

# Monte Carlo maximization of likelihood: A convergence study

Laurent Younes  
CMLA, ENS de Cachan  
61 av. du Président Wilson  
94 235 Cachan CEDEX  
younes@cmla.ens-cachan.fr

**Abstract** We propose a rigorous study of an iterative maximization algorithm introduced by Geyer and Thompson for maximum likelihood estimation of Markov random fields. One step of the algorithm consists in a Monte-Carlo approximation of the likelihood, followed by a local maximization in the neighborhood of the current parameter. We study convergence properties of the induced process, and bound the computational complexity of the procedure. The main tool involved in the stochastic analysis are deviation inequalities and concentration of measure bounds applied to empirical processes.

**Keywords** Empirical processes, Gibbs distributions, Markov-Chain Monte Carlo, maximum likelihood estimator, Stochastic approximation

## 1 Introduction

In this paper, we study a numerical scheme which has been proposed in [5] for maximum likelihood parameter estimation of exponential families. The framework is the following: let  $\Omega$  be a finite set of very large cardinality (typically,  $\Omega = F^N$ , where  $F$  is finite and  $N$  is of order of hundreds or thousands). The considered models are indexed by  $\theta \in \mathbb{R}^d$ , and defined by

$$\pi_\theta(x) = \frac{1}{Z_\theta} e^{-\langle \theta, H(x) \rangle}, \quad x \in \Omega$$

where  $H : \Omega \rightarrow \mathbb{R}^d$  is a function and  $\langle \cdot, \cdot \rangle$  is the usual inner-product on  $\mathbb{R}^d$ .

The computation of the maximum likelihood for such models is an intricate numerical problem, since it requires finding the solution of

$$\sum_{x \in \Omega} H(x) e^{-\langle \theta, H(x) \rangle} = H(x_0) \sum_{x \in \Omega} e^{-\langle \theta, H(x) \rangle}$$

both terms being impossible to compute because of the overwhelming cardinality of  $\Omega$ . Feasible solutions to this problem pass by Monte-Carlo sampling and probabilistic numerical approximations.

There are essentially two ways for introducing such an approximation. The first one is to work on the gradient descent algorithm, and estimate the gradient step by Monte-Carlo simulations. This approach provides a Robbins-Monro stochastic approximation procedure, which has been studied in [17], [18] and [20]. Other references for practical applications of this approach are [7], [10], [11], [19], ...

The other point of view is to directly approximate the log-likelihood

$$f(\theta) := \log \pi_\theta(x_0) = -\log Z_\theta - \langle \theta, H(x_0) \rangle,$$

by Monte-Carlo sampling. More precisely, if  $X_1, \dots, X_k$  are random samples of a distribution  $\pi_* > 0$  on  $\Omega$ ,  $f(\theta)$  can be estimated by

$$-\langle \theta, H(x_0) \rangle - \log \frac{1}{k} \sum_{q=1}^k \frac{e^{-\langle H(X_q), \theta \rangle}}{\pi_*(X_q)}$$

and maximized by standard deterministic methods. The issue of course is how good this approximation can be, for reasonable (feasible) values of  $k$ . Finding a good probability  $\pi_*$  is the subject of *importance sampling* theory. It is easy to show that the best choice, for a fixed  $\theta$  (the one for which the variance of

$$\sum_{q=1}^k \frac{e^{-\langle H(X_q), \theta \rangle}}{\pi_*(X_q)}$$

is the smallest) for independent samples  $X_1, \dots, X_k$  with law  $\pi_*$  is provided by  $\pi_*(x) = \pi_\theta(x)$  for all  $x$ . This is however useless for our purposes, since  $f(\theta)$  has to be maximized in  $\theta$ : the distribution  $\pi_*$  must provide a small variance for all  $\theta$ .

However, because of this result, it can be expected that, for  $\psi \in \mathbb{R}^d$  the distribution  $\pi_* = \pi_\psi$  will provide good approximations of  $f(\theta)$  for  $\theta$  close to  $\psi$ . This yields the algorithm proposed in [4], (an ancestor of which being in [12] in the case of spatial points processes), which consists in iteratively maximizing an approximate log-likelihood in the neighbourhood of some current parameter  $\theta_n$ , to obtain the update  $\theta_{n+1}$ .

The object of this paper is to study this algorithm, and in particular to enlight the trade-off between the expected precision of the estimations and the overall numerical complexity of the algorithm. This will be the subject of section 6. Technical results will be obtained before this, in sections 3 and 4. The former will contain deterministic results related to the convergence of this kind of algorithm, under the hypothesis of a uniform control of the difference between the exact gradient of the likelihood and its stochastic approximation. Section 4 is devoted to the estimation of the probability of validity of such a uniform control, and is based on rough application of results on the speed of convergence of empirical processes.

Throughout this paper, we make the simplifying assumption that the samples  $(X_q, q = 1, \dots, k)$  which are used for the estimation of the likelihood are

independent. This is almost never the case in practical experiments, in which they typically follow a Markov chain controlled by the current parameter  $\theta_n$ . We describe how our results can be extended to this framework in section 7. We consider, however, that most of the interesting features of the analysis already appear in the simplified setting, while requiring less technicalities than what would be needed in the Markovian case. We therefore restrict to it in the main line of the paper. For the same reason, we work with rough bounds in the application of the theorems on empirical processes, some details on how getting better bounds being also provided in section 7.

## 2 Algorithm

We fix the notation, and give the precise definition of the algorithm under study. Let  $\Omega$  be a finite space and  $H$  a function  $H : \Omega \rightarrow \mathbb{R}^d$ . For each  $\theta \in \mathbb{R}^d$ , for each  $x \in \Omega$ , let

$$\pi_\theta(x) = \frac{e^{-\langle \theta, H(x) \rangle}}{Z_\theta}$$

where  $Z_\theta = \sum_{x \in \Omega} e^{-\langle \theta, H(x) \rangle}$ .

The algorithm proposed in [5] aim at maximizing  $\pi_\theta(x_0)$  for a given  $x_0 \in \Omega$  using a probability renormalization approach. Fix  $x_0$  and let  $H_0 = H(x_0)$ . Set

$$f(\theta) = -\langle \theta, H_0 \rangle - \log Z_\theta$$

This function is concave in  $\theta$ , and we shall assume, in the following, that it is strictly concave, and that it admits a unique maximum  $\theta^* \in \mathbb{R}^d$ . These assumptions are discussed in [5]: strict concavity is equivalent to the fact that there is no  $u \in \mathbb{R}^d$  such that  $x \mapsto \langle u, H(x) \rangle$  is constant, and the existence of the maximum to the fact that  $H_0$  lies in the interior of the convex hull of the points  $\{H(x), x \in \Omega\} \subset \mathbb{R}^d$ .

If  $\psi \in \mathbb{R}^d$ , a straightforward computation yields

$$f(\theta) = -\langle \theta, H_0 \rangle - \log Z_\psi - \log E_\psi \left[ e^{-\langle \theta - \psi, H \rangle} \right] \quad (1)$$

where  $E_\psi$  refers to the expectation with respect to  $\pi_\psi$ . Thus, if  $X_1, \dots, X_k$  is a sequence of iid variables with distribution  $\pi_\psi$ ,  $f(\theta)$  can be approximated by

$$\hat{f}_{\psi,k}(\theta) = -\langle \theta, H_0 \rangle - \log Z_\psi - \log \frac{1}{k} \sum_{p=1}^k e^{-\langle \theta - \psi, H(X_p) \rangle}$$

For any fixed  $\theta$ , and for large  $k$ , this converges to  $f(\theta)$ . Moreover, it can be shown that, if  $\theta_k$  is a maximizer of  $\hat{f}_{\psi,k}$  ( $\psi$  being fixed), and the sequence  $\theta_k$  is bounded, then, it converges, when  $k$  tends to infinity, to  $\theta^*$ , the maximizer of  $f$  (see [4]). However, for fixed  $k$ ,  $\hat{f}_{\psi,k}(\theta)$  is a good approximation of  $f(\theta)$  only for  $\theta$  in a neighborhood of  $\psi$ , the size of this neighborhood depending on

the number of samples,  $k$ . For a given precision, it appears that  $k$  should grow exponentially with the norm  $|\theta - \psi|$ , so that the computations would become intractable when  $\theta^*$  is too far away from  $\psi$ .

For this reason, the authors in [5] have introduced a recursive algorithm which only performs local maximizations of  $\hat{f}_{\psi,k}$ . This can be described as follows: fix  $\theta_0 \in \mathbb{R}^d$ , the starting point. Fix also two sequences:  $a_0, a_1, \dots$  of positive real numbers, and  $k_0, k_1, \dots$  of positive integers. At time  $n$ , let  $\theta_n$  be the current parameter, and define  $\theta_{n+1}$  such that

$$\hat{f}_{\theta_n, k_n}(\theta_{n+1}) = \max\{\hat{f}_{\theta_n, k_n}(\theta), |\theta - \theta_n| \leq a_n\} \quad (2)$$

where  $\hat{f}_{\theta_n, k_n}$  is computed like in (1), using  $X_1^n, \dots, X_{k_n}^n$ , a sequence of iid samples of  $\pi_{\theta_n}$ .

We study convergence properties of this algorithm. More precisely, we try to answer the question whether the parameter  $\theta_n$  will eventually pass in a neighborhood of  $\theta^*$ . Note that we are not interested in letting  $k_n$  tend to infinity, which would be too costly, nor letting  $a_n$  tend to 0, because the algorithm will, in the end, be equivalent to a stochastic Newton algorithm which can be studied using the methods of [20]. Also, the practical advantage of the algorithm lies in the fact that maximization is performed with positive  $a_n$ , hopefully as large as possible, and reasonably large values of  $k_n$ .

The study of the behaviour of the sequence  $(\theta_n)$  is split in two parts: we first analyze the consequences of the concavity of  $f$  to prove that a control of

$$\sup\{|\hat{f}'_{\theta_n, k_n}(\theta) - f'(\theta)| : |\theta - \theta_n| \leq a_n\}$$

is enough to ensure convergence, and then try to bound the probability that this supremum is small.

### 3 Deterministic bounds

In this section, we use the concavity of  $f$  and  $\hat{f}_{\theta,k}$  to show how a uniform control of the approximation of  $f'(\theta)$  by  $\hat{f}'_{\theta_n, k_n}(\theta)$  can be used to show that the algorithm eventually approaches a neighbourhood of  $\theta^*$ , the maximum likelihood estimator. This is essentially contained in the next lemma. For  $\psi \in \mathbb{R}^d$  and  $a > 0$ , we let

$$g_a(\psi) = \inf\{|f'(\theta)| : |\theta - \psi| \leq a\}$$

We have

**Lemma 1** *Let  $(\theta_p, p \geq 0)$  be generated by (2). Fix  $n \geq 0$  and let*

$$\eta_n = \sup\{|\hat{f}'_{\theta_n, k_n}(\theta) - f'(\theta)| : |\theta - \theta_n| \leq a_n\}$$

and

$$g_n = g_{a_n}(\theta_n) = \inf\{|f'(\theta)| : |\theta - \theta_n| \leq a_n\}$$

*Assume that  $\delta_n := g_n - 2\eta_n > 0$ : then  $f(\theta_{n+1}) - f(\theta_n) > \delta_n a_n$ .*

*Proof:* By concavity, we have:

$$\begin{aligned} f(\theta_{n+1}) - f(\theta_n) &\geq \langle f'(\theta_{n+1}), \theta_{n+1} - \theta_n \rangle \\ &\geq \langle f'_{\theta_n, k_n}(\theta_{n+1}), \theta_{n+1} - \theta_n \rangle - a_n \eta_n \end{aligned}$$

The fact that  $\delta_n > 0$  implies that  $|f'_{\theta_n, k_n}(\theta)| > g_n - \eta_n > 0$  for  $|\theta - \theta_n| \leq a_n$ , which implies that  $\theta_{n+1}$  cannot be a global maximizer of  $f'_{\theta_n, k_n}$  and therefore that  $|\theta_{n+1} - \theta_n| = a_n$ . Thus, because  $\theta_{n+1}$  is a maximizer,  $f'_{\theta_n, k_n}(\theta_{n+1})$  must be colinear to  $\theta_{n+1} - \theta_n$  (which is normal to the ball centered at  $\theta_n$ ), and more precisely

$$\langle f'(\theta_{n+1}), \theta_{n+1} - \theta_n \rangle = |f'(\theta_{n+1})| |\theta_{n+1} - \theta_n| \geq (g_n - \eta_n) a_n$$

so that  $f(\theta_{n+1}) - f(\theta_n) > \delta_n a_n$ .  $\square$

**Corollary 1** *Assume that  $\sum_n a_n = +\infty$ . Assume that for some  $n_0 > 0$  and some constants  $a$  and  $\eta$ , one has  $\eta_n < \eta$  and  $a_n < a$ , for all  $n \geq n_0$ . Define*

$$q_\alpha = \inf \{g_a(\theta) : f(\theta) \leq \alpha\}$$

*Then, if  $q_\alpha - 2\eta > 0$ , there exists  $n_1 \geq n_0$  such that  $f(\theta_n) \geq \alpha - |H|a$  for all  $n \geq n_1$ , where*

$$|H| = \max \{|H(x) - H(y)|, x, y \in \Omega\}$$

*Proof:* Set  $U_\alpha = \{\theta : f(\theta) \leq \alpha\}$ . By the previous lemma, if  $q_\alpha > 2\eta$ , the sequence  $\theta_n$  may only stay a finite time in  $U_\alpha$ , since, if  $\theta_n, \dots, \theta_{n+p}$  stay in  $U_\alpha$ ,  $f(\theta_{n+p}) - f(\theta_n) \geq (q_\alpha - 2\eta) \sum_{i=n}^{n+p-1} a_i$  (by definition of  $q_\alpha$ ), and the lower bound can be made arbitrary large.

So, there is an integer  $n_1$  such that  $f(\theta_{n_1}) > \alpha$ . In addition, each time a  $\theta_n$  goes back to  $U_\alpha$ , the sequence  $f(\theta_{n+p})$  increases until  $\theta_{n+p}$  goes out of  $U_\alpha$ . Since  $|f(\theta_n) - f(\theta_{n+1})| \leq |H| |\theta_n - \theta_{n+1}| \leq a|H|$ , for any  $n \geq n_1$ ,  $f(\theta_n)$  can never become smaller than  $\alpha - |H|a$ , as announced.  $\square$

Let  $\alpha^* = \max(f) = f(\theta^*)$ . The set  $U_\alpha^c = \mathbb{R}^d \setminus U_\alpha$  for  $\alpha < \alpha^*$  is an open neighborhood of  $\theta^*$ . The previous corollary implies then that, if the errors  $\eta_n$  are small and  $a$  is small, the sequence  $(\theta_n)$  will stay close to  $\theta^*$ .

## 4 Stochastic bounds

Fix  $\psi \in \mathbb{R}^d$ . Our next step is to bound the probability

$$P \left( \sup \left\{ |\hat{f}'_{\psi, k}(\theta) - f'(\theta)| : |\theta - \psi| \leq a \right\} > \eta \right)$$

in function of  $k$ ,  $\eta$  and  $a$ . This will provide a confidence estimate for the validity of lemma 1. More precisely, we prove

**Theorem 1** For  $\psi \in \mathbb{R}^d$ ,

$$P \left( \sup \left\{ |\hat{f}'_{\psi, k}(\theta) - f'(\theta)| : |\theta - \psi| \leq a \right\} > \eta \right) \\ \leq 6d \exp \left[ -\frac{1}{K} \frac{k\eta - v(a)\sqrt{k}}{b(a)} \log \left( 1 + \frac{k\eta - v(a)\sqrt{k}}{kb(a)} \right) \right]$$

with  $b(a) = e^{a|H|}(|H| + \eta) + \eta$  and  $v(a) = 2K''\sqrt{d}|H|(|H| + \eta)e^{a|H|}$  where  $K$  and  $K''$  are universal constants.

Here, we have used the notation  $|H| = \max \{ H^{(i)}(x) - H^{(i)}(y) : i = 1, \dots, d, x, y \in \Omega \}$ . The proof uses deviations inequalities for empirical processes which are recalled in the next section.

## 4.1 Deviations inequalities

We use a concentration of measure theorem, proved by Talagrand:

**Theorem 2 (Talagrand)** Let  $Y_1, \dots, Y_n$  be independent random variables on a probability space  $\Omega$ . Let  $\mathcal{F}$  be a countable family of functions  $f : \Omega \rightarrow \mathbb{R}$ . Denote by  $Z$  the random variable

$$Z = \sup_{f \in \mathcal{F}} \left\{ \left| \sum_{i=1}^n f(Y_i) \right| \right\}$$

Then,

$$P[Z \geq E[Z] + t] \leq 3 \exp \left[ -\frac{1}{K} \frac{t}{C} \log \left( 1 + \frac{Ct}{\sigma^2} \right) \right] \quad (3)$$

where  $C = \sup_{f \in \mathcal{F}} \|f\|_\infty$ ,

$$\sigma^2 = E \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)^2 \right]$$

and  $K$  is a universal constant.

Note that, since  $\sigma^2 \leq kC^2$ , the weaker bound is valid

$$P[Z \geq E[Z] + t] \leq 3 \exp \left[ -\frac{1}{K} \frac{t}{kC} \log \left( 1 + \frac{t}{kC} \right) \right] \quad (4)$$

which will be enough for our purposes.

To use this theorem, one must also control the value of the expectation  $E(Z)$ . We handle this with the following estimate (see, for example, [16] for a proof):

**Theorem 3 (Dudley)** Let  $(Y_t, t \in T)$  be a family of random variables indexed by a metric space  $T$ , provided with a distance  $\gamma$ , such that

$$\forall s, t \in T, \forall \lambda > 0, P(|Y_s - Y_t| > \lambda) \leq 2 \exp\left(-\frac{\lambda^2}{\gamma(s, t)^2}\right) \quad (5)$$

One has, for all finite subset  $F \subset T$

$$E \sup_{t \in F} Y_t \leq K' \int_0^\infty \sqrt{\log N(T, \gamma, \epsilon)} d\epsilon \quad (6)$$

where  $N(T, \gamma, \epsilon)$  is the maximal number cardinality of the subsets of  $T$  which contain only elements at distance larger than  $\epsilon$  of each other, and  $K'$  is a universal constant.

Since the right-hand term in (6) is independent of  $F$ , the left-hand term can be replaced by

$$\sup \left\{ E \sup_{t \in F} Y_t : F \subset T, F \text{ finite} \right\}$$

which will be equal to  $E \sup_{t \in T} X_t$  as soon as an approximation argument by a denumerable dense subset can be used (for example if  $t \mapsto X_t$  is continuous, which will be the case when we shall apply this theorem).

## 4.2 Proof of theorem 1

We have

$$\hat{f}'_{\psi, k}(\theta) = \frac{\sum_{p=1}^k H(X_p) e^{-(\theta - \psi, H(X_p))}}{\sum_{p=1}^k e^{-(\theta - \psi, H(X_p))}} - H_0$$

and

$$f'(\theta) = E_\theta(H) - H_0 = \frac{Z_\psi}{Z_\theta} E_\psi [H e^{-(\theta - \psi, H)}] - H_0$$

Letting  $H^{(i)}$  be the  $i$ th component of  $H : \Omega \rightarrow \mathbb{R}^d$ , we have

$$\begin{aligned} & P \left( \sup \left\{ |\hat{f}'_{\psi, k}(\theta) - f'(\theta)| : |\theta - \psi| \leq a \right\} > \eta \right) \\ &= P \left( \sup \left\{ \left| \frac{\sum_{p=1}^k H(X_p) e^{-(\theta - \psi, H(X_p))}}{\sum_{p=1}^k e^{-(\theta - \psi, H(X_p))}} - E_\theta(H) \right| : |\theta - \psi| \leq a \right\} > \eta \right) \\ &\leq d \max_i P \left( \sup \left\{ \left| \frac{\sum_{p=1}^k H^{(i)}(X_p) e^{-(\theta - \psi, H(X_p))}}{\sum_{p=1}^k e^{-(\theta - \psi, H(X_p))}} - E_\theta(H) \right| : |\theta - \psi| \leq a \right\} > \eta \right) \end{aligned}$$

We now fix  $i$  and set  $T_\theta = e^{-(\theta - \psi, H)}$  and  $S_\theta = H^{(i)} T_\theta$ . Let

$$A(k, a, \eta) = P \left( \sup \left\{ \left| \frac{\sum_{p=1}^k S_\theta(X_p)}{\sum_{p=1}^k T_\theta(X_p)} - \frac{E_\psi(S_\theta)}{E_\psi(T_\theta)} \right| : |\theta - \psi| \leq a \right\} > \eta \right)$$

We have

$$\begin{aligned}
& A(k, a, \eta) = \\
& P \left( \sup_{|\theta - \psi| \leq a} \left[ \left| E_\psi(T_\theta) \sum_{p=1}^k S_\theta(X_p) - E_\psi(S_\theta) \sum_{p=1}^k T_\theta(X_p) \right| - \eta E_\psi(T_\theta) \sum_{p=1}^k T_\theta(X_p) \right] > 0 \right) \\
& \leq P \left( \sup_{|\theta - \psi| \leq a} \left[ E_\psi(T_\theta) \sum_{p=1}^k S_\theta(X_p) - E_\psi(S_\theta) \sum_{p=1}^k T_\theta(X_p) - \eta E_\psi(T_\theta) \sum_{p=1}^k T_\theta(X_p) \right] > 0 \right) \\
& + P \left( \sup_{|\theta - \psi| \leq a} \left[ E_\psi(S_\theta) \sum_{p=1}^k T_\theta(X_p) - E_\psi(T_\theta) \sum_{p=1}^k S_\theta(X_p) - \eta E_\psi(T_\theta) \sum_{p=1}^k T_\theta(X_p) \right] > 0 \right)
\end{aligned}$$

Letting

$$U_\theta^+(x) = \frac{S_\theta(x)}{E_\psi(T_\theta)} - \frac{E_\psi(S_\theta)}{E_\psi(T_\theta)} \frac{T_\theta(x)}{E_\psi(T_\theta)} - \eta \frac{T_\theta(x)}{E_\psi(T_\theta)} + \eta$$

and

$$U_\theta^-(x) = \frac{E_\psi(S_\theta)}{E_\psi(T_\theta)} \frac{T_\theta(x)}{E_\psi(T_\theta)} - \frac{S_\theta(x)}{E_\psi(T_\theta)} - \eta \frac{T_\theta(x)}{E_\psi(T_\theta)} + \eta$$

this can be written

$$\begin{aligned}
A(k, a, \eta) & \leq P \left( \sup \left\{ \sum_{p=1}^k U_\theta^+(X_p) : |\theta - \psi| \leq a \right\} > k\eta \right) \\
& + P \left( \sup \left\{ \sum_{p=1}^k U_\theta^-(X_p) : |\theta - \psi| \leq a \right\} > k\eta \right).
\end{aligned}$$

Since  $E_\psi(U_\theta^+) = E_\psi(U_\theta^-) = 0$ , we now are in position to apply theorem 2: let

$$Z^+ = \sup \left\{ \sum_{p=1}^k U_\theta^+(X_p) : |\theta - \psi| \leq a \right\}$$

By theorem 2,

$$P[Z^+ \geq k\eta] \leq 3 \exp \left[ -\frac{1}{K} \frac{t}{C^+} \log \left( 1 + \frac{t}{kC^+} \right) \right]$$

where  $t = k\eta - E(Z^+)$ ,  $C^+ = \sup \{U_\theta^+(x), x \in \Omega, |\theta - \psi| \leq a\}$ . Since  $\frac{t}{C^+} \log \left( 1 + \frac{t}{C^+} \right)$  is decreasing in  $C^+$  and increasing in  $t = k\eta - E(Z^+)$ , we need upper bounds for  $C^+$  and  $E(Z^+)$ . For  $C^+$ , we have

$$\begin{aligned}
U_\theta^+(x) & = \frac{Z_\psi}{Z_\theta} H^{(i)}(x) e^{-(\theta - \psi, H(x))} - E_\theta(H^{(i)}) \frac{Z_\psi}{Z_\theta} e^{-(\theta - \psi, H(x))} \quad (7) \\
& + \eta \left( 1 - \frac{Z_\psi}{Z_\theta} e^{-(\theta - \psi, H(x))} \right) \\
& = \frac{Z_\psi}{Z_\theta} e^{-(\theta - \psi, H(x))} \left( H^{(i)}(x) - E_\theta(H^{(i)}) - \eta \right) + \eta
\end{aligned}$$

We have

$$|\log Z_\psi - \log Z_\theta - \langle \theta - \psi, H(x) \rangle| \leq |\theta - \psi| \cdot |H| \quad (8)$$

since the differential of the left hand term is  $E_\theta(H) - H(x)$ . We thus get

$$|U_\theta^+(x)| \leq b(a)$$

with  $b(a) = e^{a|H|}(|H| + \eta) + \eta$ , which gives an upper bound for  $C^+$ .

We now provide an upper-bound for  $E(Z^+)$ , based on theorem 3. We apply this theorem to  $T = \{\theta : |\theta - \psi| < a\}$ , and to  $Y_\theta = \sum_{p=1}^k U_\theta^+(X_p)$ . We first show that (5) is true for  $\gamma(\theta, \theta') = L|\theta - \theta'|$  for a suitable constant  $L$ . Indeed, we have

**Lemma 2** *For all  $x \in \Omega$ ,  $U_\theta^+(x)$  is a Lipschitz function of  $\theta$ , the Lipschitz constant over  $\{\theta : |\theta - \psi| < a\}$  being bounded by  $|H|(|H| + \eta)e^{a|H|}$ .*

*Proof:* The fact that  $U_\theta^+$  is Lipschitz is obvious, since it is differentiable in  $\theta$ , so we only need to estimate the derivative. Fix  $\theta, \theta'$  in the ball with center  $\psi$  and radius  $a$ . We have, using (8) in the first estimate

$$\left| \frac{d}{d\theta} \left( \frac{Z_\psi}{Z_\theta} e^{-\langle \theta - \psi, H(x) \rangle} \right) \right| = |E_\theta(H) - H(x)| \frac{Z_\psi}{Z_\theta} e^{-\langle \theta - \psi, H(x) \rangle} \leq |H| e^{a|H|}$$

$$\left| \frac{d}{d\theta} E_\theta(H^{(i)}) \right| = \text{Var}_\theta(H^{(i)}) \leq |H|^2$$

which implies, using equation (7), that the derivative of  $U_\theta^+(x)$  is uniformly bounded by

$$2|H|(|H| + \eta)e^{a|H|}$$

□

Let  $\lambda > 0$  be given. We now apply Hoeffding's inequality: if  $V_1, \dots, V_k$  are iid centered random variables such that  $|V_i| \leq M$  almost surely then

$$P\left(\left|\sum_{i=1}^k V_i\right| > \lambda\right) < 2 \exp\left(-\frac{\lambda^2}{2kM^2}\right) \quad (9)$$

This can be applied to  $Y_p = U_\theta^+(X_p) - U_{\theta'}^+(X_p)$ , taking  $M = 2|H|(|H| + \eta)e^{a|H|}|\theta - \theta'|$ . This yields

$$P(|Y_\theta - Y_{\theta'}| > \lambda) \leq 2 \exp\left(-\frac{\lambda^2}{2kM^2}\right)$$

Let  $L = 2\sqrt{2k}|H|(|H| + \eta)e^{a|H|}$ . We have (5), and thus (6) for  $\gamma(\theta, \theta') = L|\theta - \theta'|$ . Moreover,

$$N(T, \gamma, \epsilon) \leq \max\left[1, \left(\frac{L}{\epsilon}\right)^d\right]$$

so that theorem 3 yields

$$\begin{aligned} E(Z^+) &\leq K' \int_0^L \sqrt{-d \log \frac{\epsilon}{L}} \\ &= 2K' \sqrt{dL} \int_0^1 \sqrt{-\log \epsilon} d\epsilon \end{aligned}$$

and thus

$$E(Z^+) \leq 2K'' \sqrt{2kd} |H| (|H| + \eta) e^{a|H|} \quad (10)$$

for the universal constant  $K'' = \sqrt{\pi} K'$ .

We thus have, letting

$$v(a) = 2K'' |H| (|H| + \eta) e^{a|H|} \sqrt{2d}$$

$$P[Z^+ \geq k\eta] \leq 3 \exp \left[ -\frac{1}{K} \frac{k\eta - v(a)\sqrt{k}}{b(a)} \log \left( 1 + \frac{k\eta - v(a)\sqrt{k}}{kb(a)} \right) \right].$$

Since the same inequality is obviously true for  $Z^-$ , this yields theorem 1.

## 5 Return times in a neighborhood of $\theta^*$

In this section, we assume that the algorithm is run with fixed  $a$  and  $k$ , and consider the sequence  $(\theta_n, n \geq 0)$  generated by (2). This sequence is a Markov chain, since it can be written

$$\theta_{n+1} = F(\theta_n, X_1^n, \dots, X_k^n)$$

with  $X_1^n, \dots, X_k^n$  independent, with distribution  $\pi_{\theta_n}$ , and

$$F(\theta_n, x_1, \dots, x_k) = \operatorname{argmax} \left\{ \hat{f}_{\theta_n}(\theta) : |\theta - \theta_n| \leq a \right\}$$

with

$$\hat{f}_{\theta_n}(\theta) = -\langle \theta, H_0 \rangle - \log Z_\psi - \log \sum_{p=1}^k e^{-\langle \theta - \phi, H(x_p) \rangle}.$$

Recall the notation

$$g(\psi) = \inf \{ |f'(\theta)| : |\theta - \psi| \leq a \}$$

**Proposition 1** *Let  $V$  be a neighborhood of  $\theta_*$  such that*

$$\inf \{ g(\psi), \psi \notin V \} > 0.$$

*Then there exists a constant  $k_V$  such that, if  $k \geq k_V$ , the sequence  $\theta_n$  returns to  $V$  infinitely often.*

*Proof:* This proposition therefore states that the Markov chain  $(\theta_n)$  is recurrent in the considered neighborhood,  $V$ . Notice that this neighborhood is small when  $a$  is small.

Denote by  $A_n^p$  the event that  $\theta_n$  stays out of  $V$  at least  $p$  times before step  $n$ :

$$A_n^p = \left[ \sum_{q=0}^n \mathbf{1}_{[\theta_n \notin V]} \geq p \right]$$

Fix  $\eta > 0$  such that,  $\inf \{g(\psi), \psi \notin V\} > 3\eta$ , so that, if  $\psi \notin V$ ,  $g(\psi) - 2\eta > \eta$ . Define the events

$$U_i = \left[ \sup \left\{ |\hat{f}'_{\theta_i, k}(\theta) - f'(\theta)| : |\theta - \theta_i| \leq a \right\} > \eta \right]$$

and

$$B_n^p = \left[ \sum_{q=0}^n \mathbf{1}_{U_i} \geq p \right]$$

We first prove ( $[x]$  being the integer part of  $x \in \mathbb{R}$ ):

**Lemma 3** *Assume that  $\theta_0 \in V$ . There exists  $\rho > 0$  such that  $A_n^p \subset B_n^{[\rho p]}$  for all  $n$  and  $p$*

*Proof:* Introduce the random times:

$$\sigma_0 = \min \{q \geq 0 : \theta_q \notin V\}$$

and for  $p > 0$

$$\tau_p = \min \{q \geq \sigma_{p-1} : \theta_q \in V\}$$

and

$$\sigma_p = \min \{q \geq \tau_p : \theta_q \notin V\}$$

Let  $p_n$  be such that  $\sigma_{p_n} = \max \{\sigma_p : \sigma_p \leq n\}$ . Lemma 1 implies that, if  $\theta_q \in V$  and  $\theta_{q+1} \notin V$ , then

$$\sup \{ |\hat{f}'_{\theta_q, k}(\theta) - f'(\theta)| : |\theta - \theta_q| \leq a \} > \eta$$

This already implies that, for each  $p \leq p_n$ , the event  $U_{\sigma_p}$  is true. We now give a lower bound of the number of times  $U_i$  is true for  $\sigma_p < i < \min \{\tau_p, n\}$ .

Again from lemma 1, if  $\theta_q \notin V$ , then either

$$\sup \{ |\hat{f}'_{\theta_q, k}(\theta) - f'(\theta)| : |\theta - \theta_q| \leq a \} > \eta$$

or  $f(\theta_{q+1}) - f(\theta_q) > \eta a$ . Since  $|f'(\theta)| < |H|$ , we have, in any case,  $f(\theta_{q+1}) - f(\theta_q) > -|H|a$ . Setting  $r_p = \min \{\tau_p, n\}$ , and  $\nu_p$  the number of  $q$  between  $\sigma_p$  and  $r_p$  such that  $U_q$  is true, we have

$$f(\theta_{r_p}) - f(\theta_{\sigma_p}) \geq (r_p - \sigma_p - \nu_p)\eta a - \nu_p |H|a$$

Let  $f_V = \max\{f(\theta), \theta \notin V\}$ : we have  $f(\theta_{r_p}) \leq f_V + a|H|$  and  $f(\theta_{\sigma_p}) \geq f_V - a|H|$  which yields

$$2a|H| \geq (r_p - \sigma_p)\eta a - \nu_p(\eta + |H|)a$$

or

$$\nu_p + 2 \geq \frac{(r_p - \sigma_p)\eta - 2|H|}{\eta + |H|} \geq (r_p - \sigma_p) \frac{\eta}{\eta + |H|}$$

which yields, since  $\nu_p + 2 \leq 2(\nu_p + 1)$

$$\nu_p + 1 \geq \frac{\eta}{2(\eta + |H|)}$$

Now, assume that  $A_n^s$  is true, which yields

$$s = \sum_{p=0}^{p_n} (r_p - \sigma_p)$$

The number of occurrences of  $U_i$  between 0 and  $n$  is  $t = p_n + \sum_{p=0}^{p_n} \nu_p$  and we get:

$$t \geq \frac{\eta}{2(\eta + |H|)} s$$

which yields the lemma with  $\rho = \frac{\eta}{2(\eta + |H|)}$  □

This yields an upper bound on  $P(A_n^p)$ . Indeed, a direct application of theorem 1 and of the Markov property for  $\theta_n$  yields the fact that

$$P(B_n^p) \leq \binom{n}{p} \epsilon(k)^p$$

with

$$\epsilon(k) = 6d \exp \left[ -\frac{1}{K} \frac{k\eta - v\sqrt{k}}{b} \log \left( 1 + \frac{k\eta - v\sqrt{k}}{kb} \right) \right]$$

$b$  and  $v$  being given in theorem 1. This yields

$$P(A_n^p) \leq \binom{n}{[\rho]p} \epsilon^{[\rho]p}$$

In particular, we can take  $p = [\lambda n]$  for some  $\lambda \in ]0, 1[$ . From Stirling's formula,  $\binom{n}{[\rho]p}$  is equivalent to

$$\frac{1}{2\pi n \alpha (1 - \alpha)} \beta^n$$

with  $\alpha = \lambda\rho$  and

$$\beta = \exp[-(1 - \alpha) \log(1 - \alpha) - \alpha \log \alpha]$$

If we take  $k$  large enough such that

$$\rho \log \epsilon(k) - (1 - \rho) \log(1 - \rho) - \rho \log \rho < 0$$

then, for some  $\lambda < 1$ ,

$$\alpha \log \epsilon(k) - (1 - \alpha) \log(1 - \alpha) - \alpha \log \alpha < 0$$

(with  $\alpha = \lambda\rho$ ). This implies that  $P(A_n^{[\lambda n]})$  tends to 0 at exponential speed. From Borel-Cantelli lemma, this implies that, almost surely,  $A_n^{[\lambda n]}$  is true for a finite number of  $n$ , and thus that for  $n$  large enough, the proportion of instances of  $\theta_p \in V$  for  $p \leq n$  is larger than  $(1 - \lambda)n$ .  $\square$

## 6 Comparison of strategies and cost optimization

Because

$$\begin{aligned} & |\sup \{f(\theta) : |\theta - \psi| \leq a\} - \sup \{f_{\psi,k}(\theta) : |\theta - \psi| \leq a\}| \\ & \leq a \sup \left\{ |\hat{f}'_{\psi,k}(\theta) - f'(\theta)| : |\theta - \psi| \leq a \right\}, \end{aligned}$$

theorem 1 implies that, when  $k$  tends to infinity, and for  $a > |\psi - \theta_*|$ , the probability for the Monte-Carlo supremum to be far from the maximum likelihood decreases exponentially fast. However, the values of  $k$  for which this bound becomes small increases exponentially fast in  $a$ , as does  $b((a))$ : thus, for large  $a$ , the required number of samples is likely to be untractably large for inducing a reasonably accurate estimation.

Iterating small steps, as suggested in [5], seems to be a better choice from the point of view of our estimates. In fact, we shall see that the computational cost of the algorithm increases only as a polynomial in  $|H|$  if  $a$  is selected adaptively.

A *strategy*, for algorithm 2 is the choice of a number of steps  $n$ , and of the values of  $a_p$  and  $k_p$  at each step. We write  $S = (n, a_1, \dots, a_n, k_1, \dots, k_n)$ . Several notions can be associated to  $S$ . Define its range to be the maximal distance which can be covered by the algorithm, that is

$$R(S) = \sum_{p=1}^n a_p$$

The computational cost of the strategy can be defined by

$$\Gamma(S) = \sum_{p=1}^n k_p$$

Finally, define the accumulated risk at level  $\eta$  by

$$\Sigma_\eta(S) = \sum_{p=1}^n \epsilon(\eta, k_p, a_p)$$

with

$$\epsilon(\eta, k, a) = 6d \exp \left[ -\frac{1}{K} \frac{(k\eta - v\sqrt{k})^+}{b} \log \left( 1 + \frac{(k\eta - v\sqrt{k})^+}{kb} \right) \right]$$

$v = v(\eta, a)$  and  $b = b(\eta, a)$  being given in theorem 1.

One may ask the following natural question: given a maximal acceptable value  $\sigma$  for  $\Sigma_\eta$  and a fixed value  $r$  of  $R(S)$  (based, for example, on the maximal distance the starting point of the algorithm is expected to be from  $\theta_*$ ), what strategy  $S$  will provide the minimal cost  $\Gamma(S)$ ? Answering this question in full generality seems quite intractable. To simplify, we assume, here again that the values of  $a_p$  and  $k_p$  are fixed to  $a$  and  $k$  for all  $p$ : in this case, we have  $r = na$ ,  $\gamma := Ga(S) = nk$ . We thus have  $n = r/a$ ,  $k = (\gamma/r)a$  and

$$\Sigma_\eta = \frac{r}{a} \epsilon(\eta, \frac{\gamma}{r}a, a)$$

Clearly,  $k$  should be larger than the minimal value for which the evaluation of  $\Sigma_\eta$  is efficient, namely,  $k\eta - v\sqrt{k} > 0$ ; this yields

$$\gamma \geq \gamma_0(a) = \frac{r}{a} \left( \frac{v}{\eta} \right)^2$$

So, take  $\gamma = \gamma_0(a)(1 + \delta)$  for some  $\delta > 0$ . We have, since  $k = (\gamma/r)a$

$$\begin{aligned} k\eta - v\sqrt{k} &= \gamma\eta a/r - \sqrt{(\gamma a/r)}v \\ &= \frac{v^2}{\eta}(1 + \delta - \sqrt{1 + \delta}) \end{aligned}$$

and

$$\begin{aligned} \frac{k\eta - v\sqrt{k}}{kb} &= \frac{v^2}{\eta kb}(1 + \delta - \sqrt{1 + \delta}) \\ &= \frac{\eta}{b} \left( 1 - \frac{1}{\sqrt{1 + \delta}} \right) \end{aligned}$$

so that, for such a  $\gamma$ , letting  $g(\delta) = (1 + \delta) - \sqrt{1 + \delta}$ ,

$$\frac{r}{a} \epsilon(\eta, \frac{\gamma}{r}a, a) = \exp \left[ -g(\delta) \frac{v^2}{\eta b} \log \left( 1 + \frac{\eta}{b} \left( 1 - \frac{1}{\sqrt{1 + \delta}} \right) \right) - \log(a/r) \right]$$

so that, for any value of  $a$ ,  $\delta$  should be selected such that

$$g(\delta) \frac{v^2}{\eta b} \log \left( 1 + \frac{\eta}{b} \left( 1 - \frac{1}{\sqrt{1 + \delta}} \right) \right) \geq -\log(a/r) - \log \sigma_\eta$$

Noting that  $g(\delta) \geq \delta/2$ , the condition can be further simplified in

$$\delta \log \left( 1 + \frac{\eta}{b} \left( 1 - \frac{1}{\sqrt{1 + \delta}} \right) \right) \geq -\frac{2b}{v^2} (\log(a/r) + \log \sigma_\eta) \quad (11)$$

Since the left-hand term in increasing in  $\delta$ , a suitable value  $\delta_0$  can be selected. An upper bound for  $\delta$  can be taken to be (for any  $q \in ]0, 1[$ )

$$\delta = \max\left(\frac{1}{(1-q)^2} - 1, -\frac{2b}{v^2 \log(1 + q\eta/b)}(\log(a/r) + \log \sigma_\eta)\right)$$

which yields an explicit expression of  $\gamma$  which can be plotted as a function of  $a$ , in order to select the minimum value. Of course, because the constants in our bounds are far from being optimal (in particular for the estimate of  $E(Z^+)$ ), the numerical values of the obtained numerical costs are too high to be practical. However, the plots presented in figures 1 and 2 provide a qualitative picture of the situation, showing, in particular, the existence of an optimal value for  $a$ , and that variations in the choice of  $a$  can dramatically affect the required numerical cost for a given risk  $\sigma$  (which was fixed to 0.01).

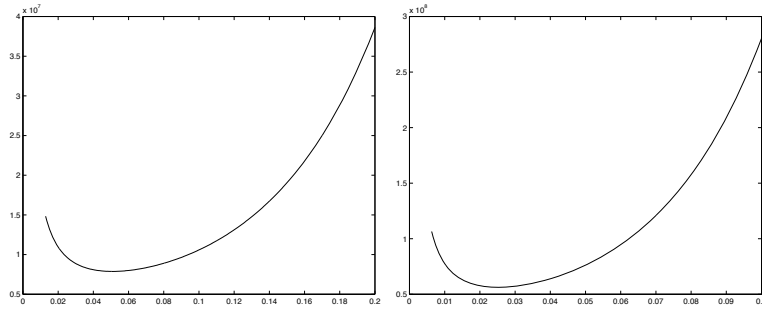


Figure 1: Plots of the minimal values of  $\gamma$  in function of  $a$  for two choices of values for  $|H|, r, \sigma_\eta$

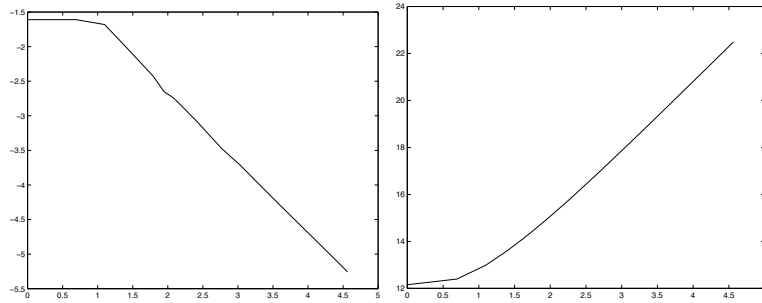


Figure 2: Plots of the optimal  $a$  (left) and  $\gamma$  (right) in a log-scale in function of  $|H|$

Figure 2 shows that the optimal computational cost  $\gamma$  grows at most as a polynomial of  $|H|$ . This can be readily checked by computation. This is very

important, since, for fixed  $a$ , this cost should grow exponentially fast: tuning a correct value of this parameter is essential for the efficiency of the algorithm.

## 7 Refining the estimates

As noticed before, our estimates in theorem 1 are too rough to yield more than a qualitative analysis, which was enough for our purposes. However, based on other standard inequalities, they can be significantly improved. As indicated by the computations of the previous sections, the main room for improvement does not lie in the use of the concentration of measure theorem (theorem 2) but on the estimation of the expectation of the supremums, and thus on the use of theorem 3. Using, for example, better estimates for  $\sigma^2$ , or the sharp constants provided by Massart in ([9]), would not do us much good unless we significantly improve the estimation of the expectations.

Controlling the expectations by subgaussian inequalities like (5), is, when (5) is obtained by Hoeffding's inequality (9), the easiest choice. In fact, as soon as an exponential inequality of the kind

$$\forall s, t \in T, \forall \lambda > 0, P(|Y_s - Y_t| > \lambda) \leq K \exp\left(-\Psi\left(\frac{\lambda}{\gamma(s, t)}\right)\right)$$

is available, in which  $\Psi$  is an increasing convex function, an estimate of the expectation of the supremum can be obtained by a chaining argument (see [16]), together with a control of the Orlicz norm associated to  $\Psi$  (see [13]). When the variance of  $Y_s - Y_t$  can be controled as a function of  $d(s, t)$ , sharper inequalities than Hoeffdings are available (cf. [13], [6]), like Bennett's or Bernstein's. Note that, in general, the control of this variance by explicit constants is far from obvious in the setting in which we are using these inequalities, but computer evaluations can be devised for a given model. It is not sure, however, that even with these refined inequalities, one would be able to obtain practical estimates, which could be used directly while designing an experiment.

Another direction into which our results should be generalized in order to correspond to the actual contexts in which (2) is used in practice, should be the case when the generated samples  $X_1^n, \dots, X_{k_n}^n$  are dependent (typically forming a Markov chain). In fact, generating independant samples of  $\pi_\theta$ , which we have choosed to assume for the simplicity of the exposition, is most of the time untractable, for the same reasons for which the maximum likelihood estimator cannot be computed by deterministic methods.

However, a machinery similar to the one we have used for independent samples is available for Markov chains: concentration inequalities for empirical processes have been proved in [15] (see also [8]). Inequalities similar to Hoeffding's or Bernstein's can be found in [1], [3], [2] or [14]. In fact, exactly the same line of proof can be applied to this more realistic setting, to yield almost similar results. To be more precise, let us quote the theorems which would replace theorem 1 and inequality 9 in the context of Markov chains.

Start with deviation inequalities, and fix some notation. Let the sequence  $Y_1, \dots, Y_k$  be generated by an ergodic and aperiodic stationary Markov chain with transition  $P$  on  $\Omega$  (the invariant distribution is uniquely defined and provides the law of  $X_1$ ). As before, let  $\mathcal{F}$  be a countable family of functions  $f : \Omega \rightarrow \mathbb{R}$  and

$$Z = \sup_{f \in \mathcal{F}} \left\{ \left| \sum_{i=1}^n f(Y_i) \right| \right\}$$

then

**Theorem 4 (Samson, [15])** *There exists a constant  $\Gamma(P)$  such that*

$$P[Z > E(Z) + t] \leq \exp \left( -\frac{1}{8\Gamma(P)} \min \left( \frac{t}{C}, \frac{t^2}{4\sigma^2} \right) \right) \quad (12)$$

where  $C = \sup_{f \in \mathcal{F}} \|f\|_\infty$ ,

$$\sigma^2 = E \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)^2 \right]$$

The constant  $\Gamma(P)$  is related to the mixing properties of the Markov chain. If  $\rho < 1$  and  $\mu > 0$  are such that, for all  $x, y \in \Omega$ , and for all  $p > 0$

$$\|P^p(x, \cdot) - P^p(y, \cdot)\| \leq \mu \rho^p \quad (13)$$

(the norm being the total variation norm), then one may take (see [15])

$$\Gamma(p) = \frac{\mu}{(1 - \sqrt{\rho})^2}$$

The following estimate of the probability of deviation for a single variable may be found in [14].

**Theorem 5 (Rio)** *Let  $\mathcal{F}_1, \dots, \mathcal{F}_k$  be an increasing sequence of  $\sigma$ -algebras. Let  $V_1, \dots, V_k$  be a sequence of real-valued random variables such that, for some sequence  $M_1, \dots, M_k$ ,*

$$\sup \left\{ \|X_i\|_\infty + \left\| 2V_i \sum_{k=i+1}^j E(V_k | \mathcal{F}_i) \right\| : j = i, \dots, n \right\} \leq M_i \quad (14)$$

Then, for all  $\lambda > 0$ ,

$$P(|V_1 + \dots + V_k| > \lambda) \leq \sqrt{e} \exp(-x^2 / (2M_1 + \dots + 2M_k)) \quad (15)$$

If  $X_1, \dots, X_k$  is a stationary Markov-chain on a finite state, with transition probability  $P$  satisfying (13), and  $V_k = U(X_k) - E(U)$  for some real-valued

function  $U$ , where the expectation refers to the invariant distribution  $\pi$  of the chain, and if  $\mathcal{F}_i$  is the  $\sigma$ -algebra generated by  $X_1, \dots, X_i$ , we have

$$E(V_k | \mathcal{F}_i)(x) = \int U(y) P^{k-i}(\cdot, dy) - \int U(y) P^{k-i}(x, dy) \pi(dx)$$

so that  $\|E(V_k | \mathcal{F}_i)\|_\infty \leq \mu |U| \rho^{k-i}$ . This implies that, in this case, one can take

$$M_i = \frac{2\mu \max(|U|^2, 1)}{1 - \rho}$$

which implies

$$P(|V_1 + \dots + V_k| > \lambda) \leq \sqrt{e} \exp\left(-\frac{x^2}{2|U|} \times \frac{1 - \rho}{\mu}\right) \quad (16)$$

It is clear that the proof of theorem 1 can be carried on exactly the same way when  $X_1, \dots, X_k$  in the definition of  $\hat{f}_{\psi, k}$  is an ergodic Markov chain in stationary regime, with invariant distribution  $\pi_\psi$ , satisfying (13) for some constants  $\rho_\psi$  and  $\mu_\psi$ , yielding

**Theorem 6** For  $\psi \in \mathbb{R}^d$ ,

$$\begin{aligned} P & \left( \sup \left\{ |\hat{f}'_{\psi, k}(\theta) - f'(\theta)| : |\theta - \psi| \leq a \right\} > \eta \right) \\ & \leq 6d \exp \left[ -\frac{(1 - \sqrt{\rho_\psi})^2}{\mu_\psi} \min \left( \frac{k\eta - v(a, \psi)\sqrt{k}}{b(a)}, \frac{(k\eta - v(a, \psi)\sqrt{k})^2}{4kb(a)^2} \right) \right] \end{aligned}$$

with  $b(a) = e^{a|H|}(|H| + \eta) + \eta$  and  $v(a, \psi) = 2K'' \sqrt{d\mu_\psi} |H| (|H| + \eta) e^{a|H|} / \sqrt{1 - \rho_\psi}$  where  $K$  and  $K''$  are universal constants.

In addition to the different form taken by Samson's deviation inequality compared to Talagrand's, which have little qualitative impact, the essential new feature for Markov chains is that the bound now has to depend on the mixing speed of the chain, that is on the constants  $\rho_\psi$  and  $\mu_\psi$ . If the mixing speed were uniformly bounded in  $\psi$ , this would have no consequence, but it is typically not the case in the standard practical situation when the sampled distribution is a Gibbs field: one should rather expect  $\rho_\psi$  to degenerate and tend rapidly to 1 when  $\psi$  tends to infinity (typical available estimates are  $\rho_\psi \leq 1 - \exp(-K(H)|\psi|)$  for large  $\psi$ , where  $K(H)$  is a constant which depends on  $H$ ). This implies that it would not be possible to obtain proposition 1, using fixed  $k$  and  $a$ , at least with the inequalities we have proved so far. In fact, it should be necessary to assume that  $k$  adapts to  $\theta_k$ , yielding a condition of the kind

$$\sqrt{k}(1 - \sqrt{\rho_{\theta_k}}) \geq k_0$$

for a large enough constant  $k_0$ . This is similar to the results which have been obtained in [20] in the case of stochastic gradient estimation.

## References

- [1] R. BLUM, J. L. HANSON, D. AND H. KOOPMANS, L, *On the strong law of large numbers for a class of stochastic processes*, Z. Wahrscheinlichkeitstheorie, 3 (1963), pp. 1–11.
- [2] P. DOUKHAN, *Mixing: properties and examples*, vol. 85, Springer-Verlag, 1995.
- [3] P. DOUKHAN, P. MASSART, AND E. RIO, *Invariance principles for absolutely regular empirical processes*, Annales de L'Institut Henri Poincaré, 31 (1995), pp. 393–427.
- [4] C. GEYER, *On the convergence of monte carlo maximum likelihood calculations*, J. R. Statist. Soc, 56 (1994), pp. 261–274.
- [5] C. J. GEYER AND E. THOMPSON, *Constrained monte carlo maximum likelihood for dependent data (with discussion)*, J. of Royal Stat. Soc., 54 (1992).
- [6] M. LEDOUX AND M. TALAGRAND, *Probability in Banach spaces. Isoperimetry and processes*, Springer-Verlag, 1991.
- [7] A. LIPPMAN, *A Maximum Entropy Method for Expert Systems*, PhD thesis, Brown University, 1986.
- [8] K. MARTON, *Measure coconcentration fir a class of random processes*, Probability Theory and rel. fields, 110 (1998), pp. 427–439.
- [9] P. MASSART, *About the constants in talagrand's deviation inequalities for empirical processes*, tech. rep., Laboratoire de statistiques, Université Paris Sud, 1998.
- [10] A. J. MOYEED, R. A. AND BADDELEY, *Stochastic approximation of the mle for a spatial point pattern*, Scand. J. Statist., 18 (1991), pp. 39–50.
- [11] A. PENTINEN, *Stochastic approximations in the maximum likelihood inference for markov random fields*, in Frontiers in Pure and Applied Probabilities, H. N. et al., ed., 1993, pp. 197–205.
- [12] A. PENTTINEN, *Modeling interactions in spatial point patterns: parameter estimation by the maximum likelihood method*, PhD thesis, University of Jyväskylä, 1984.
- [13] D. POLLARD, *Empirical processes: theory and applications*, NSF-CMBS Regional conference series in probability and statistics, 1990.
- [14] E. RIO, *Théorie asymptotique des processus altoires faiblement dépendants*, Springer Verlag, 2000.

- [15] P.-M. SAMSON, *Concentration of measure for markov chains and  $\phi$ -mixing processes*, Annals of probability, (To appear).
- [16] M. TALAGRAND, *Majorizing measures: the generic chaining*, The Annals of Probability, 24 (1996), pp. 1049–1103.
- [17] L. YOUNES, *Couplage de l'estimation et du recuit pour des champs de gibbs*, C. R. Acad. Sc. Paris, série I, (1886).
- [18] ———, *Estimation and annealing for gibbsian fields*, Ann. de l'Inst. Henri Poincaré, 2 (1988).
- [19] ———, *Maximum likelihood estimation for gibbs fields*, in Spatial Statistics and Imaging: Proceedings of an AMS-IMS-SIAM Joint Summer Research Conference, A. Possolo, ed., Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, California, 1991.
- [20] ———, *On the convergence of markovian stochastic algorithms with rapidly decreasing ergodicity rates*, Stochastics and Stochastics Reports, 65 (1999), pp. 177–228.