theta_1 \in \mathcal{V}$, $\theta_2 \in \Theta$, for $\pi_{\theta_{\star}}$ almost all y*

a)

$$\lim_{D \rightarrow \mathbf{Z}^d} \frac{1}{|D|} \sum_s E_{\theta_1, D}^y [\text{var}_{\theta_2}(H_s | X_t, t \neq s)] = E_{\theta_{\star}} \{E_{\theta_1}^y [\text{var}_{\theta_2}(H_0 | X_t, t \neq 0)]\}$$

The indices θ, D under expectations or covariances mean that they are computed accorded to some given approximate (conditional) marginal on D (the result is independent of the boundary condition). The limit $D \rightarrow \mathbf{Z}^d$ should be understood in the sense of Van Hove, or, for simplicity, for any sequence of d -dimensional cubes increasing to \mathbf{Z}^d .

b)

$$\lim_{D \rightarrow \mathbf{Z}^d} \frac{1}{|D|} \text{cov}_{\theta_1, D}^y \left[\sum_s H_s - \hat{H}_s(\theta_2), \sum_t K_t \right] = \sum_t E_{\theta_{\star}} \{ \text{cov}_{\theta_1} [H_0 - \hat{H}_0(\theta_2), K_t] \}$$

Boundary condition has an effect of order $\sqrt{|D|}$, which is negligible in the preceding limits.

To prove a), remark that, by i), the difference between $E_{\theta_1, D}^y [\text{var}_{\theta_2}(H_s | X_t, t \neq s)]$ and the exact expectation $E_{\theta_1}^y [\text{var}_{\theta_2}(H_s | X_t, t \neq s)]$ tends to 0; ii) and the ergodic theorem give the conclusion.

To prove b), remark that, by i) and iii), it suffices to consider the limit of

$$\frac{1}{|D|} \text{cov}_{\theta_1} \left[\sum_s H_s - \hat{H}_s(\theta_2), \sum_t K_t \right]$$

which, by ii) has the same limit as

$$\sum_t \frac{1}{|D|} \sum_s \text{cov}_{\theta_1}^{T_s y} [H_0 - \hat{H}_0(\theta_2), K_t],$$

and this converges to

$$\sum_t E_{\theta_{\star}} \{ \text{cov}_{\theta_1} [H_0 - \hat{H}_0(\theta_2), K_t] \}. \diamond$$

In addition, it can be proven by similar techniques that these functions are continuous in θ_1, θ_2 . We shall denote by W_1 and W_2 the expectations given in the right-hand terms of a) and b) in the preceding proposition, computed for $\theta_1 = \theta_2 = \theta_*$. Note that

$$W_2 = W_1 - \sum_t \text{cov}_{\theta_*} \{E_{\theta_*}^y[H_0 - \hat{H}_0(\theta_*)], E_{\theta_*}^y[K_t]\}.$$

We have the following (under the same hypothesis as in proposition 1)

Proposition 2 For π_{θ_*} , almost all y ,

a) If $W_1 - W_2$ is invertible, there exists a neighborhood of θ_* in which for all large enough D , there exists a unique solution $\hat{\theta}_D$ of $E_{\theta, D}^y[p'(\theta)] = 0$ and θ_D converges to θ_* .

b) if the largest eigenvalue (in modulus) of $W_1^{-1}W_2$ is strictly smaller than 1, there exists a neighborhood of θ_* such that any starting point in this neighborhood, the EMG algorithm converges to θ_D defined in a).

a) is a straightforward application of the inverse mapping theorem and of the fact that $\frac{1}{|D|}E_{\theta_*, D}[p'(\theta_*)|y]$ tends to 0 a.s. and b) is obvious from the preceding results. \diamond

Unfortunately, a) and b) in proposition 2 appear to be very difficult to check, even for very simple models. They are however true for *weak enough noise* (because of the continuity of W_1, W_2 in the parameters), with the advantage over remark 2 above that this statement is uniform with respect to $D \subset \mathbf{Z}^d$.

We have tried to check b) for a very simple non trivial case of imperfect observations on which some computation could be made. The original field is an isotropic nearest neighbor 2-D Ising field, and observation is made on even sites, ie. sites $(i, j) \in \mathbf{Z}^d$ such that $i + j$ is even. Even in that simple case, we couldn't go all the way down, relying on simulations at some point. Since computations are quite cumbersome, and this example has little practical interest, because we can precisely describe the law of the observations, and use any estimation method for complete observations, we do not give details here, just saying that, in trying to check $|W_2| < W_1$ (in that case, these quantities are real numbers), we had a rigorous proof of $W_2 < W_1$ and used simulation to check the other inequality that appeared to be true.

We conclude this paper by giving some numerical results in some situations that are closer to reality. All simulations (except when specified) are done on a 64×64 lattice.

The first one is the case of an Ising field with Bernoulli noise. In that case, the joint energy is:

$$U(x^h, y) = \sum_{\langle st \rangle} \beta x_s^h x_t^h + \alpha \sum_s y_s x_s.$$

Where the first sum is over all pairs of neighbor sites, and x_s and y_s take values in $\{-1, 1\}$. X^h is therefore a regular isotropic Ising field and Y is obtained by flipping each site independently with probability $p = e^\alpha / (e^\alpha + e^{-\alpha})$. We assumed that the value of α was known. To illustrate the description of section 3.3.3, we explicitly express the algorithm. Equation (7) becomes:

$$g_n(\theta) = \sum_s E_{\theta_n} [v_s(x_s + \tanh(\beta v_s)) | y]$$

where v_s is the sum of x_t for all t nearest neighbors of s .

Since x_s takes values $-1, +1$, and v_s is one of $-4, -2, 0, 2, 4$, we define 5 sets $\Omega_k(s)$, $k = 1, \dots, 5$, and the associated value $f_k(s)$ of $v_s(x_s + \tanh(\beta v_s))$ on these sets by:

- $\Omega_1(s) = \{(x_s, v_s) = (-1, 4) \text{ or } (x_s, v_s) = (1, -4)\}$, $f_1 = -4(1 - \tanh 4\beta)$.
- $\Omega_2(s) = \{(x_s, v_s) = (-1, 2) \text{ or } (x_s, v_s) = (1, -2)\}$, $f_2 = -2(1 - \tanh 2\beta)$.
- $\Omega_3(s) = \{(x_s, v_s) = (-1, 0) \text{ or } (x_s, v_s) = (1, -0)\}$, $f_3 = 0$.
- $\Omega_4(s) = \{(x_s, v_s) = (-1, -2) \text{ or } (x_s, v_s) = (1, 2)\}$, $f_4 = 2(1 + \tanh 4\beta)$.
- $\Omega_5(s) = \{(x_s, v_s) = (1, 4) \text{ or } (x_s, v_s) = (1, 4)\}$, $f_5 = 4(1 + \tanh 4\beta)$.

Put $\pi_k = \sum_s \pi_{\theta_s}[\Omega_k(s)]$, that can be computed by simulation; the equation to solve is

$$4(\pi_1 + \pi_5) \tanh 4\beta + 2(\pi_2 + \pi_4) \tanh 2\beta + 4(\pi_5 - \pi_1) + 2(\pi_4 - \pi_2) = 0.$$

We computed by simulation the mean values, that we still denote by W_1 and W_2 , of the partial derivatives of ϕ at point (β, β) , for several values of β and p . Even if propositions 1 and 2 were proven in situations that here correspond to small β or large α , we believe that these expectations still give interesting information in other cases. We present several figures that give the value of W_2/W_1 in function of α for fixed values of β . This value appears to always be between 0 and 1, but gets very close to 1 as soon as the noise becomes significant (note that these are mean values, and that we observed in the simulations situations when the observed value of η' was larger than 1, in which case the algorithm might diverge).

This case of imperfect observations was studied in [13], where estimators for β and p were proposed. This paper shows the same kind of feature for their method, and the estimation of β for high noise was also very unaccurate. We also could observe this difficulty while experimenting the algorithm of section 3.1 for that situation. Bernoulli noise seems much more critical for estimation than, for example, additive gaussian noise.

The second example concerns precisely Gaussian noise. The hidden field is still an isotropic Ising field, and the observed field Y_s is defined by its conditional law given x^h which is gaussian with mean μ and variance σ^2 if $x_s^h = 1$ and with mean μ' and variance σ'^2 if $x_s^h = -1$. We shall also consider a submodel for which $\sigma = \sigma'$ (this submodel includes additive Gaussian noise). We computed, for various values of the parameter, the largest eigenvalue of $W_1^{-1}W_2$. It is worthwhile to remark that for that model, this matrix is not the same for various laws in $\mathcal{G}(\theta_*)$, and we made the simulation with boundary condition equal to 1. This problem did not appear for the preceding model, because we had to estimate expectations of even functions. The mean values W_2 and W_1 are here meaningless in the case of non-ergodic π_{θ_*} , since they are not converging to the empirical expectations they are supposed to approximate. In fact, even when the true law π_{θ_*} is not ergodic, the relevant quantities are values of W_1 and W_2 for ergodic Gibbs fields in $\mathcal{G}(\theta_*)$ of which π_{θ_*} is a mixture.

In all our simulations, this largest eigenvalue was smaller than 1 (see table 4), apparently indicating some stability properties for the algorithm (note however that these values are very close to 1). But, computer experiments with this model show that the EMG algorithm does not converge, whereas it has a completely satisfying behavior for the submodel in which variances are

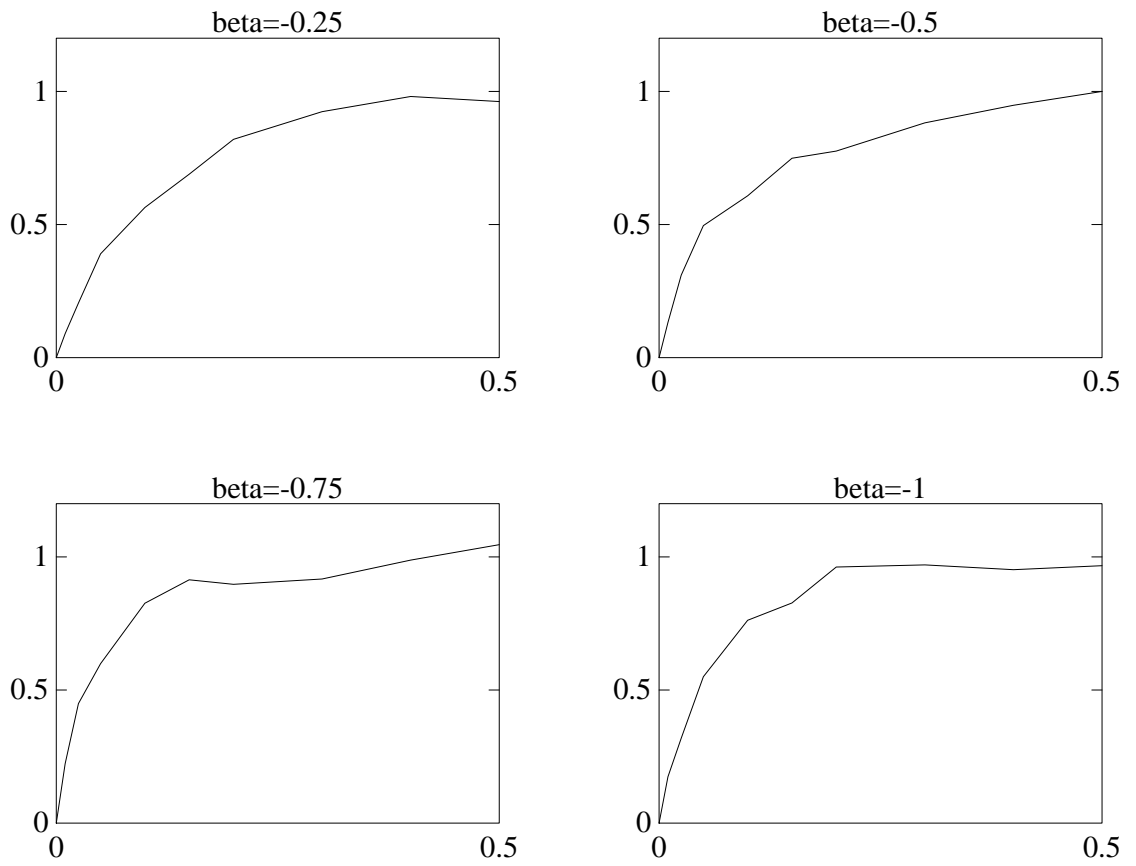


Figure 1: W_2/W_1 in function of α for 4 values of β

assumed to be equal. To understand this, several remarks can be made. First, propositions 1 and 2 are not valid for all range of the parameter, but EMG behaves poorly even for very weak noise, in which case they should apply. Secondly, proposition 1 and 2 are asymptotic results: large size of the lattice is needed for the estimated parameter $\hat{\theta}_D$ to be close enough of θ_* to ensure that W_1 and W_2 are good approximations for the derivatives of η at $\hat{\theta}_D$. The third reason for non convergence is related to the remark we made at the beginning of this section, concerning the problem of accuracy of the estimation of the $\pi^y(\Omega_k(s))$, and consequence it may have in divergence of the algorithm. This clearly appears in the expressions of the estimated μ' and σ' at step n which are:

$$\begin{aligned}\mu' &= \frac{\sum_s y_s \pi^y(x_s = -1)}{\sum_s \pi^y(x_s = -1)} \\ \sigma'^2 &= \frac{\sum_s y_s^2 \pi^y(x_s = -1)}{\sum_s \pi^y(x_s = -1)} - \mu'^2\end{aligned}$$

The fact that, for Ising fields below phase transition, the number of -1 is very small (edge condition is 1), implying imprecision in the estimation of $\pi^y(x_s = -1)$, yielding wrong updating of μ' and particularly of σ' (in the case of the sub-model, the variance is estimated using all sites and is therefore accurate, improving stability of the algorithm). In addition, we can note that if during the algorithm, it happens that $\mu = \mu'$ and $\sigma = \sigma'$, the field X^h does not depend on Y anymore, and the algorithm is lost and has no chance to find its way back on the right track.

To illustrate this we ran the EMG algorithm, estimating two variances for lattices of size 64×64 and 128×128 , and various numbers N of iterations of a Swendsen-Wang procedure to estimate the $\pi^y(\Omega_k(s))$. Here are the results ($\beta = -0.8$, $\mu = 5$, $\mu' = 9$); it seems hard to determine which of the preceding reasons is relevant (probably all of them).

- size: 128×128 , $\sigma = \sigma' = 2.5$ $N = 60$. After 21 iterations, $\sigma' = 0$ and the image degenerates in all 1's.
- size: 128×128 , $\sigma = \sigma' = 2.5$ $N = 100$. After 24 iterations, $\sigma' = 0$ and the image degenerates in all 1's.
- size: 128×128 , $\sigma = \sigma' = 1$, $N = 60$. After 24 iterations, one gets $\mu = \mu'$, $\sigma = \sigma'$, and the rest of the algorithm becomes independent on the observations...
- size: 128×128 , $\sigma = \sigma' = 1$, $N = 100$. The algorithm seems stable.
- size: 64×64 , $\sigma = \sigma' = 2.5$ $N = 60$. After 24 iterations, the image degenerates in all 1's.
- size: 64×64 , $\sigma = \sigma' = 2.5$ $N = 100$. After 16 iterations, the image degenerates in all 1's.
- size: 64×64 , $\sigma = \sigma' = 1$, $N = 60$. The algorithm seems to converge, with a very unaccurate value for σ'^2 : 0.05.
- size: 64×64 , $\sigma = \sigma' = 1$, $N = 100$. $\mu = \mu'$, $\sigma = \sigma'$ after 16 iterations.

When trying smaller values of $-\beta$ that yield a larger number for -1 's, performances are improved although the same kind of phenomena can still be observed from time to time.

Table 1: Largest eigenvalue, λ , of $W_1^{-1}W_2$ for various parameters

β	μ	μ'	σ^2	σ'^2	λ
-1	-1	1	2	2	0.93
-1	-.5	.5	2	2	0.95
-1	-.05	.05	2	2	0.99
-1	1	5	1	1	0.8
-1	1	5	1	2	0.88
-1	1	5	1	5	0.92
-1	1	5	1	10	0.87
-0.8	5	9	1	1	0.68
-0.8	5	9	5	5	0.75
-0.8	5	9	7.5	7.5	0.97
-0.8	5	9	15	15	0.94

References.

- [1] Almeida, B. Gidas: A variational Method for Estimating the Parameters of MRF from Complete or Incomplete Data. (to appear in *Annals of Applied Probability* (1991).
- [2] A. Benveniste, M. Métivier, P.Priouret: *Algorithmes Adaptatifs et Approximations Stochastiques, Théorie et Application. Techniques Stochastiques.* Masson (1987).
- [3] J. Besag: Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. of Roy. Stat. Soc.* B-36 pp 192-236 (1974).
- [4] J.Besag: On the Statistical Analysis of Dirty Pictures. *J. of Roy. Stat. Soc.* B-48 n³, pp259-303 (1986) (with discussion).
- [5] B. Chalmond: Image Restoration Using an Estimated Markov Model. *Signal Processing* 15 pp. 115-129 (1987)
- [6] B. Chalmond: An iterative Gibbsian Technique for reconstruction of m-ary images. *Pattern Recognition*.Vol. 22 No. 6 pp. 747-761.(1989)
- [7] T.S. Chiang, Y. Chow and C.R. Hwang : Diffusion for Global Optimization in \mathbf{R}^n . *SIAM J. control and optimization*, 25 pp 737-753.
- [8] F. Comets and B. Gidas: Parameter estimation for Gibbs distributions, II: Partially Observed Data (to appear in *Annals of Applied Probability* (1991).
- [9] F. Comets: On the Consistency of a Class of Estimators for Exponential Families of MRF on the Lattice. preprint univ. Paris X.
- [10] R.L. Dobrushin: The Description of a Random Field by Means of Conditional Probabilities and Conditions of its Regularity. *Thry. Prob. Appl.* XIII-2 p. 197-224 (1968).
- [11] R.L. Dobrushin: Prescribing a System of Random Variables by Conditional Distributions. *Thry. Prob. Appl.* XV-3 p.458-486 (1970).
- [12] R.L. Dobrushin and S. B. Schlosman in *Statistical Mechanics and Dynamical Systems* Ed. Fritz, Jaffe, Szasz. Birkauser- Boston (1985).
- [13] A. Frigessi - M. Piccioni: Parameter Estimation for the two-dimensional Ising Fields Corrupted by Noise. *Stoch. Proc. and Appl.* 34 pp. 297-311. (1990)

- [14] S.B. Gelfand and S. Mitter: Recursive Stochastic Algorithm for Global Optimization in \mathbf{R}^d . preprint (1990).
- [15] D. and S. Geman: Stochastic Relaxation, Gibbs Distribution and Bayesian Restoration of Images. *IEEE TPAMI*. Vol PAMI-6 pp 721-741 (1984).
- [16] D. Geman: Parameter Estimation for MRF with Hidden Data and Experiments With the EM Algorithm. Preprint. (1984)
- [17] D. Geman: Random Fields and Inverse Problems in Imaging. Lecture notes at the Ecole D'Eté de Saint Flour.
- [18] D. Geman and D.E. Mc Clure: Statistical Methods for tomographic image reconstruction. In: *Proceedings of the 46-th session of the International Statistical Institute* Bulletin of the ISI, vol 52. (1987).
- [19] S. Geman and C. Graffigne : Markov Random Field Image Models and Their Applications to Computer Vision. In *Proceedings of the International Congress of Mathematicians 1986* Ed. A.M. Gleason. American Mathematical Society, Providence (1987).
- [20] S. Geman and C-R Hwang :Diffusion for Global Optimization. *SIAM J. control and optimization*, 24 pp 1031-1043.
- [21] B. Gidas: Consistency of Maximum Likelihood and Pseudo Maximum Likelihood for Gibbs Distributions. In *Stochastic Differential Systems With Appl. Elec. Comp. Eng., Op. Res.* Springer Verlag.
- [22] B. Gidas: Parameter estimation for Gibbs distributions, I: Fully Observed Data (to appear in *Markov Random Fields: Theory and Applications*, Academic Press (1991).
- [23] X.Guyon: Pseudo Maximum de Vraisemblance et Champs Markoviens. In *Spatial Processes and Spatial Time Series Analysis. Proc. 6th. Franco-Belgian Meeting of Statisticians*. Dreesbeke F. (Ed.). (1987)
- [24] C.R. Hwang and C.R. Sheu: On the behavior of a stochastic algorithm with Annealing. preprint (1990)
- [25] H.Künsch: Asymptotically Unbiased Inference for Ising Models. *J. of Appl. Prob.* 19 A 345-357 (1982).
- [26] H. Künsch et H. Dalhaus: Edge Effect and Efficient Parameter Estimation for Stationary Random Fields. Preprint 49. ETH Zentrum Zürich. (1986).
- [27] H.L. Kushner: Asymptotic global behavior for Stochastic Approximations and Diffusion with Slowly Decreasing Noise Effects: Global Minimization via Monte Carlo. *SIAM J. of Appl. Math* 47 pp. 169-185 (1987).
- [28] S. Lakshamanan and H. Derin: Simultaneous Parameter Estimation and Segmentation of Gibbs Random Fields. *IEEE trans. Pattern ana. and Machine Intell.* 11 pp 799-813.
- [29] O.E. Landford et D. Ruelle: Observable at Infinity and States with Short Range Correlations in Statistical Mechanics. *Comm. Math. Phys.* 13 p.194-215 (1969)
- [30] A.Lippman: A Maximum Entropy Method for Expert Systems. Brown University Thesis (1986).
- [31] J. L. Marroquin, S. Mitter and T. Poggio: Probabilistic solution of Ill Posed Problems in Computational Vision. *J. Am. Stat. Assoc.* 82 pp 76-89 (1987).
- [32] M. Métivier et P. Priouret: Théorèmes de Convergence p.s. pour une Classe d'Algorithmes Stochastiques à Pas Décroissants. *Prob. Th. Rel. Fields* 74, pp. 403-428. (1987).
- [33] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller et E. Teller: Equations of State Calculation for Fast Computing Machines. *J. Chem. Phys.* Vol 21. p.1087-1091. (1953).

- [34] Nguyen Xuan Xanh et H. Zessin: Ergodic Theorems for Spatial Processes. *Z. Wahr. Verw. Geb.* 48; p. 133-158 (1979).
- [35] D. Pickard: Asymptotic Inference for an Ising Lattice. *J. Appl. Prob.* 13, pp 486-497, 1976.
- [36] D. Pickard: Asymptotic Inference for an Ising Lattice II. *Adv Appl. Prob.* 9, pp 479-501 (1977).
- [37] D. Pickard: Asymptotic Inference for an Ising Lattice III. *J. Appl. Prob.* 16, pp 12-24 (1979).
- [38] D. Pickard: Inference for General Ising Models. *J. Appl. Prob.* 19 A, pp 345-357 (1982).
- [39] A. Possolo: Estimation of Binary Markov Random Fields. University of Washington, Technical Report (1986).
- [40] C. Preston: *Random Fields*. In Lect. Notes in Math. Vol. 534. Berlin, heidelberg, Newyork. Springer (1976).
- [41] B. Prum: *Processus sur un réseau et mesures de Gibbs; applications*. Techniques stochastiques. Masson. (1986).
- [42] D. Ruelle: *Thermodynamics Formalism*. In Encyclopedia of Mathematics and its applications. Vol. 5. Addison-Wesley. (1978).
- [43] B. Simon: A Remark on Dobrushin Uniqueness Theorem. *Comm. Math. Phys.* p.183. (1979).
- [44] A. Sokal: Monte Carlo Methods in Statistical Mechanics. Lecture notes: Cours de troisième cycle de la physique en Suisse Romande.
- [45] L. Younes: Couplage de l'estimation et du recuit pour des champs de Gibbs. *C. R. Acad. Sc. Paris, t. 303*, série I, No. 13 (1986).
- [46] L. Younes: Estimation and Annealing for Gibbsian Fields. *Ann. de l'Inst. Henri Poincaré* vol 2 (1988).
- [47] L. Younes: Parametric Inference For Imperfectly Observed Gibbsian Fields. *Prob. Thry. Rel. Fields* 82, pp 625-645. (1989)
- [48] L. Younes: Maximum of likelihood estimation for Gibbs Fields. To appear in *Proceeding of the 1988 AMS-IMS-SIAM joint conference on Spatial Statistics and Imaging* (1988).
- [49] L. Younes: Problèmes d'estimation paramétrique pour des champs de Gibbs Markoviens. Application au traitement d'images. Thesis at Université Paris Sud. (1988).