

Stochastic gradient estimation strategies for Markov random fields

Laurent Younes^a

^a CMLA, ENS-Cachan, 61 av. du Président Wilson, 94235 Cachan CEDEX

ABSTRACT

This communication presents new results about convergence of stochastic gradient algorithms for maximum likelihood estimation of Markov random fields. We first present theoretical results dealing with the convergence of a generalized Robbins-Monro procedure. These results provide rigorous justifications for simple numerical strategies which can be employed in practice; they are illustrated by numerical experiments.

1. INTRODUCTION

Markov random field models have become a standard, efficient tool, for low-level Bayesian analysis of images. In many cases, a good estimation of the prior distribution brings significant improvement to subsequent analyses based on the posterior. We describe here some strategies which evaluate with accuracy the true parameters of a statistical model of random fields, on the basis of an observed training data.

The statistical procedure which will be used is maximum likelihood estimation. For Markov random fields, this is a hard numerical problem which can be solved efficiently by stochastic gradient algorithms, and the aim of this paper is to consider a class of stochastic gradient algorithms which can be proved to be consistent (theorem 4.1), and try to evaluate their performance *for a given amount of computational effort*. Such stochastic algorithms indeed exhibit a natural trade-off, because at each step, the parameter is updated on the basis of an estimated gradient. The issue is then to decide whether one is ready to dedicate a lot of time to the refinement of the estimation of the gradient at each step, or rather use a crude estimation, at the risk of having to make more steps. We do not give a definite answer to this dilemma but we provide computer simulations which might indicate a few hints on the way this trade-off can be resolved.

We therefore consider a statistical model of Markov random fields. We restrict to the exponential family of models, and set

$$\pi_{\theta}(x) = \exp\{-\langle \theta, U(x) \rangle - \log Z_{\theta}\} \quad (1)$$

where θ is a d -dimensional parameter, x belongs to a (finite) configuration space Ω , U is a sufficient statistics defined on Ω taking values in \mathbb{R}^d , $\langle \theta, U(\cdot) \rangle$ is the usual scalar product and Z_{θ} is adjusted so that $\sum_{x \in \Omega} \pi_{\theta}(x) = 1$.

It is well-known that, for such models, the log-likelihood $l_{\theta}(x) = \log \pi_{\theta}(x)$ is a concave function of θ , so that its maximization is, in theory, a well specified problem; the maximum is either a unique local maximum, or belongs to a direction of constancy of the log-likelihood, or does not exist because it is attained at a direction of recession of the log-likelihood (cf. ref. 1). The last two cases being in general quite exceptional ($U(x)$ must belong to a face of the convex hull of the collection $\{U(y), y \in \Omega\}$), we restrict here to the case of a unique local maximum, which also is the unique solution of

$$\frac{d}{d\theta} l_{\theta}(x) = 0$$

which writes

$$E_{\theta}(U) - U(x) = 0 \quad (2)$$

the expectation E_{θ} being taken with respect to π_{θ} .

The numerical issue of finding the maximum of $l_{\theta}(x)$, or equivalently of solving (2) is much less favorable. The expectation $E_{\theta}(U)$ is not computable, and can only be approximated by Monte-Carlo sampling; as a consequence, the standard gradient descent procedure

$$\theta_{n+1} = \theta_n + \gamma_{n+1}(E_{\theta_n}(U) - U(x)) \quad (3)$$

E-mail: younes@cmla.ens-cachan.fr; internet: <http://www.cmla.ens-cachan.fr/~younes>

which converges, for example, if γ_n is a small enough constant positive gain, cannot be used, since $E_{\theta_n}(U)$ is not computable. However, since π_{θ_n} can be sampled from, this expectation can be approached by, say, k_{n+1} Monte-Carlo averages, ie. one may write

$$E_{\theta_n}(U) \simeq \frac{1}{k_{n+1}} \sum_{p=1}^{k_{n+1}} U(Y^{n+1,p})$$

where $Y^{n+1,p}, p = 1, \dots, k_{n+1}$ is a sequence of random variables which sample the distribution π_{θ_n} (we shall be more specific in the next section). Equation (3) can be replaced by

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \left[\frac{1}{k_{n+1}} \sum_{p=1}^{k_{n+1}} [U(Y^{n+1,p}) - U(x)] \right]. \quad (4)$$

This kind of approximation, with large k_{n+1} (and an approximation of the Hessian), has been proposed in several papers, for example ref. 2, 3. However, provided that the sampling is carefully organized, as described below, small values of k_{n+1} , and thus very rough estimations of the gradient can be used. The first proof of almost-sure convergence with $k_n = 1$ for all n , and small enough decreasing gains is provided in ref. 4. We here specialize theoretical results of ref. 5 to the case of Markov random fields; these results exhibit the possibility to *tune* the value of k_{n+1} depending on the value of the current parameter θ_n .

We are thus placed into the context of stochastic gradient algorithms, in the framework which has originally been described in ref. 6. However, the specificity of the problem which is addressed here is that the sampling which can be employed for Markov random fields uses dynamic algorithms, which generate a *Markov chain* which only converges after a long (theoretically infinite) run to the required distribution. This slightly complicates the form of the stochastic algorithms, and makes the theoretical analysis significantly harder. In particular, almost-sure convergence is only true with much more restrictive assumptions. Some basic tools for the theoretical analysis of such algorithms can be found in ref. 7.

2. SAMPLING FROM MARKOV RANDOM FIELDS

It is not in our intent to give here a precise description of the sampling algorithms which can be used for Markov random fields. The interested reader can refer to one of the many references which address this subject, for example ref. 8, 9, 10, 11 or 12. Here, we want to only provide a general form for such algorithms, essentially to settle some notation.

As said before, these sampling procedures generate Markov chains, and as such are completely specified by providing their transition probabilities, ie. a function P defined on $\Omega \times \Omega$ such that, for all $x \in \Omega$, $y \mapsto P(x, y)$ is a probability on Ω . The associated Markov chain with initial state x_0 is the stochastic process $(X^n, n \geq 0)$ in Ω such that $X^0 = x_0$ and the conditional distribution of X^{n+1} given all the past X^0, \dots, X^n is given by $P(X^n, \cdot)$. This Markov chain is said to *sample a distribution* π if, for all $x \in \Omega$ the probability that $X^n = x$ tends to $\pi(x)$ when n tends to infinity, that is:

$$\lim_{n \rightarrow \infty} \mathbf{P}(X^n = x) \rightarrow \pi(x).$$

We shall only consider situations when this convergence is at exponential speed, that is, situations when there exists $\lambda \in [0, 1[$ such that, for some constant K , for all n and for all $x \in \Omega$,

$$|\mathbf{P}(X^n = x) - \pi(x)| \leq K \lambda^n.$$

λ is the *rate of convergence* of the Markov chain.

So, returning to our statistical model (π_θ) , we shall assume that we are given such a sampling procedure, with transition probability P_θ , which samples from π_θ at rate λ_θ . It is typical that this rate of convergence becomes less and less efficient for large values of the parameter. In most cases, we have a bound of the kind

$$0 \leq \lambda_\theta \leq 1 - C e^{-D|\theta|} \quad (5)$$

for some constant C and $D > 0$. This inequality will be assumed hereafter.

3. DESCRIPTION OF THE STOCHASTIC GRADIENT ALGORITHM

We fix a configuration $x \in \Omega$ and assume that the function $\pi_\theta(x), \theta \in \mathbb{R}^d$ admits a unique maximum, which will be denoted $\hat{\theta}$ (the maximum likelihood estimator based on the observation x).

The stochastic gradient algorithm updates the parameter according to the rule

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \left[\frac{1}{k_{n+1}} \sum_{p=1}^{k_{n+1}} [U(Y^{n+1,p}) - U(x)] \right]. \quad (6)$$

in which the variables $Y^{n+1,p}, p = 1, \dots, k_{n+1}$ are generated by a Markov chain with transition P_{θ_n} and *initial state* $Y^{n+1,0} = Y^{n,k_n}$. Thus, for all $p \in \{1, \dots, k\}$,

$$\mathbf{P}(Y^{n+1,p} = y | Y^{n+1,q}, q = 0, \dots, p-1) = P_{\theta_n}(Y^{n+1,p-1}, y)$$

and $Y^{n+1,0} = Y^{n,k_n}$.

This last constraint is *essential* to ensure almost-sure convergence. Because of the Markovian character of the sampling, restarting the Markov chain afresh at each update of θ_n will induce a *systematic bias* in the evaluation of the gradient, and eventually bias the limit parameter.

We have two degrees of freedom in the design of this algorithm: the choice of the gain γ_{n+1} , and of the number of iterations k_{n+1} at each step. These quantities cannot be chosen arbitrarily, however, and we now give some sufficient conditions which yield sequences (θ_n) which converge almost-surely to $\hat{\theta}$

4. THEORETICAL RESULTS

Assume that $\gamma_n = an^{-b}$ with $1/2 < n \leq 1$, and l_n be the integer part of n^{1/c_1} with $c_1 < b/2$.

We assume that the sequence k_p is such that, for some constants a' and $a'' > 0$,

- i) For all p , k_{p+1} is larger than the integer part of $1 + a' \exp(2D|\theta_p|)$
- ii) If there exists no n such that $p = l_n$, then

$$\frac{1}{\sqrt{k_p}} - \frac{1}{\sqrt{k_{p-1}}} \leq a'' p^{-c_2}$$

with $c_2 > 1 - b/2$.

Then,

THEOREM 4.1. *With the previous assumptions, the sequence θ_n converges almost surely to $\hat{\theta}$*

Conditions i) and ii) for strategy 2 means that k_p is not allowed to decrease too much, unless $p = l_n$ for some n , and that k_p must be large if the current parameter is large. There is no assumptions on the sizes of the constants a, a' and a'' . In particular a' can be very small, so that k_{p+1} will be constant unless θ_p is very large; this is an adaptive behaviour, in which the algorithm reacts only when there is a hint that the trajectory $\theta_n, n \geq 0$ might enter into an exploding regime. This solution is preferable to the usual solution for stochastic gradient algorithms which simply imposes hard boundedness constraints to prevent excursions of the parameter outside of a given compact set which has been selected *a priori*.

Note that exploding trajectories (ie. trajectories for which $|\theta_n| \rightarrow \infty$) can occur with positive probability if one is not careful in designing the gains. Examples are provided in ref. 5.

5. PRACTICAL IMPLEMENTATION

5.1. Iterating the algorithm

A central limit theorem is proved in ref. 5, which states that, when the sequence θ_n comes close enough to its limit value, it starts to oscillate around it with a variance which scales as $\sqrt{\gamma_n}$. This suggests the following strategy, which iterates the previous one.

For this, let a finite sequence $1/2 < b_1 < \dots < b_m \leq 1$, and run the following algorithm:

- 1) Set $q = 1$
- 2) Run the algorithm of theorem 4.1 with $b = b_q$ and compute empirically $u_n = n^{b/2} \text{var}(\theta_n)$ during the run
- 3) Detect if u_n stabilizes; if $q = m$, stop, otherwise set $q := q + 1$

5.2. Choice of k_n

We recommend to let k_n be fixed unless $\|\theta_n\|$ is very large. The exact constant D in equation (5) is generally unknown, and the upper bounds which can generally be computed are much too large to be used in practice. However, a few trials in running the algorithm generally provide enough information to obtain good empirical values for D and other constants.

5.3. Using the Hessian

Deterministic gradient ascent algorithms can significantly be improved by using the Hessian (seconde derivative) of the maximized function (Newton algorithm). Here, this would mean replacing equation (3) by

$$\theta_{n+1} = \theta_n + \gamma_{n+1}(\lambda I - H_{\theta_n})^{-1}(E_{\theta_n}(U) - U(x))$$

for some $\lambda > 0$, I being the $d \times d$ identity matrix, and H_θ begin the Hessian of $l_\theta(x)$,

$$H_\theta = \frac{d^2}{d\theta^2} l_\theta(s).$$

For exponential families of models, this second derivative is the covariance matrix of U for the distribution π_θ :

$$H_\theta = -\mathbf{Var}_\theta(U),$$

which can be approached by Monte-Carlo sampling. For stochastic approximation, the following algorithm may be used (we write it for $k_n \equiv 1$ to simplify)

$$\begin{cases} M_{n+1} = M_n + \sigma_{n+1}\{U(Y^{n+1,1}) - M_n\} \\ V_{n+1} = V_n + \rho_{n+1}\{[U(Y^{n+1,1}) - M_n] \cdot [U(Y^{n+1,1}) - M_n] - V_n\} \\ \theta_{n+1} = \theta_n + \gamma_{n+1}(\lambda I + V_n)^{-1}\{U(Y^{n+1,1}) - U(x)\} \end{cases} \quad (7)$$

When θ_n stabilizes, one can see that M_n converges to $E_{\theta_n}(U)$, and that V_n is close to $\mathbf{Var}_{\theta_n}(U)$, and the updating of θ_n will be optimal. Thus, at least in a neighbourhood of the limit, one can expect that the performances of the algorithm will be increased (in fact, one can prove it, cf. ref.⁷). However, if one is far away from convergence, and θ_n varies abruptly, significant errors can be made in the evaluation of the Hessian and nothing guaranties that using this modified algorithm would be better: since U varies a lot, the variance generally might be overestimated, which reduces the next variations of θ_n , which might delay convergence. More over, an ill-estimated variance has no reason to yield directions of descent which would be better than with the original gradient algorithm.

6. EXPERIMENTS

6.1. Introduction

In this section, we consider an Ising model with external field on a square lattice $\Lambda \subset \mathbb{Z}^2$. This model writes, for $\theta = (\alpha, \beta) \in \mathbb{R}^2$ and $x = (x_s, s \in \Lambda), x_s = \pm 1$,

$$\pi_\theta(x) = \exp[-\alpha U_1(x) - \beta U_2(x) - \log Z_\theta]$$

where $U_1(x) = \sum_{s \in \Lambda} x_s$ and

$$U_2(x) = \frac{1}{2} \sum_{s, t \in \Lambda, |s-t|=1} x_s x_t$$

where $|s - t| = 1$ means that s and t are nearest neighbors in Λ .

We have experimented algorithm (6) with a fixed values of $k_n \equiv k_0$ (no explosion occurred with this data-set). For given values of α and β , and k_0 , we have used the following procedure:

- 1) sample, once for all, a configuration x from the distribution π_θ , with $\theta = (\alpha, \beta)$.
- 2) estimate the maximum likelihood estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ associated to the observation x with a long run of stochastic gradient algorithms
- 3) repeat K times the following operations:
 - 3.1) generate an initial parameter $\theta_0 = (\alpha_0, \beta_0)$, where $\alpha_0 = \hat{\alpha} + \Delta_1, \beta_0 = \hat{\beta} + \Delta_2$, where Δ_1, Δ_2 are random variables sampled from a uniform distribution on a small interval $[-\delta, \delta]$.
 - 3.2) run algorithm (6) with $k_n \equiv k_0$ for a fixed number N/k_0 of steps
- 4) compute (by averaging over the K previous experiments) the standard errors of the current parameter θ_p , for different fixed values of $p \leq N$.

For such algorithms, we define the computation time at step n to be nk_0 , that is, we essentially let the unit time correspond to the sampling of one $Y^{p,k}$. All the algorithms are therefore ran for a fixed computation time N .

6.2. Analysis of k_0

In the following experiments, we took $K = 200, N = 7500$, and ran the computations for $k_0 = 1, 5, 10, 15, 20$ and 25. We also ran experiments with different values of δ and β (we always let $\alpha = 0$). For each experiment, we have drawn two kinds of plots:

- The variance of α_n (left plot) and β_n (right plot) in function of the computation time; trajectories for various values of k_0 are plotted on the same frame.
- The variance of α_n (left) and β_n (right) at time N in function of k_0 .

Some general trends can be observed. First, it is always harder to estimate α than β (this has already been observed by several authors for this Ising model when $\alpha = 0$). The theory predicts that using large k_0 is always better for long runs, but that there should be an optimal k_0 for a limited number of iterations. This is more or less confirmed by our experiments, which we detail now.

In figures 1 and 2, we have $\beta = -0.5$ and $\delta = 0.05$, which corresponds to a small initial variance. For the estimation of α , the runs are quite similar. The initial steps yield a degradation from the original variance (indicating that γ_n should have been smaller), then the error decreases, and small k_0 remain better than large k_0 until about $t = 5000$. The trajectories then seem to cross in favor of larger k_0 , although this inversion seems not to be completely achieved at $t = 7500$. For β , after the degradation at the beginning, $k_0 = 1$ remains better until the end of our runs ($t = 7500$), but an extrapolation of the curves indicates that the inversion should take place some time later.

With the same $\beta = -0.5$ and $\delta = 0.1$, the results are more intricate. While $k_0 = 1$ or 5 is always better for short runs, the different trajectories become so close that it seems hard to draw any conclusion ; this is the same for $\delta = 0.25$ (figures 3 to 6).

For $\alpha = 0$, $\beta = -0.7$, we still note that the estimation of α becomes even harder. In all cases, larger k_0 are preferable from the start, and in all cases, convergence seems very slow. The trajectories for β are more interesting. In figures 7 and 9 (initial variations: 0.05 and 0.1) we do observe what is predicted by the theory: small k_0 are preferable for short runs, larger ones become progressively better. More surprisingly, the trajectories seem to have a minimum value, then increase and stabilizes to a limit value. This behaviour should probably deserve more exploration. For large initial variations ($\delta = 0.25$), the trajectories are more erratic and uneasy to interpret.

6.3. Using the Hessian

The last two experiments (figures 13 and 14) attempt to compare algorithm (6) with $k_n = 1$ and (7). The latter is slightly better, but, clearly, the improvement is very small.

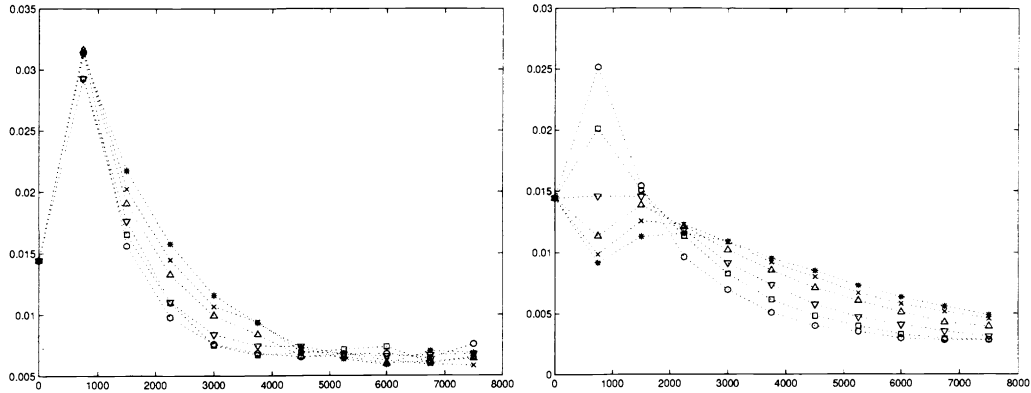


Figure 1. variance of the estimation in function of computer time. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.05$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

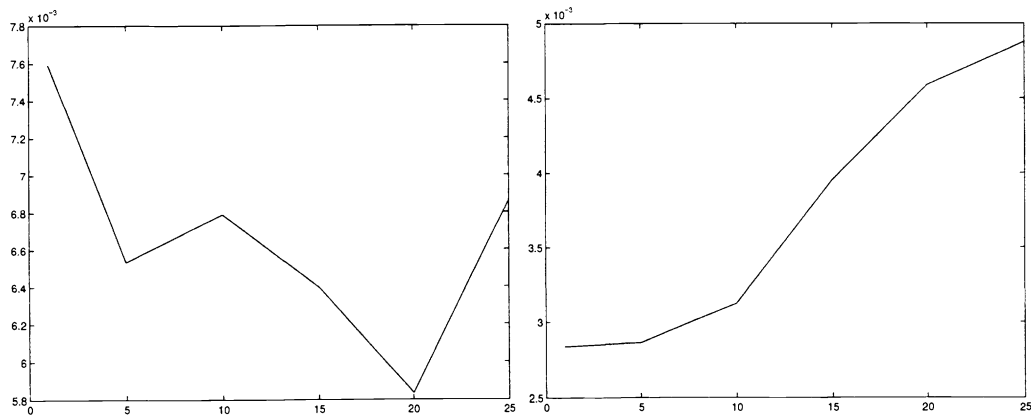


Figure 2. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.05$. standard errors for α and β in function of k_0 , at time $t = 7500$

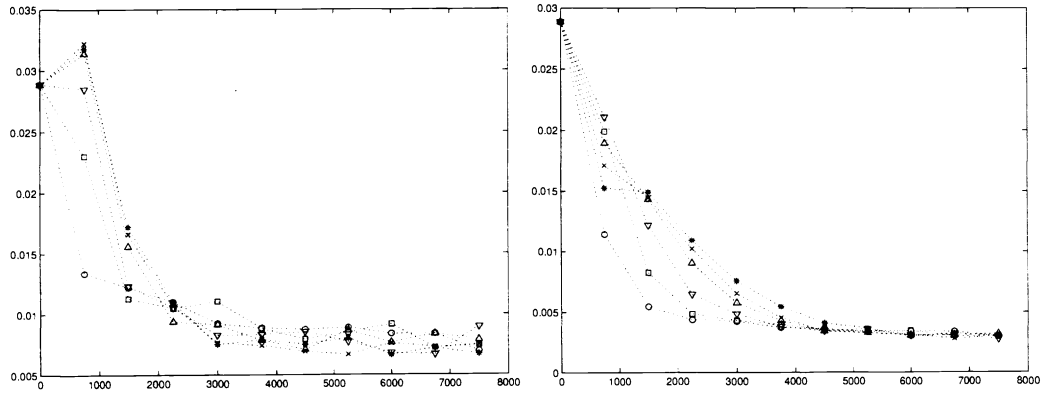


Figure 3. variance of the estimation in function of computer time. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.1$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

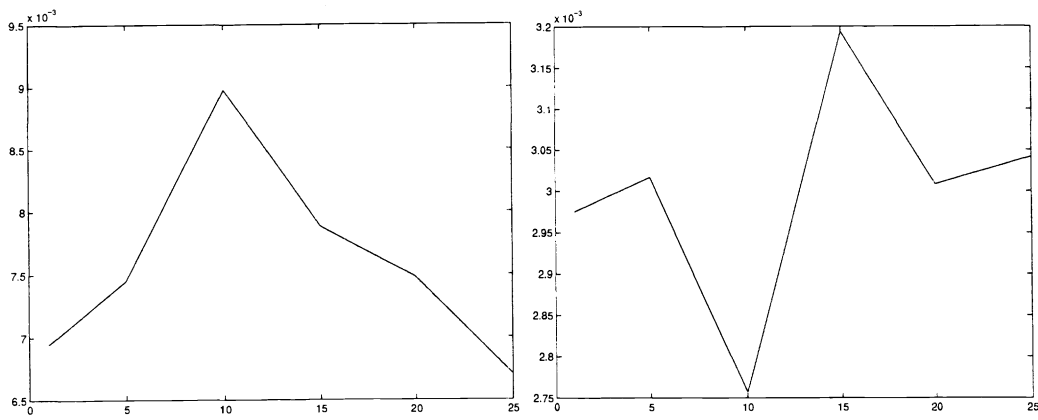


Figure 4. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.1$. standard errors for α and β in function of k_0 , at time $t = 7500$

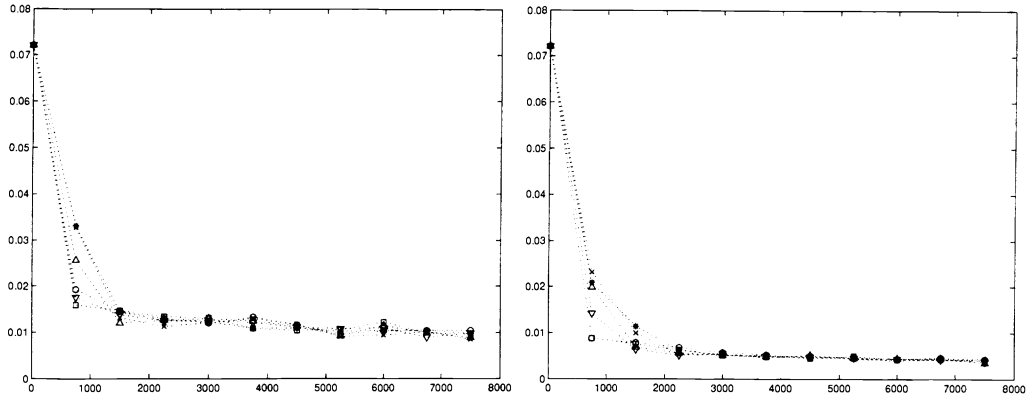


Figure 5. variance of the estimation in function of computer time. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.25$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

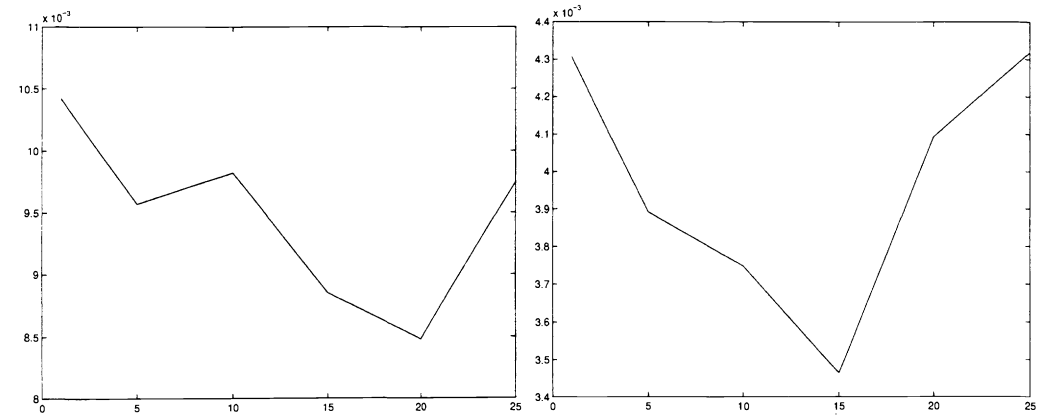


Figure 6. True parameters: $\alpha = 0$, $\beta = -0.5$; initial variation: $\delta = 0.25$. standard errors for α and β in function of k_0 , at time $t = 7500$

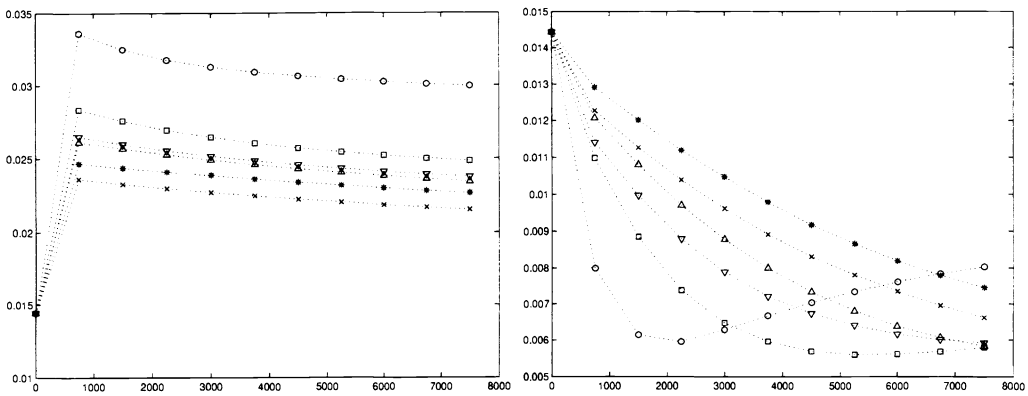


Figure 7. variance of the estimation in function of computer time. True parameters: $\alpha = 0$, $\beta = -0.7$; initial variation: $\delta = 0.05$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

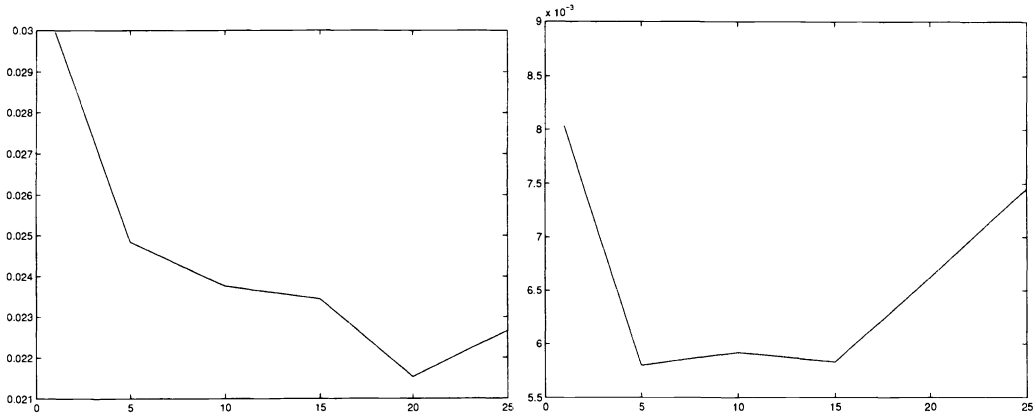


Figure 8. True parameters: $\alpha = 0$, $\beta = -0.7$; initial variation: $\delta = 0.05$. standard errors for α and β in function of k_0 , at time $t = 7500$

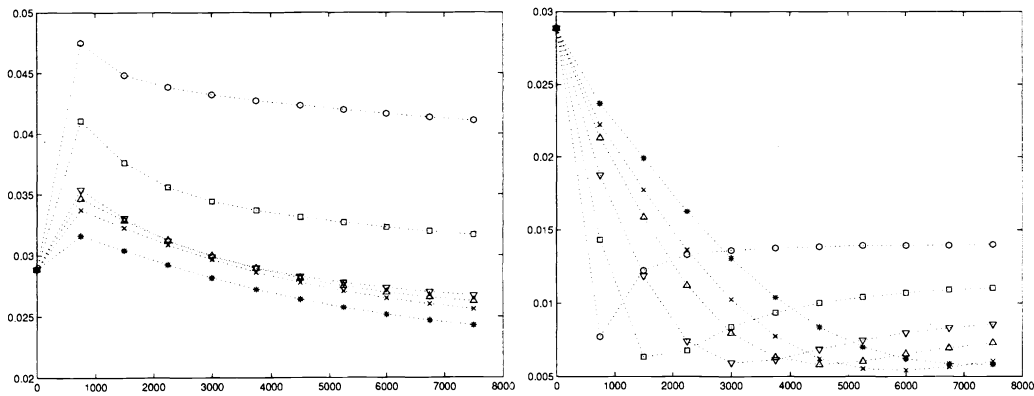


Figure 9. variance of the estimation in function of computer time. True parameters: $\alpha = 0$, $\beta = -0.7$; initial variation: $\delta = 0.1$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

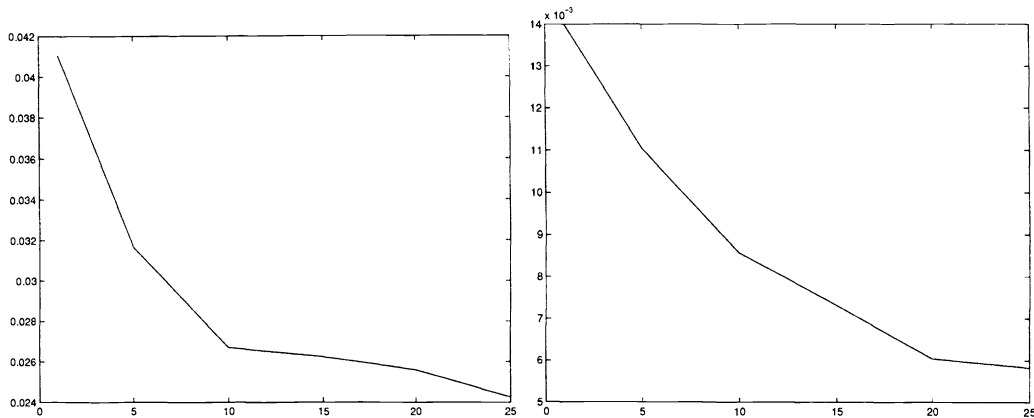


Figure 10. True parameters: $\alpha = 0$, $\beta = -0.7$; initial variation: $\delta = 0.1$. standard errors for α and β in function of k_0 , at time $t = 7500$

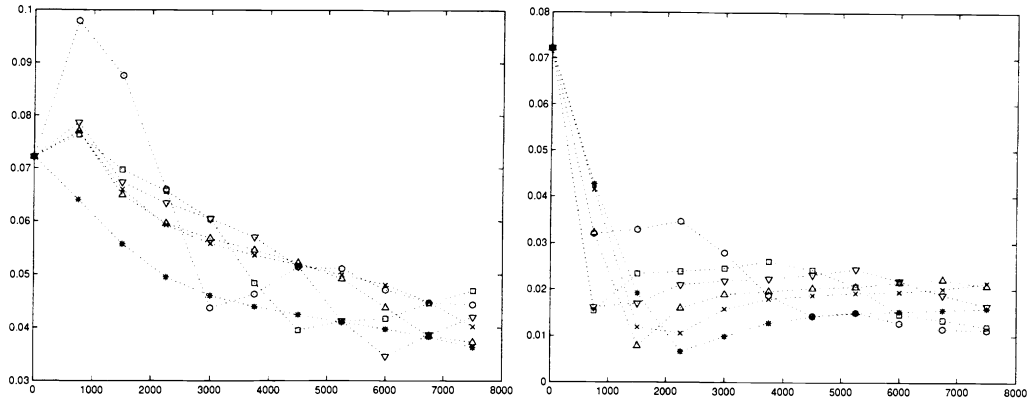


Figure 11. variance of the estimation in function of computer time. True parameters: $\alpha = 0, \beta = -0.7$; initial variation: $\delta = 0.25$. variance of α_n and β_n in function of n , for $k_0 = 1$ (\circ), $k_0 = 5$ (\square), $k_0 = 10$ (∇), $k_0 = 15$ (Δ), $k_0 = 20$ (\times), $k_0 = 25$ ($*$)

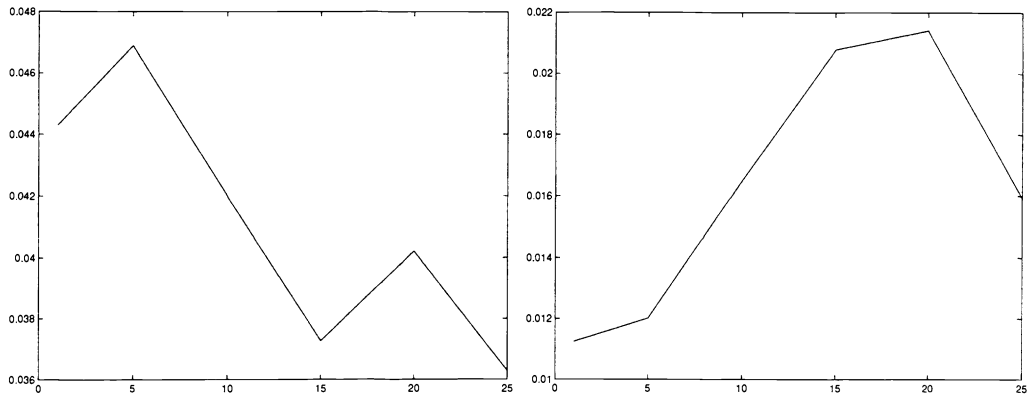


Figure 12. True parameters: $\alpha = 0, \beta = -0.7$; initial variation: $\delta = 0.25$. standard errors for α and β in function of k_0 , at time $t = 7500$

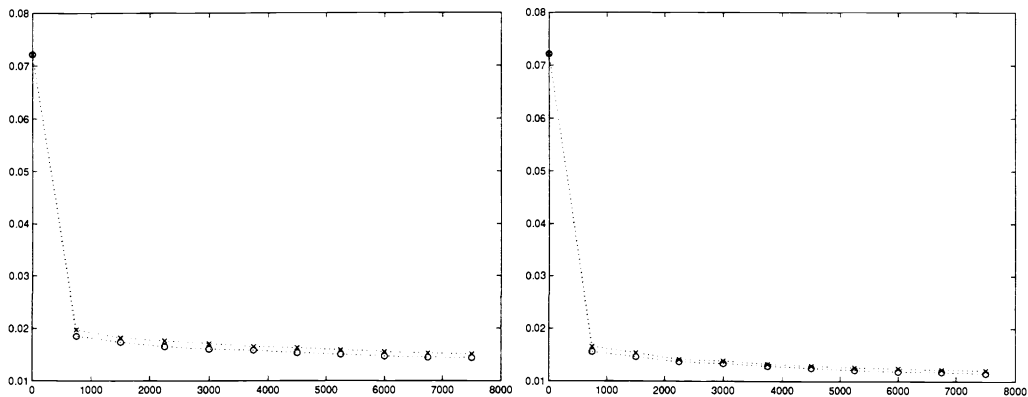


Figure 13. True parameters: $\alpha = -0.5, \beta = -0.5$; initial variation: $\delta = 0.25$. standard errors for α and β in function of the number of iterations (with hessian: \circ , without Hessian: \times)

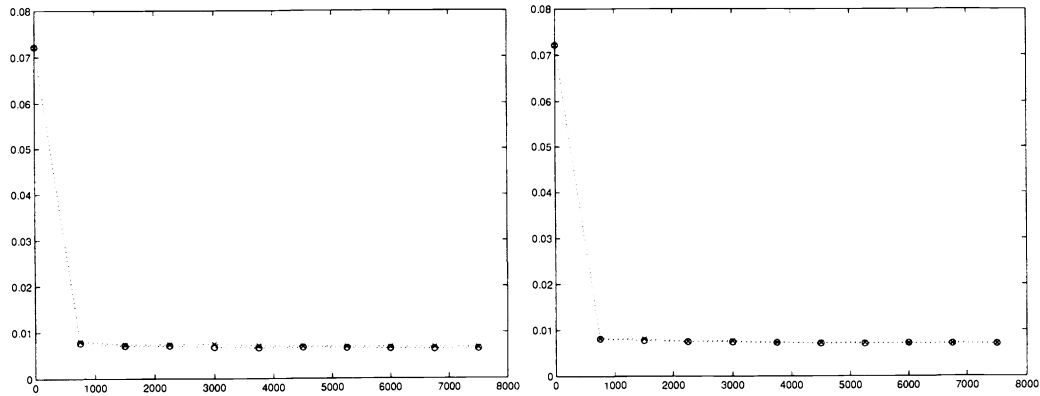


Figure 14. True parameters: $\alpha = -0.5$, $\beta = -0.5$; initial variation: $\delta = 0.1$. standard errors for α and β in function of the number of iterations (with hessian: o, without Hessian: x)

REFERENCES

1. R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.
2. A. Penttinen, "Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method," *Jy. Stud. Comput. Sci. Econ. Statist.* **7**, 1984.
3. A. Lippman, *A Maximum Entropy Method for Expert Systems*. PhD thesis, Brown University, 1986.
4. L. Younes, "Estimation and annealing for gibbsian fields," *Ann. de l'Inst. Henri Poincaré* **2**, 1988.
5. L. Younes, "On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates," to appear in *Stochastics and Stochastics Models*, 1998.
6. H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Math. Stat.* **22**, pp. 400–407, 1951.
7. A. Benveniste, M. Métivier, and P. Priouret, *Algorithmes Adaptatifs et Approximations Stochastiques, Théorie et Application*, Masson, 1987.
8. H. Peskun, P, "Optimum monte carlo sampling using markov chains," *Biometrika* **60**, pp. 607–612, 1973.
9. S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. PAMI* **6**, pp. 721–741, 1984.
10. A. Sokal, "Monte carlo methods in statistical mechanics." Lecture notes: Cours de troisième cycle de la physique en Suisse Romande, Lausanne, 1989.
11. D. Geman, "Random fields and inverse problems in imaging," in *Ecole d'été de Saint-Flour, Lecture Notes in Mathematics*, vol. 1427, Springer-Verlag, (New York), 1990.
12. A. Frigessi, C.-R. Hwang, and L. Younes, "optimal spectral structure of reversible stochastic matrices, monte carlo methods and the simulation of markov random fields," *Annals of Applied probability* **2**(3), pp. 610–628, 1992.