MAXIMUM LIKELIHOOD ESTIMATION FOR GIBBSIAN FIELDS

Laurent Younes
Université Paris Sud
Laboratoire de statistique appliquée
Bat 425, ORSAY
France

ABSTRACT

This paper presents practical remarks and numerical results on maximum likelihood estimation algorithms for perfectly and imperfectly observed Gibbsian fields on a finite lattice. These remarks are preceded by the definition of the algorithms. In the appendix, consistency of maximum likelihood estimation is proved.

Math. Reviews Classification: Primary:62F10 Secondary:60G80.

Keywords: Parametric Inference; maximum likelihood estimator; noisy data; stochastic gradient algorithm; random field; mixing properties; Gibbs sampler; consistency.

1. Introduction

Since Geman [10], Markov fields have become an efficient tool in image analysis. They have been used for image restoration, boundary detection, segmentation and appear to be a very natural way of modeling pictures. In this paper, we overview some problems related to maximum likelihood estimation for Gibbsian fields. We shall be mainly interested in giving practical remarks and numerical results about stochastic gradient algorithms, presented in Younes [32], Younes [33], that make possible the computation of maximum likelihood estimator (m.l.e.) for Markov fields.

Statistical analysis of Markov fields, using the equivalence with Gibbs field representation, has mainly been introduced in Besag [8]. A major contribution of this well-known paper, as regards parameter estimation, is the introduction of the coding techniques, which finally amounted to the definition of pseudo likelihood estimator (p.l.e.) for parameter inference. M.l.e. has not been used in this context because its computation has been considered an intractable problem, for reasons that will appear in the following presentation of modeling.

We shall consider a domain D, which will be a finite subset of $\mathbf{Z}^2$, typically a square. D is the set of sites, and a field on D is a random vector $X = (X_s, s \in D)$. For each s in D, X_s takes its values in a fixed finite set F. A Gibbs model is a family of laws on F^D,

$$\pi_\theta(x) = \exp(-\langle \theta, H(x) \rangle)/Z(\theta) \tag{1}$$

H is a function defined on $\Omega = F^D$, θ is a parameter which varies in a subset Θ of $\mathbf{R}^d$, and $Z(\theta)$ is a normalizing constant, the so-called partition function. The main problem in parameter estimation (and in other techniques involving Markov field modeling) is the fact that, because of the very large size of the set Ω, $Z(\theta)$ cannot be directly computed, neither analytically nor numerically. Indeed, this constant involves a sum of a number of terms that grows exponentially with the cardinality of D. This implies that expectations with respect to π_θ cannot be computed, although it is possible to estimate them using an iterative simulation algorithm (the Gibbs sampler); but even such an approximation can require a non negligible time.

Assume we are given a configuration x_0 in Ω as a sample of a field X under a law π_{θ_*} for an unknown θ_* in Θ. Maximum likelihood estimation in order to approximate θ_* consists in maximizing $\pi_\theta(x_0)$ in θ. The m.l.e. is a solution of the equation:

$$E_\theta(H) = H_0 \tag{2}$$

where H_0 is by definition $H(x_0)$, and E_θ is the expectation under π_θ. The preceding remarks about the difficulties that arise in the computation of expectations have led people to give up maximum likelihood estimation and look for more tractable procedures, in particular pseudo-likelihood estimation. This procedure, which first appeared (or at least its philosophy) in Besag [8], and has then been used and studied in various papers, (such as Guyon [9], Geman-Graffigne [22] ...), consists in choosing another criterion than the likelihood in order to discriminate distinct parameters. In general, the probability, under π_θ, of observing x_s at site s, conditional to the observation of x_t at site $t \neq s$, is an easily computed quantity. We shall denote it:

$$\pi_\theta^s(x_s | x_t, t \neq s) = P(X_s = x_s | X_t = x_t, t \neq s)$$

These conditional probabilities are called local specifications, and p.l.e. is the parameter θ at which

$$\prod_{s \in D} \pi_\theta^s(x_s | x_t, t \neq s)$$

is maximum, where $x = x_0$ is the observation.

Using the fact that it is easy to deal with local specifications, it is possible to define an iterative simulation procedure of Markov random fields: the Gibbs sampler (Geman [10]). With this procedure in hand, one can think of using a stochastic gradient algorithm to solve (2). This is done in Younes [31] and Younes [32]. Moreover, it is possible to make a generalization of this algorithm, to include the case of imperfectly observed data (Younes [33]). We shall here recall these procedures. It is not in our intent to give extensive proofs and details about them, as they already appear in the references; we shall rather focus on practical remarks and experimental results, which are new material. In addition, we shall give a result of consistency of maximum likelihood estimation, which, besides its own interest, appears to be important in a theorem in Younes [33] where it is stated without proof. We shall state this theorem and give more details later.

2. Parameter estimation

2.1 Fully observed data

We now give an algorithm that solves equation (2). We begin by an informal discussion to see that its definition is very natural.

We want to solve: $h(\theta) = E_\theta(H) - H_0 = 0$. In theory (we deal here with exponential models), this is a very simple problem, as $h(\theta)$ is the derivative of a concave function, and we have:

$$h'(\theta) = -var_\theta(H).$$

Let us assume that this matrix is definite for all θ.

Our issue is thus that it is impossible to compute exactly $h(\theta)$; if we could do so, a natural algorithm to solve (2) would be of the form:

$$\theta_{n+1} = \theta_n + a(E_{\theta_n}(H) - H_0), \tag{3}$$

and this algorithm will converge to the m.l.e., provided that a is a small enough positive constant. We shall now use the fact that there exists a method for simulating each π_θ. The Gibbs sampler has been introduced in Geman [10], and we now recall briefly its definition; it is an iterative algorithm, that provides a sequence $X^0, X^1, \ldots$ of configurations that converges in law to π_θ. To construct this sequence, we first give ourselves a sequence of sites, $(s_n, n \geq 0)$, which must scan the set D in an almost periodic way; more precisely, it must verify:

$$\exists R > 0 / \forall n, D \subset \{s_{n+1}, \ldots, s_{n+R}\}$$

Starting with any configuration X^0, one defines now X^{n+1} from X^n by taking $X_s^{n+1} = X_s^n$ for $s \neq s_n$, and taking $X_{s_n}^{n+1}$ at random according to the law $\pi_\theta^{s_n}(.|X_s^n, s \neq s_n)$.

In other terms, (X^n) is defined as a non-homogeneous Markov chain, and the transition kernel from X^n to X^{n+1} is

$$p^{n,n+1}(x,y) = \bigotimes_{s \neq s_n} \delta_{x_s}(y_s) \otimes \pi_\theta^{s_n}(y_{s_n}|y_s, s \neq s_n) \tag{4}$$

The Gibbs sampler has the following important ergodic property:

$$|P(X^n = x) - \pi_\theta(x)| = O(r^n)$$

with $0 \leq r < 1$. This implies in particular that one can estimate $E_\theta(H)$, with fixed θ by using the fact that:

$$E_\theta(H) = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} H(X^k)$$

where (X^k) is obtained by Gibbs sampling with parameter θ.

We can now replace $E_{\theta_n}(H)$ by an approximation in (3). But problems still subsist. We should know how to choose n_0 such that: $\sum_{k=1}^{n_0} H(X^k)/n_0$ is close enough to its limit. Such n_0 can be, in many cases, very large, so that a procedure involving such a computation would take too long. A second remark is that, in any case, we shall have to replace the expectation in (3) by a stochastic approximation, which will fluctuate; to take this into account, we must, (this is standard for stochastic algorithms), replace the constant a in (3) by a sequence that will tend (not too fast) to 0.

With this in mind we modify the initial algorithm, to reach a feasible procedure, as follows:

$$\theta_{n+1} = \theta_n + \sigma_{n+1}(\widehat{E}_n(H) - H_0) \tag{5}$$

With: $\sigma_n = \frac{1}{nU}$, where U is a fixed positive constant, and

$$\widehat{E}_n(H) = \sum_{k=1}^{n_0} H(X^{k,n+1})/n_0,$$

where $(X^{k,n+1})_{k=1,\ldots,n_0}$ is a sequence obtained by Gibbs sampling, with parameter θ_n, and taking as a starting point $X^{0,n+1} = X^{n_0,n}$.

If is easy to verify that, if the parameter converges, to $\overline{\theta}$, say, then the sequence $(X^{k,n})$ will follow, for large n, approximately the law $\pi_{\overline{\theta}}$, and $\overline{\theta}$ must be a solution of (2), and must then be equal to the m.l.e. $\widehat{\theta}$. Note that in Younes [32], we took $n_0 = 1$, and proved essentially the following result, that we state for any n_0:

Theorem 1. *There exists an $\alpha > 0$ such that, if $\frac{1}{n_0 U} < \alpha$, the sequence θ_n defined by (5) converges almost surely to $\widehat{\theta}$. (cf. Younes [32])*

Note that we do not need $\widehat{E}_n(H)$ to be close to $E_{\theta_n}(H)$ at each step, and moreover, one cannot expect $X^{k,n}$ to have law π_{θ_n} unless the algorithm is near convergence.

In Younes [32], we gave a lower bound for α in Theorem 1. This lower bound, which can be easily computed in most cases, appears to be too small in the applications, and we generally allow $1/(n_0 U)$ to be larger. But one must note that if this quantity is too large, the algorithm may diverge and simple counter-examples can be found (cf. Künsch, personal communication). However, in all cases, one may apply Métivier and Priouret's results on stochastic algorithms that once adapted to our context, say the following:

Theorem 2. *If $\mathcal{Q}$ is a fixed compact set in $\mathbf{R}^d$, that contains $\widehat{\theta}$ we have:*

conditionally on the event $\{\theta_p \in \mathcal{Q}\}$, the probability of the event: $\{\theta_n \in \mathcal{Q}$ for all $n \geq p$ and $\theta_n \to \widehat{\theta}\}$ is greater than $1 - \frac{C}{p}$, where C is a constant that may depend

on $\mathcal{Q}$, and on the constant U, but not on p. (cf. Benveniste-Métivier-Priouret [15], Métivier-Priouret [17])

This enables us to obtain, for any value of U, a convergent algorithm, by fixing an *a priori* compact set $\mathcal{Q}$, and forcing θ_n to come back into $\mathcal{Q}$ every time it goes out (by projection onto a smaller compact).

2.2 Imperfect observations

We now give an extension of (5) to imperfect observations. For this purpose, we must introduce some additional notations. X is still the same, with the same model (π_θ), but it no longer is observed. We only observe a function Y of X which will be constructed in the following manner: let b be a function from F onto another set G, and put $Y_s = b(X_s)$, for all $s \in D$. We shall denote this $Y = b_D(X)$.

Therefore, although there still is a realization x_0 of X, under a law π_{θ_*}, for an unknown θ_*, we can only make statistical inference from the realization $y_0 = b_D(x_0)$ of Y. Before going further, we remark that this framework contains the case which is of our main interest: noisy data. Indeed, let X_o be the original field, and N be "noise". We shall define X as the couple (X_o, N), and the laws π_θ will model the *joint distribution* of X_o and N. Now, the observation Y will be a function of X, for example $Y = X_o + N$ in the case of additive noise.

We call $\psi_\theta = \pi_\theta b_D^{-1}$ the likelihood of Y. It is easy to check that if θ maximizes $\psi_\theta(y_0)$, it must be a solution of:

$$E_\theta(H) - E_\theta(H|Y = y_0) = 0 \tag{6}$$

where the first expectation is under π_θ and the second is under $\pi_\theta^{y_0}$ which, by definition, is the law π_θ conditional on $Y = y_0$.

Note that (6) may have several solutions, as $\log \psi_\theta(y_0)$ needs not to be a concave function of θ. The algorithm we propose is a stochastic gradient algorithm, which, like any gradient-descent algorithm may converge to a local maximum of the likelihood. In fact, the result we are going to state, following Younes [33] is a local convergence result. We now proceed to the definition of the algorithm we shall use; this definition is mainly based on the following remark:

call $\widetilde{X}$ the field (X_1, X_2), with independent X_1 and X_2, X_1 of law π_θ and X_2 of law $\pi_\theta^{y_0}$; $\widetilde{X}$ is a Markov field, of which it is easy to compute the energy function, and which has the same neighborhood system as π_θ. Let us call $\widetilde{E}_\theta$ the expectation with respect to the law of $\widetilde{X}$, and note $f(x_1, x_2) = H(x_1) - H(x_2)$. Equation (6) can now be rewritten as:

$$\widetilde{E}(f) = 0 \tag{7}$$

This allows us to try the same kind of procedure as before, simply replacing X by $\widetilde{X}$; θ_{n+1} will be obtained from θ_n by

$$\theta_{n+1} = \theta_n + \sigma_n \widehat{E}_n(f) \tag{8}$$

where $\widehat{E}_n(f)$ is an "approximation" of $\widetilde{E}_{\theta_n}(f)$, obtained, as in (5), by averaging the values of f after n_0 steps of a Gibbs sampler simulating the law of $\widetilde{X}$. Note that such

Gibbs sampler is equivalent to two independent simulation algorithms, one for X_1 (i.e. for the law π_θ) and one for X_2 (i.e. for the law $\pi_\theta^{y_0}$).

We shall now state asymptotic results in terms of the size of the set of sites D. For this, we need to define, in a proper way, a law π_θ for all D, and this is possible after introducing the notion of potential. Description of this notion, and definition of laws on D, are made in the appendix. For the sake of brevity, we shall only state here the result without recalling all definitions that are needed, as its general sense will still be apparent. (Refer to the appendix for more details)

In Younes [33], we proved that, if the true parameter was θ_*, then, for homogeneous potential (cf. appendix) and under Dobrushin's mixing conditions (Dobrushin [2]), the function $\log \psi_\theta(y_0)$ was concave in a neighborhood of θ_* for large enough D, and we gave the expression of the second derivative of the likelihood. This, combined with the fact that the m.l.e. is consistent (cf. appendix), and after applying results in Métivier-Priouret [17] gives the following theorem:

Theorem 3. *Assume X has homogeneous potential, that under θ_* satisfies Dobrushin's mixing conditions. Assume moreover that the m.l.e. is consistent. Then, there exists an open subset of $\mathbf{R}^d$, $\mathcal{D}$, with $\theta_* \in \mathcal{D}$, such that,*

if D (the set of sites) is large enough, then the m.l.e. $\widehat{\theta}$ is in $\mathcal{D}$, and for all compact $\mathcal{Q}$ included in $\mathcal{D}$ and containing $\widehat{\theta}$, one has: conditionally on the event $\{\theta_p \in \mathcal{Q}\}$, the probability of the event: $\{\theta_n \in \mathcal{Q}$ for all $n \geq p$ and $\theta_n \to \widehat{\theta}\}$ is greater than $1 - \frac{C}{p}$.

We make now some practical remarks and give numerical results.

3. Experiments

3.1 Generalities

Implementation of algorithms (5) and (8) requires some care. As it could have been expected, the former is far more stable than the latter, since in the first case, the function to maximize is concave. But even in this case, there are good choices of several constants that must be done. For example, the best U in Theorem 1, is to be found. We shall often use matrices to replace σ_n, and also define a stopping procedure; this will also require fitting some constants or thresholds. These numbers must in general be found by experimenting with the algorithm on a given model, as they will highly depend on the situation in which inference is made. For (5), there can be a significant difference of computation time between good and bad choices of these constants, and for (8), a bad choice can simply prevent the algorithm from converging.

Another important point is the need of having a good initialization for the algorithm. This is important first as regards reduction of computation, and also for imperfect observations, to prevent non-convergence, or convergence to a local maximum. Unfortunately, in this last case there exist very few methods for parameter estimation, even rough parameter estimation. They are in general very dependant on the application (Chalmond [20]), or on the model (Frigessi [21]). In most cases, one must rely on *a priori* knowledge and heuristic considerations on the model to have an idea of what kind of parameter is to be expected. Of course, when the data are perfectly observed, the pseudo-likelihood estimate should be chosen as starting point.

3.2 Choice of gains

The algorithms we have defined so far were of the kind:

$$\theta_{n+1} = \theta_n + \sigma_{n+1}\widehat{E}_n(f)$$

where the gain σ_n was real. In general, one obtains a significant improvement by using matrices. These gains can be taken from what is done for exact gradient algorithms: in order to maximize a function l with known derivatives, "optimal" gains are of the kind:

$$\theta_{n+1} - \theta_n = [a.||l'(\theta_n)||.I - l''(\theta_n))]^{-1}l'(\theta_n)$$

In our case, the function l is the logarithm of the likelihood, for which derivatives are not explicitly known. In the imperfect observation case, for example, we have

$$l'(\theta) = E_\theta(H) - E_\theta(H|Y = y_0),$$

and

$$-l''(\theta) = var_\theta(H) - var_\theta(H|Y = y_0).$$

In the perfectly observed case, both conditional expectation and variance degenerate into constants.

For fixed θ, one can approximate these derivatives by simulation. Of course, it is out of question to compute them accurately at each step of the algorithm, as one of these estimations might be almost as costly as the initial algorithm itself. However, one can use the sequence(s) (X^n) that are simulated during the procedure, to have a rough approximation of these differentials. In the fully observed case, for example, $l''(\theta_n)$ will be estimated by the sample variance of the $H(X^{k,p})$, by averaging over a certain past of time n (typically a few scans of D). These approximations are of course anything but accurate, as they may even be based on random variables that do not follow the law under the parameter θ_n. But, provided one takes good care of problems related to matrix inversion, use of these approximations can provide a fair increase in the speed of convergence of the algorithm. Note also that these variances are more accurately estimated when we approach convergence. This will be important for the stopping criterion in the next section.

To conclude this section we give the precise form of the algorithm we actually use:

$$\theta_{n+1} = \theta_n + \sigma_n.[a||d_n||.I + \Gamma_n]^{-1}\widetilde{E}_n(f); \tag{9}$$

d_n is a vector that "approximates" the derivative of the log likelihood, Γ_n is a matrix that "approximates" the opposite of the second derivative; a is a real number that must be chosen, and (σ_n) is a sequence of real numbers that tends to 0.

It is often useful, in particular when the initial value of the algorithm is far for the m.l.e., to let σ_n be constant during the first few iterations (once again, this constant will have to be properly chosen). Then, one will let σ_n tend to 0; in our applications, we took $\sigma_n = 1/(b + cn)$, b et c depending on the situation.

We now focus on the choice of an automatic stopping rule for the algorithm.

3.3 Stopping Rule

The maximum likelihood estimator, both for perfect and imperfect observations, is a solution of an equation of the form:

$$\widetilde{E}_\theta(f) = 0. \tag{10}$$

Our stopping rule consists of stopping the algorithm as soon as this equation is satisfied. Once again, computational constraints do not allow us to check at each step if the current parameter θ_n is a solution of (10). We could do here as we have done before: compute an approximation of $\widetilde{E}(f)$ by averaging on some past values of f computed in the simulations done during the algorithm; but, by doing this, we would have had to face some complicated problems. The main one is that, for our purpose, it would be important to have rather good approximations of the expectation, or at least, to have a measure of the level of accuracy that can be expected; even in the simple case of Gibbs sampling with constant parameter, it is still an open problem to define a precise criterion to check how near the estimated expectation is from the true value. Another drawback of this method is that it doesn't use the fact that, in general, domains on which fields are observed are large, and laws may have some ergodic properties that imply that $f/|D|$ is close to $\widetilde{E}(f)/|D|$, where $|D|$ is the number of sites. Under some additional hypothesis (typically mixing), one can even say that $f/\sqrt{|D|}$ follows a Gaussian law.

When the algorithm is close to convergence, one can consider that the simulated configuration $X^{k,p}$ follows approximately the law under the limit parameter $\widehat{\theta}$. Hence, if the preceding hypotheses are true, $f(X^{k,p})/\sqrt{|D|}$ is approximately centered and Gaussian. As the variance of f can be estimated during the procedure, and is rather accurate when we are close to convergence, it is possible to make a χ^2-test to check if (10) is true. In fact, we used this stopping rule even when the hypothesis that we made during this discussion were not satisfied, and it appeared to still be giving rather good results. Since approximations of the variances can be very bad when the algorithm has not converged, it is better, each time the test gives a positive answer, to make some additional iterations to check if this answer was really true. All this is summarized next.

3.4 Summary

Here is the procedure we propose for using the estimation algorithm.

1. Start with a preliminary estimation, θ_0, of the parameter, as good as possible for a small cost.
2. Let the algorithm run with constant σ. Compute during this time sample variances that are involved in the second derivative of the likelihood, as well as in the χ^2-test of the stopping rule (in the case of perfectly observed data there is in fact only one variance to compute). Stop, when the χ^2-statistics is lower than a constant C_1, provided it hasn't overpassed another constant C_1' during a few control iterations.
3. Same as the preceding one, with decreasing gain and new constants C_2 and C_2' for the stopping rule. The constants C_1, C_1', C_2, C_2' must be empirically chosen.

3.5 Simulations: Fully observed data

We chose, for our simulations, an Ising model, taking binary values (0 or 1), with external field. The energy is:

$$\alpha \sum x_{ij} + \beta_1 \sum x_{ij} x_{i-1j} + \beta_2 \sum x_{ij} x_{ij-1} \tag{11}$$

For several values of $(\alpha, \beta_1, \beta_2)$, we simulated 100 realizations of this model on a 64×64 lattice. For each realization, we computed p.l.e. and m.l.e., and estimated bias and normalized mean square errors on the basis of the 100 results. We also computed the Fisher information matrix, which is, for exponential models, the inverse of the covariance matrix of the normalized sufficient statistics. This covariance matrix was also computed on the basis of the 100 simulations.

1. True parameters: 1,-1,-1.

   Bias:

ple:	mle:
−0.046	−0.076
−0.086	0.019
0.031	0.039

   Mean square errors:

   ple:
   $$\begin{pmatrix} 725 & -195 & -235 \\ -195 & 130 & -10 \\ -235 & -10 & 150 \end{pmatrix}$$
   mle:
   $$\begin{pmatrix} 685 & -175 & -225 \\ -175 & 120 & -15 \\ -225 & -15 & 150 \end{pmatrix}$$
   Fisher:
   $$\begin{pmatrix} 335 & -95 & -95 \\ -95 & 75 & -20 \\ -95 & -20 & 70 \end{pmatrix}$$

2. True parameters: 0,-0.8, -0.8.

   Bias:

ple:	mle:
0.014	0.007
−0.039	−0.032
0.021	0.019

   Mean square errors:

   ple:
   $$\begin{pmatrix} 1770 & -465 & -510 \\ -465 & 265 & 0 \\ -510 & 0 & 280 \end{pmatrix}$$
   mle:
   $$\begin{pmatrix} 1720 & -465 & -480 \\ -465 & 255 & 5 \\ -480 & 5 & 260 \end{pmatrix}$$
   Fisher:
   $$\begin{pmatrix} 1090 & -330 & -265 \\ -330 & 195 & -10 \\ -265 & -10 & 160 \end{pmatrix}$$

3. True parameters: .15, 2, 2.

   Bias:

ple:	mle:
0.003	−0.009
0.018	−0.042
0.032	−0.036

   Mean square errors:

   ple:
   $$\begin{pmatrix} 15 & -15 & -5 \\ -15 & 135 & 0 \\ -5 & 0 & 110 \end{pmatrix}$$
   mle:
   $$\begin{pmatrix} 15 & 10 & 0 \\ 10 & 85 & -35 \\ 0 & -35 & 65 \end{pmatrix}$$
   Fisher:
   $$\begin{pmatrix} 15 & -10 & -20 \\ -10 & 90 & 0 \\ -20 & 0 & 110 \end{pmatrix}$$

4. Trues parameters: 1, 2, -2.

   Bias:

ple:	mle:
−0.015	−0.021
0.002	0.001
0.004	0.012

   Mean square errors:

   ple:
   $$\begin{pmatrix} 125 & -70 & -70 \\ -70 & 80 & 35 \\ -70 & 35 & 50 \end{pmatrix}$$
   mle:
   $$\begin{pmatrix} 95 & -50 & -55 \\ -50 & 55 & 30 \\ -55 & 30 & 35 \end{pmatrix}$$
   Fisher:
   $$\begin{pmatrix} 75 & -40 & -45 \\ -40 & 50 & 25 \\ -45 & 25 & 30 \end{pmatrix}$$

5. True parameters: 0.15, 2,-2.

   Bias:

ple:	mle:
0.003	0.003
0.001	0.000
−0.001	0.002

   Mean square errors:

   ple:
   $$\begin{pmatrix} 335 & -165 & -160 \\ -165 & 100 & 75 \\ -160 & 75 & 90 \end{pmatrix}$$
   mle:
   $$\begin{pmatrix} 275 & -130 & -130 \\ -130 & 80 & 60 \\ -130 & 60 & 75 \end{pmatrix}$$
   Fisher:
   $$\begin{pmatrix} 255 & -115 & -120 \\ -115 & 65 & 55 \\ -120 & 55 & 65 \end{pmatrix}$$

The average time for estimation by maximum of likelihood is approximately 5 minutes on a Vax 750, the maximum of pseudo likelihood only takes a few seconds. In general, m.l.e. is more efficient than p.l.e., and always is at least as efficient. Both as very bad for "ferromagnetic fields" ($\beta_i < 0$), and more accurate for "anti-ferromagnetic" fields. Difference in efficiency between the two estimators becomes larger as interactions become stronger ($|\beta_i|$ increases).

Finally, we point out that the preceding matrices were only empiric estimations. Hence, it may happen that the theoretical Cramer-Rao inequality is not always satisfied between mean square error matrices and Fisher information.

3.6 Simulation: Noisy Data

We now add noise to the preceding observations. In fact, we used three kinds of noise:

- Forgotten data: an Ising model provides binary values (0 or 1); 1 can be, for example, interpreted as a presence of a certain particle on a site; we assumed that some of these particles have not been seen by the observer so that the true field has been perturbed by replacing by 0 some values with a probability p, which is assumed to be known. If we call X_o the original field, and N is the noise, X_o follows an Ising law, N is field of i.i.d. Bernoulli r.v. of parameter $1 - p$, which is

independent of X_o, and the function b in the section 2.2 is

$$b(x, n) = xn$$

- Flipped data: now, we assume that some of the values have been flipped, with probability p. X and N are as before, except that the parameter of the Bernoulli laws is p; b is now:

$$b(x, s) = x(1 - n) + n(1 - x).$$

- Gaussian additive noise: We add to the binary field Gaussian white noise, so that the observed data are real. The variance of the noise is assumed to be known. b is

$$b(x, n) = x + n$$

We now give some estimated mean square errors for some values of the parameter. We still give three matrices, *but they must not be compared in the same way as in the preceding section.* P.l.e. is computed on the noisy data, as if there were no noise. It thus gives very biased results; the only interest of this computation is that it shows that the perturbation of the data was significant. We used this very bad estimation as starting point of the procedure of section 3.4. The Fisher information matrices are computed on the original data, before adding of noise. They are thus not to be considered as an efficiency criterion, but they give information on the loss of accuracy of estimations because of the noise. Here are the results:

- Forgotten data:

(a) True parameters: 0.5, 1, 1. $p = 0.3$

Bias:

ple:	mle:
0.74	0.02
−0.12	−0.004
−0.13	−0.014

Mean square errors:

ple:

$$\begin{pmatrix} 2260 & -400 & -415 \\ -200 & 155 & 75 \\ -415 & 75 & 150 \end{pmatrix}$$

mle:

$$\begin{pmatrix} 36 & -30 & -30 \\ -30 & 125 & 0 \\ -30 & 0 & 100 \end{pmatrix}$$

Fisher:

$$\begin{pmatrix} 20 & -15 & -15 \\ -15 & 50 & -5 \\ -15 & -5 & 55 \end{pmatrix}$$

(b) True parameters: 0.5, -0.5, 0.5. $p = 0.3$

Bias:

ple:	ple:
0.5	0.008
0.16	−0.002
−0.11	0.002

Mean square errors:

ple:
$$\begin{pmatrix} 1040 & 320 & -240 \\ 320 & 125 & -75 \\ -240 & -75 & 75 \end{pmatrix}$$
mle:
$$\begin{pmatrix} 60 & -35 & -35 \\ -35 & 40 & 5 \\ -35 & 5 & 45 \end{pmatrix}$$
Fisher:
$$\begin{pmatrix} 35 & -20 & -15 \\ -20 & 20 & 5 \\ -15 & 5 & 20 \end{pmatrix}$$

(c) True parameters: 0.5,-0.5, 0.5. $p = 0.4$

Bias:

ple:	mle:
0.71	0.07
0.19	−0.04
−0.15	−0.05

Mean square errors:

ple:
$$\begin{pmatrix} 2085 & 550 & -460 \\ 550 & 175 & -125 \\ -460 & -125 & 130 \end{pmatrix}$$
mle:
$$\begin{pmatrix} 110 & -75 & -60 \\ -75 & 85 & 10 \\ -60 & 10 & 75 \end{pmatrix}$$
Fisher:
$$\begin{pmatrix} 30 & -15 & -20 \\ -15 & 20 & 0 \\ -20 & 0 & 25 \end{pmatrix}$$

(d) True parameters: 0.5, -1, 1. $p = 0.3$

Bias:

ple:	ple:
0.48	0.03
−0.25	−0.02
0.35	0.01

Mean square errors:

ple:
$$\begin{pmatrix} 970 & -520 & 665 \\ -520 & 305 & -365 \\ 665 & -365 & 505 \end{pmatrix}$$
mle:
$$\begin{pmatrix} 75 & -45 & -45 \\ -45 & 50 & 15 \\ -45 & 15 & 40 \end{pmatrix}$$
Fisher:
$$\begin{pmatrix} 35 & -15 & -25 \\ -15 & 15 & 5 \\ -25 & 5 & 25 \end{pmatrix}$$

- Flipped Data.

(a) True parameters: 0.5, 1, 1. $p = 0.1$

Bias:

ple:	mle:
−0.003	−0.03
−0.58	−0.07
−0.58	−0.05

Mean square errors:

ple:
$$\begin{pmatrix} 20 & -5 & -5 \\ -5 & 1390 & 1370 \\ -5 & 1370 & 1405 \end{pmatrix}$$
mle:
$$\begin{pmatrix} 45 & -50 & -40 \\ -50 & 230 & -30 \\ -40 & -30 & 265 \end{pmatrix}$$
Fisher:
$$\begin{pmatrix} 20 & -15 & -15 \\ -15 & 55 & 0 \\ -15 & 0 & 50 \end{pmatrix}$$

(b) True parameters: 0.5, 0.5, -0.5. $p = 0.1$

Bias:

ple:	mle:
−0.11	0.02
−0.19	−0.01
0.20	−0.02

Mean square errors:

ple:
$$\begin{pmatrix} 80 & -105 & 75 \\ -105 & 170 & -155 \\ 75 & -155 & 175 \end{pmatrix}$$

mle:
$$\begin{pmatrix} 70 & -45 & -35 \\ -45 & 50 & 5 \\ -35 & 5 & 45 \end{pmatrix}$$

Fisher:
$$\begin{pmatrix} 25 & -15 & -15 \\ -15 & 15 & 0 \\ -15 & 0 & 15 \end{pmatrix}$$

(c) True parameters: 0.5, 1.5, -1.5. $p = 0.1$

Bias:

ple:	mle:
−0.20	0.01
−0.61	−0.01
0.59	−0.02

Mean square errors:

ple:
$$\begin{pmatrix} 195 & 455 & -485 \\ 455 & 1535 & -1460 \\ -485 & -1460 & 1420 \end{pmatrix}$$

mle:
$$\begin{pmatrix} 105 & -50 & -55 \\ -50 & 40 & 20 \\ -55 & 20 & 35 \end{pmatrix}$$

Fisher:
$$\begin{pmatrix} 40 & -20 & -20 \\ -20 & 20 & 10 \\ -20 & 10 & 15 \end{pmatrix}$$

- Gaussian noise.

(a) True parameters: 0.5, 1, 1. Variance of the noise: 0.25

Bias:

ple:	mle:
−0.04	0.04
−0.73	−0.03
−0.73	−0.04

Mean square errors:

ple:
$$\begin{pmatrix} 30 & 105 & 105 \\ 105 & 2230 & 2200 \\ 105 & 2200 & 2220 \end{pmatrix}$$

mle:
$$\begin{pmatrix} 70 & -45 & -70 \\ -45 & 240 & -65 \\ -70 & -65 & 300 \end{pmatrix}$$

Fisher:
$$\begin{pmatrix} 15 & -10 & -15 \\ -10 & 50 & -5 \\ -15 & -5 & 60 \end{pmatrix}$$

(b) True paramaters: 0.5, -1, 1. Variance of the noise: 0.25

Bias:

ple:	mle:
−0.20	0.01
0.54	0.006
−0.57	−0.02

Mean square errors:

ple:
$$\begin{pmatrix} 175 & -425 & 420 \\ -425 & 1195 & -1255 \\ 420 & -1255 & 1355 \end{pmatrix}$$

mle:
$$\begin{pmatrix} 90 & -55 & -45 \\ -55 & 60 & 10 \\ -45 & 10 & 55 \end{pmatrix}$$

Fisher:
$$\begin{pmatrix} 30 & -20 & -15 \\ -20 & 20 & 5 \\ -15 & 5 & 15 \end{pmatrix}$$

(c) True parameters: 0.15, 1.5, -1.5. Variance of the noise: 0.25

Bias:

ple:	mle:
−0.09	−0.007
−0.80	0.006
0.80	−0.004

Covariance Matrices:

ple:
$$\begin{pmatrix} 75 & 290 & -325 \\ 290 & 2655 & -2655 \\ -325 & -2655 & 2695 \end{pmatrix}$$

mle:
$$\begin{pmatrix} 100 & -45 & -50 \\ -45 & 30 & 20 \\ -50 & 20 & 40 \end{pmatrix}$$

Fisher:
$$\begin{pmatrix} 65 & -25 & -30 \\ -25 & 15 & 10 \\ -30 & 10 & 20 \end{pmatrix}$$

Appendix: Consistency of M.L.E.

A.1 Introduction

We now study consistency of m.l.e. for imperfectly observed data. For this, we shall need some definitions and notation, in particular suitably to define laws on any finite subset D of $\mathbf{Z}^2$.

A potential is a family (on which we shall put a parameter) of functions $(\lambda_C(\theta,.))_C$, where C runs over all finite subsets of $\mathbf{Z}^2$, $\lambda_C(\theta,.)$ is defined on $F^{\mathbf{Z}^2}$, takes its values in $\mathbf{R}$, and only depends on coordinates indexed by elements of C.

We shall make the following hypotheses on this family:

- Homogeneity: let us call T_s the shift operator on $F^{\mathbf{Z}^2}$ defined by: $(T_s x)_t = x_{s+t}$ for x in $F^{\mathbf{Z}^2}$ and $t \in \mathbf{Z}^2$. We assume that, for all C, θ, s and x:

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

- Uniform bounded range: we assume that there exists a $\gamma \geq 0$ such that, for all θ, $\lambda_C(\theta,.) = 0$ if $\mathrm{diam}(C) > \gamma$. The diameter of C is taken with respect to the distance on $\mathbf{Z}^2$:

$$d((i,j),(i',j')) = \max(|i-i'|, |j-j'|)$$

- Regularity: we assume that λ_C is continuously differentiable in θ for all θ.

Once we are given a family of potentials, we define for each $D \subset \mathbf{Z}^2$ a family of conditional laws in the following manner. Let $x \in F^D$, and $x' \in F^{D^c}$. Call $z = x.x'$ the element of $F^{\mathbf{Z}^2}$, such that $z_s = x_s$ for $s \in D$ and $z_s = x'_s$ for $s \in D^c$. Define on F^D the law $\pi_\theta(.|x')$ by:

$$\pi_\theta(x|x') = e^{-\Lambda_D^{x'}(\theta,x)}/Z^{x'}(\theta) \tag{12}$$

where

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

x' is called the boundary condition. We can also define a free boundary condition law on F^D by replacing the sum in $\Lambda_D^{x'}$ by a sum over those C that are included in D, that is, use the energy:

$$\Lambda_D(\theta,x) = \sum_{C\subset D} \lambda_C(\theta, x.x'). \tag{13}$$

The preceding formula does not depend on x', which can be arbitrary, because λ_C only depends on coordinates indexed by C.

In general, assume that we have, for all $D \subset \mathbf{Z}^2$, a family $\pi_{\theta,D}$ of laws on D, defined by (12), replacing $\Lambda_D^{x'}$ by another energy $\overline{\Lambda}_D$. We say that we have a suitable family of approximate laws (s.f.a.l.) if:

$$\lim_{D\to\mathbf{Z}^2} \|\Lambda_D(\theta,.) - \overline{\Lambda}_D(\theta,.)\|_\infty/|D| = 0 \tag{14}$$

uniformly on compact sets in θ. The preceding limit, like all limits over subsets of $\mathbf{Z}^2$ in this paper, is to be understood as being true for any sequence (D_n) of squares in $\mathbf{Z}^2$ that contains any finite set for large enough n. Because of the hypotheses we made, it is easy to check that, if we arbitrarily fix a boundary condition for each D, then the resulting family is a s.f.a.l.

The family of laws given by (12) form a consistent system of conditional laws. And in fact, there exists a law π_θ on $F^{\mathbf{Z}^2}$ such that $\pi_\theta(.|x')$ is (as indicated by the notation) the conditional law of π_θ on D when the outside of D is known, and equal to x'. In fact, there may even exist several laws on $F^{\mathbf{Z}^2}$ which satisfy to this property.

Our conditions on the potential imply the following results:

Result 1. *For each θ, there exists at least one (spatially) homogeneous law on $F^{\mathbf{Z}^2}$ associated to the potential $(\lambda_C(\theta,.))$. We call the family of these laws: $\mathcal{G}_0(\theta)$. $\mathcal{G}_0(\theta)$ is convex, compact. Its extremal points are ergodic measures on $\mathbf{Z}^2$, and any two of them are singular. Each element of $\mathcal{G}_0(\theta)$ is a convex combination of extremal points.*

These facts can be found in Ruelle [7]. Existence of a field has first been proved in Dobrushin [2]. Note that there can exist non-homogeneous laws associated with a given potential. The set $\mathcal{G}(\theta)$ of all these laws, which contains $\mathcal{G}_0(\theta)$ has the same properties as $\mathcal{G}_0$. But, as it will appear in the hypothesis that we shall give later on, we shall only be concerned with homogeneous laws. There exist sufficient conditions that ensure uniqueness of a law associated to the potential, ie. $|\mathcal{G}| = |\mathcal{G}_0| = 1$. These conditions given in Dobrushin [2], and simplified in Simon [8], are used in theorem 3 of section 2.2. In our context, they are:

Let θ be fixed. There exists an $a \in [0,1]$ such that:

$$\sum_{C,0\in C} (|C|-1)\|\Lambda_C(\theta,.)\| \leq \alpha$$

We shall not assume these conditions to prove consistency; in fact our framework is the following : give ourselves a family of potentials satisfying the hypotheses given at the beginning of the section. The true law of X is an element of $\mathcal{G}_0(\theta_*)$, for an unknown θ_*. Our purpose is to estimate θ_* on the basis of imperfect observations of X on a finite square D.

Description of the way imperfect observations are obtained is given in section 2.2; we recall that we gave ourselves a function b from F to G such that, for all s, the observation at site s is $Y_s = b(X_s)$. We shall call $\mathbf{b}$ the application from $F^{\mathbf{Z}^2}$ to $G^{\mathbf{Z}^2}$, with all components equal to b, so that $Y = \mathbf{b}(X)$. Similarly, the function b_D defined on F^D is treated in the same way.

If $z = (z_s, s \in \mathbf{Z}^2)$ is a configuration on the whole plane, we denote $z_D = (z_s, s \in D)$ its restriction to a subset D of $\mathbf{Z}^2$. For a law ϕ_D on H^D, where H can be F or G, and $z \in H^{\mathbf{Z}^2}$, we shall note: $\phi_D(z)$ instead of $\phi_D(z_D)$ to simplify formulas.

Now, we describe how parameter identification is done: we fix a s.f.a.l. for X, choosing, for each D, an energy that can be boundary free, but may be anything else, provided that (14) is satisfied. $\pi_{\theta,D}(x)$ denotes the approximate law of X on D (it is not a marginal of a law in $\mathcal{G}_0(\theta)$), and $\psi_{\theta,D}(y)$ denotes the image of this law under b_D. We assume that there exists a configuration x_0 of X, which is a realization of a law of unknown parameter. The observable configuration is $y_0 = \mathbf{b}(x_0)$. If D is a finite subset of $\mathbf{Z}^2$, the maximum likelihood estimator on D is any parameter that maximizes $\psi_{\theta,D}(y_0)$ on a given set $\Theta \subset \mathbf{R}^2$, which we assume to be compact.

Whenever possible, we leave our subscripts such as D or θ.

In general, when we talk about true marginal laws, we use an subscript a (suggesting "absolute"), and write π_a or ψ_a. They cannot be put into form (12), and thus cannot be used for parameter estimation. This is why we introduced the notion of s.f.a.l. Note that, in the first part of the paper, we only used exponential models, which are particular cases of the ones we use here.

We need to make a last, but crucial, assumption before stating and proving consistency: identifiability of the parameter on the basis of Y.

To a parameter value θ, we can associate the family $\mathcal{G}_0(\theta)$, which is the set in which the law of X may vary, if the true parameter is θ. Now, the law of Y must be an element of the set $\mathcal{H}_0(\theta) = \mathbf{b}(\mathcal{G}_0(\theta))$.

Identifiability then means that for θ, $\theta' \in \Theta$,

$$\theta \neq \theta' \Longrightarrow \mathcal{H}_0(\theta) \cap \mathcal{H}_0(\theta') = \emptyset \tag{15}$$

We impose that two different parameters give two different laws.

$\mathcal{F}_D$ denotes the σ-algebra induced in $F^{\mathbf{Z}^2}$ by the coordinates included in D ($\mathcal{F}_D = \sigma(X_s, s \in D)$). $\overline{\mathcal{F}}_D$ denotes the σ-algebra on $G^{\mathbf{Z}^2}$, induced by Y_D, which is the image of $\mathcal{F}_D$ under b_D.

Elements de $\mathcal{H}_0(\theta)$ are characterized by the property: $Q \in \mathcal{H}_0(\theta)$ if and only if there exists $P \in \mathcal{G}_0(\theta)$ such that

$$\forall D \subset \mathbf{Z}^2, \text{ finite}, \forall A \in \overline{\mathcal{F}}_D, Q(A) = \int \pi_\theta(b_D^{-1}(A)|x_{D^c})P(dx_{D^c}). \tag{16}$$

This only says that Q is the image of a law P associated to the potential indexed by θ. The true parameter being θ_*, the law of X is an element P_* of $\mathcal{G}_0(\theta_*)$; its image under $\mathbf{b}$ will be denoted Q_*; E_*^X and E_*^Y denote expectations with respect to P_* and Q_*.

A.2 Statement of the theorem

Theorem 4. *If Θ is compact, $\theta_* \in \Theta$, the identifiability condition (15) is true, and that the law of X is homogeneous, then*

the maximum likelihood estimator on Θ is consistent.

This theorem is a consequence of the following proposition, which will be proven next:

Proposition 1.

$$\lim_{D \to \mathbf{Z}^2} \frac{1}{|D|} \log \frac{\psi_{\theta,D}(y)}{\psi_{\theta,D}(y)} = -h(\theta, \theta_*) \tag{17}$$

exists Q_* almost surely, and $h(\theta, \theta_*)$ is positive, continuous in θ, and vanishes only for $\theta = \theta_*$.

In addition, if Θ is compact and convex in $\mathbf{R}^d$, then there exists a constant K depending on θ such that, for all θ, θ' in Θ,

$$\frac{1}{|D|}|\log \psi_\theta(y_D) - \log \psi_{\theta'}(y_D)| \leq K||\theta - \theta'||$$

To prove Proposition 1, we shall proceed as follows. We first check that the asymptotic behavior of (17) does not depend on the s.f.a.l. that has been chosen. We then prove the equivalent of Theorem 7.1 in Preston [5], adapted to imperfect observation context. We shall then be able to prove Proposition 1.

A.3 Proof

A.3.1 First step

We first show that the quantity in interest has asymptotic behavior that is independent on the choice of the s.f.a.l. The parameter θ is fixed, and we omit it in the formulas. Let (ψ_D) and $(\widetilde{\psi}_D)$ be two s.f.a.l. We need to show that

$$\max_y \log \frac{\psi_D(y)}{\widetilde{\psi}_D(y)} = o(|D|). \tag{18}$$

We denote Λ_D the energy associated with ψ_D, $\widetilde{\Lambda}_D$ the one associated with $\widetilde{\psi}_D$, and define

$$\epsilon_E = \frac{1}{|D|}||\Lambda_D - \widetilde{\Lambda}_D||_\infty.$$

ϵ_D tends to 0 by assumption. In addition, we have:

$$\psi_D(y) = \sum_{b_D(x)=y_D} e^{-\Lambda_D(x)} / \sum_x e^{-\Lambda_D(x)}$$

And the same formula for $\widetilde{\psi}$, replacing Λ by $\widetilde{\Lambda}$. But

$$\log\left[(\sum e^{-\Lambda_D(x)})/(\sum e^{-\widetilde{\Lambda}_D(x)})\right]$$

lies between $-|D|\epsilon_D$ and $|D|\epsilon_D$, as soon as the sum is made over the same x's in the numerator as in the denominator. This implies that:

$$\left|\frac{1}{|D|}\log\frac{\psi_D(y)}{\widetilde{\psi}_D(y)}\right| \leq 2\epsilon_D$$

and thus tends to 0 uniformly in y if D tends to $\mathbf{Z}^2$.

Note that one can show in the same way:

$$\max_{y,x,x'} \log\frac{\psi_D^x(y)}{\psi_D^{x'}(y)} = o(|D|). \tag{19}$$

where ψ^x (resp. $\psi^{x'}$) is the law with boundary condition x (resp. x'). This implies that $\log\psi_D(y) - \log\psi_{a,D}(y)$ is $o(|D|)$, where $\psi_{a,D}$ is the marginal of some law in $\mathcal{G}(\theta)$. Indeed, we have:

$$\psi_{a,D}(y) = \mathbf{E}[\psi_D^{x'}(y)]$$

(the expectation is with respect to the chosen absolute law) hence:

$$\inf_{x'} \psi^{x'}(y) \leq \psi_a(y) \leq \sup_{x'} \psi^{x'}(y)$$

A.3.2 Second step

Lemma 1. *For a given θ let P_θ be any element of $\mathcal{G}_0(\theta)$, and Q_θ its image by* **b**. *Let π_{a_D} and ψ_{a_D} be the marginals of these laws on finite domains D; let $\pi_{*,D}$ and $\psi_{*,D}$ be the marginals of P_* and Q_* on D.*

If

$$\liminf \frac{1}{|D|} E_*^Y[\log\frac{\psi_{a_D}(y)}{\psi_{*,D}(y)}] = 0,$$

then $\theta = \theta_*$.

Note the behavior of the preceding ratio does not depend on the choice of P_θ in $\mathcal{G}_0(\theta)$.

Theorem 7.1 in Preston [5] states this result for the case of perfectly observed data ($Y = X$). The proof can be adapted to our context. Consequently, we shall only detail those points in the proof that have to be modified, referring to Preston [5] for complete information.

If we put $g_D(y) = \frac{\psi_{a_D}(y)}{\psi_{*,D}(y)}$, and

$$H_D = H_D(\theta, \theta_*) = E_*^Y[\log \frac{\psi_{a_D}(y)}{\psi_{*,D}(y)}],$$

we can check that $H_D = E_\theta^Y[f(g_D(y))]$ where $f(t) = t\log(t) + 1 - t$. Because f is convex and positive, H_D is increasing with D, and is positive.

If we put, for finite D and $\overline{D}$ $q_{D,\overline{D}}(y) = \frac{g_{D\cup\overline{D}}(y)}{g_{\overline{D}}(y)}$, we have (Preston [5]; Lemma 7.1), for A $\overline{\mathcal{F}}_D$-measurable:

$$Q_*(A) = \int_A q_{D,\overline{D}}(y) g_{\overline{D}}(y) Q_\theta(dy)$$

Moreover, if D if finite and A is $\overline{\mathcal{F}}_D$-measurable, then for large enough, finite $\overline{D} \subset D^c$:

$$\int_A g_{\overline{D}}(y) Q_\theta(dy) = \int \pi_\theta(b_D^{-1}(A)|x_{D^c}) P_*(dx_{D^c})$$

This equality corresponds to lemma 7.2 in Preston [5]; its proof is based on the identity:

$$E_\theta^Y(\mathbf{1}_A g_{\overline{D}}) = E_*^X\left[E_\theta^X(\mathbf{1}_A \circ \mathbf{b}|\mathcal{F}_{\overline{D}})\right],$$

and on the fact that, as we have assumed bounded range:

$$E_\theta^X(\mathbf{1}_A \circ \mathbf{b}|\mathcal{F}_{\overline{D}}) = E_\theta^X(\mathbf{1}_A \circ \mathbf{b}|\mathcal{F}_{D^c})$$

as soon as $\overline{D}$ is large enough.

We get from this that

$$Q_*(A) - \int \pi_D(b_D^{-1}(A)|x_{D^c}) P_*(dx_{D^c}) = \int_A (q_{D\overline{D}} - 1) g_{\overline{D}} dQ_\theta$$

as soon as $\overline{D}$ is large enough.

To prove Lemma 1, it suffices to show that this quantity vanishes for all A, as (16) would imply that $Q_* \in \mathcal{H}_0(\theta)$ and thus $\theta = \theta_*$ because of condition (15).

To obtain this, the rest of the proof in Preston [5] can be applied without modification (lemmas 7.3 to 7.6).

A.3.3 Third step

We now proceed to the main part of the proof. We shall study the limit of

$$\frac{1}{|D|} E_*^Y(\log(\psi_{\theta,D}(y))$$

for a fixed θ, and a given s.f.a.l. Because of the first step of the proof, this limit does not depend on the choice of the s.f.a.l.; therefore we shall choose, for each D, ψ_D as being associated with free boundary energy, which is:

$$\Lambda_D(\theta, .) = \sum_{C \subset D} \lambda_C(\theta, .).$$

Since θ is fixed, we leave it out of the formulas.

One can write

$$\psi_D(y) = \frac{Z_D(y)}{Z_D}$$

with $Z_D(y) = \sum_{b_D(x_D)=y_D} e^{-\Lambda_D(x)}$, and $Z_D = \sum_{x_D} e^{-\Lambda_D(x)}$.

It is well known, from the theory in the case of complete observations, that

$$\log(Z_D)/|D|$$

converges when D tends to $\mathbf{Z}^2$; we thus only have to study the limit of

$$E_*^Y(\log(Z_D(y))/|D|.$$

Note also that convergence of the former is a consequence from convergence of the latter.

Fix a square $D_0 = [0, k_0]^2$. Let $\mathcal{R}$ denote the sub-lattice of $\mathbf{Z}^2$ generated by the vertices of D_0. ($\mathcal{R}$ is the set of pairs of integers of the kind $(p.k_0, q.k_0)$.)

For a square D included in $\mathbf{Z}^2$, we call $\mathcal{R}_D$ the set of elements s of $\mathcal{R}$ for which $s + D_0 \subset D$.

We first study the ratio:

$$r_D = \frac{Z_D(y)}{\prod_{s\in\mathcal{R}_D} Z_{D_0+s}(y)}$$

For this, note $D_1 = \bigcup_{s\in\mathcal{R}_D}(s + D_0)$, $D_2 = D\backslash D_1$, and

$$\overline{\Lambda}_D(x) = \sum_{s\in\mathcal{R}_D} \Lambda_{s+D_0}(x),$$

r_D now can be written:

$$r_D = \frac{\sum_{x_D} e^{-\Lambda_D(x)}}{\sum_{x_{D_1}} e^{-\overline{\Lambda}_D(x)}}$$

All sums being taken over x such that $b(x) = y$.

We have:

$$\overline{\Lambda}_D(x) = \sum_{s\in\mathcal{R}_D} \sum_{C\subset s+D_0} \lambda_C(x)$$

Let $\mathcal{C}_D$ be the set of all C that are included in D, but not in any of the $s + D_0$, and of diameter lower than γ. (Recall that γ is the range of the family of potential; λ_C vanishes if C has a diameter larger than γ.)

One can write:

$$\Lambda_D(x) = \sum_{s\in\mathcal{R}_D} \sum_{C\subset s+D_0} \lambda_C(x) + \sum_{C\in\mathcal{C}_D} \lambda_C(x) \tag{20}$$

As the λ_C are uniformly bounded (by a constant K, say) we have $|\Lambda_D - \overline{\Lambda}_D| \leq K|\mathcal{C}_D|$.

$\mathcal{C}_D$ only contains sets that intersect some $D_0 + s$ without being included in one of them, and sets included in D_2. The former are fewer than (constant $.|\mathcal{R}_D|\sqrt{|D_0|}$). Indeed any element of $\mathcal{C}_D$ that meets an $s + D_0$ must be a subset of a square of diameter γ centered at an element t of $s + D_0$, but not included in $s + D_0$. If s is fixed, the number of such squares is lower than (constant $.\sqrt{|D|}$), the constant only depending on γ (it can be taken as 2γ); the number of subsets of such squares can now be bounded in the same way, for fixed s; taking all possible s gives the preceding estimate. If one notes that the cardinal of $\mathcal{R}_D$ is less than $|D|/|D_0|$, this estimate can be seen to be of the same order as $|D|/\sqrt{|D_0|}$.

If we now count the sets in $\mathcal{C}_D$ that are included in D_2, we see that there can be no more than (constant $.|D_2|$), the constant depending once again on γ. We can remark now that D_2 only contains sites that are at a distance lower than an edge of D_0 from the outside of D. With this in mind, it is easy to convince oneself that $|D_2|$ is smaller that: (constant $.\sqrt{|D|.|D_0|}$), and thus smaller than (constant $.|D|/\sqrt{|D_0|}$).

Therefore, we have shown that:

$$|\Lambda_D - \overline{\Lambda}_D| \leq M_0|D|/\sqrt{|D_0|}$$

M_0 being a constant that depends on K and on γ. Coming back to r_D, we get:

$$r_D = \frac{\sum_{x_{D_1}} e^{-\overline{\Lambda}_D(x)} \sum_{x_{D_2}} e^{-\Lambda_D(x)+\overline{\Lambda}_D(x)}}{\sum_{x_{D_1}} e^{-\overline{\Lambda}_D(x)}}$$

hence:

$$|F|^{|D_2|}e^{-M_0|D|/\sqrt{|D_0|}} \leq r_D \leq |F|^{|D_2|}e^{M_0|D|/\sqrt{|D_0|}}$$

Using once again the estimate on the cardinality of D_2, we obtain the fact that there exists a constant M such that:

$$|\frac{1}{|D|}\log Z_D(y) - \frac{1}{|D|}\sum_{s\in\mathcal{R}_0}\log Z_{s+D_0}(y)| \leq \frac{M}{\sqrt{|D_0|}} \tag{21}$$

If we note that $Z_{s+D}(y) = Z_D \circ T_s(y)$, and that P_* was assumed to be homogeneous, we deduce from (21) that

$$|\frac{1}{|D|}E_*^Y[\log Z_D(y)]| \leq \frac{|\mathcal{R}_D||D_0|}{|D|}|\frac{1}{|D_0|}E_*^Y[\log Z_{D_0}(y)] + \frac{M}{\sqrt{|D_0|}}$$

Calling $a_D = E_*^Y \frac{1}{|D|}\log Z_D(y)$, we get:

$$\limsup a_D \leq a_{D_0} + \frac{M}{\sqrt{|D_0|}}$$

and thus $\limsup a_D \leq \liminf a_D$ and a_D converges.

If we now assume that P_* is ergodic, we get from (21):

$$\limsup \frac{1}{|D|}\log Z_D(y) \leq a_{D_0} + \frac{M}{\sqrt{|D_0|}}$$

and thus

$$\limsup \frac{1}{|D|} \log Z_D(y) \leq \lim a_D$$

In the same manner, we get

$$\liminf \frac{1}{|D|} \log Z_D(y) \geq \lim a_D$$

which implies $\frac{1}{|D|} \log Z_D(y)$ converge P_* p.s.

Finally, if we do not assume that P_* is ergodic, we know that P_* is a convex combination of the extremal points of $\mathcal{G}_0$, which are ergodic and mutually singular. This means that there exists a probability measure ω on the set $\mathcal{E}$ of extremal points such that:

$$P_* = \int_{\mathcal{E}} P.\omega(dP).$$

Let us call S_P the support of $P \in \mathcal{E}$. We know that, on each S_P, $\frac{1}{|D|} \log Z_D(y)$ converges (y is a function of x); but the support of P_* is equal to the union of the S_P, up to a set of null P_*-measure. Hence, here again $\frac{1}{|D|} \log Z_D(y)$ converges P_* a.s. Its limits under P_* is in the set of its limits under the various extremal ergodic measures.

A.3.4 Conclusion

The preceding result shows that $\lim \frac{1}{|D|} \log \frac{\psi_{*,D}(y)}{\psi_{\theta,D}(y)}$ exists, P_* almost surely ($y = \mathbf{b}(x)$). In addition, the foregoing discussion shows that the limit is one of the limits for one of the extremal point P of $\mathcal{G}_0(\theta_*)$. This $P = P(x)$ depends on x, but not on θ. The limit, under $P(x)$, of $\frac{1}{|D|} \log \frac{\psi_{x,D}(y)}{\psi_{\theta,D}(y)}$ is the same as the limit of the expectation of the same quantity under $P(x)$, where $y = \mathbf{b}(x)$. Lemma 1, taking $P(x)$ instead of P_*, now shows that this limit can be 0 only if $\theta = \theta_*$.

For a fixed x, we denote this limit $h(\theta, \theta_*)$. The second part of the proposition, as well as the continuity of h can be proved by noting that, for all D, $\frac{1}{|D|} \log(\psi_\theta(y))$ is continuously differentiable in θ, and that its derivative can be estimated uniformly in y and D and uniformly on compact sets in Θ. We get from this a Lipschitz inequality for $\frac{1}{|D|} \log(\psi_\theta(y))$ which is uniform in D and in θ on compact sets, and thus get the same inequality for h by letting D tend to $\mathbf{Z}^2$.

This ends the proof of proposition 1, and the proof of the consistency of m.l.e.

References

- Statistics.

  [1] Dacunha Castelle, D., & Duflo, M. (1983). *Probabilité et Statistiques* (2 Tomes), Masson.

- Random Fields.

  [2] Dobrushin, R. L. (1968). The Description of a Random Field by Means of Conditional Probabilities and Conditions of its Regularity. *Thry. Prob. Appl.* **XIII-2**, 197-224.

[3] Kunsch, H. (1982). Decay of Correlations under Dobrushin's Uniqueness Condition and its Applications. *Comm. Math. Phys.* **84**, 207-222.

[4] Xanh, Nguyen Xuan & Zessin, H. (1979). Ergodic Theorems for Spatial Processes. *Z. Wahr. Verw. Geb.* **48**, 133-158.

[5] Preston, C. (1976). *Random Fields.* In Lect. Notes in Math. **534**. Berlin, heidelber, Newyork, Springer.

[6] Prum, B. (1986). *Processus sur un réseau et mesures de Gibbs; applications.* Techniques stochastiques, Masson.

[7] Ruelle, D. (1978). *Thermodynamics Formalism.* Encyclopedia of Mathematics and its applications **5**, Addisson-Wesley.

[8] Simon, B. (1979). A Remark on Dobrushin Uniqueness Theorem. *Comm. Math. Phys.* **183**.

- Non stationary Markov Chains and Gibbs Sampler.

[9] Dobrushin, R. L. (1956). Central Limit Theorem for Non-stationary Markov Chains. *Thry. Prob. Appl.* **I**, 329-383.

[10] Geman, D., & Geman, S. (1984). Stochastic Relaxation, Gibbs Distribution and Bayesian Restoration of Images. *IEEE TPAMI.* **PAMI-6**, 721-741.

[11] Gidas, B. (1985). Nonstationary Markov Chains and Convergence of the Annealing Algorithm. *J. Stat. Phys.* **39**, 73-131.

[12] Isaacson, D., & Madsen, R. (1976). *Markov Chains: Theory and Applications.* New York, Wiley.

[13] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of State Calculation for Fast Computing Machines. *J. Chem. Phys.* **21**, 1087-1091.

[14] Mitra, D., Romeo, F., & Sangiovanni-Vincentelli, A.L. (1985). Convergence and Finite Time Behaviour of Simulated Annealing. *Proc. 24th Conf. on Decision and Control. Ft. Lauderdale*, 761-767.

- Stochastic algorithms

[15] Benveniste, A., Métivier, M., Priouret, P. (1987). *Algorithmes Adaptatifs et Approximations Stochastiques. Théorie et Application.* Techniques Stochastiques, Masson.

[16] Lippman, A. (1986). *A Maximum Entropy Method for Expert Systems.* Brown University Thesis.

[17] Métivier, M., & Priouret, P. (1984). Théorèmes de Convergence p.s. pour une Classe d'Algorithmes Stochastiques a Pas Décroissants. *Probab. Th. Rep. Fields* **74**, 403–428.

- Estimation For Gibbs Fields.

[18] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. of Roy. Stat. Soc.* **B-36**, 192-236.

[19] Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *J. of Roy. Stat. Soc.* **B-48**, 259-303, (with discussion).

[20] Chalmond, B. (1988). Image Restoration Using an Estimated Markov Model. *Signal Processing* **15**.

[21] Frigessi, A., & Piccioni, M. (1988). Parameter Estimation for 2-D Ising Fields corrupted by noise, (preprint).

[22] Geman, S., & Graffigne, C. (1987). Markov Random Field Image Models and their Applications to Computer Vision. Gleason. A. M. (1987, ed.), *Proceeding of the internatonal Congress of Mathematicians*, American Mathematical Society.

[23] Guyon, X. (1987). Pseudo Maximum de Vraisemblance et Champs Markoviens. Spatial Processes and Spatial Time Series Analysis. Droesbeke, F. (1987, ed.), *Proc. 6th. Franco-Belgian Meeting of Statisticians.*

[24] Kunsch, H. (1982). Asymptotically Unbiased Inference for Ising Models. *J. of Appl. Prob.* **19** (A), 345-357.

[25] Kunsch, H. ( ). Discussion on Besag's paper [19].

[26] Pickard, D. (1976). Asymptotic Inference for an Ising Lattice. *J. Appl. Prob.* **13**, 486-497.

[27] Pickard, D. (1977). Asymptotic Inference for an Ising Lattice II. *Adv Appl. Prob.* **9**, 479-501.

[28] Pickard, D. (1979). Asymptotic Inference for an Ising Lattice III. *J. Appl. Prob.* **16**, 12-24.

[29] Pickard, D. (1982). Inference for General Ising Models. *J. Appl. Prob.* **19** (A), 345-357.

[30] Possolo, A. (1986). *Estimation of Binary Markov Random Fields.* University of Washington, Technical Report.

[31] Younes, L. (1986). Couplage de l'estimation et du recuit pour des champs de Gibbs. *C. R. Acad. Sc. Paris, t.* **303**, sèrie I, n° 13.

[32] Younes, L. (1988). Estimation and Annealing for Gibbsian Fields. *Ann. de l'Inst. Henri Poincaré* **2**, 269-294.

[33] Younes, L. ( ). Parameter Estimation for Imperfectly observed Gibbsian fields. *Probab. Th. Ref. Fields* **82**, 625–645.

[34] Younes, L. (1988). *Problémes d'Estimation Paramétrique pour des Champs de Gibbs Markoviens; Application au Traitement d'Images.* Thesis. Université Paris-Sud.